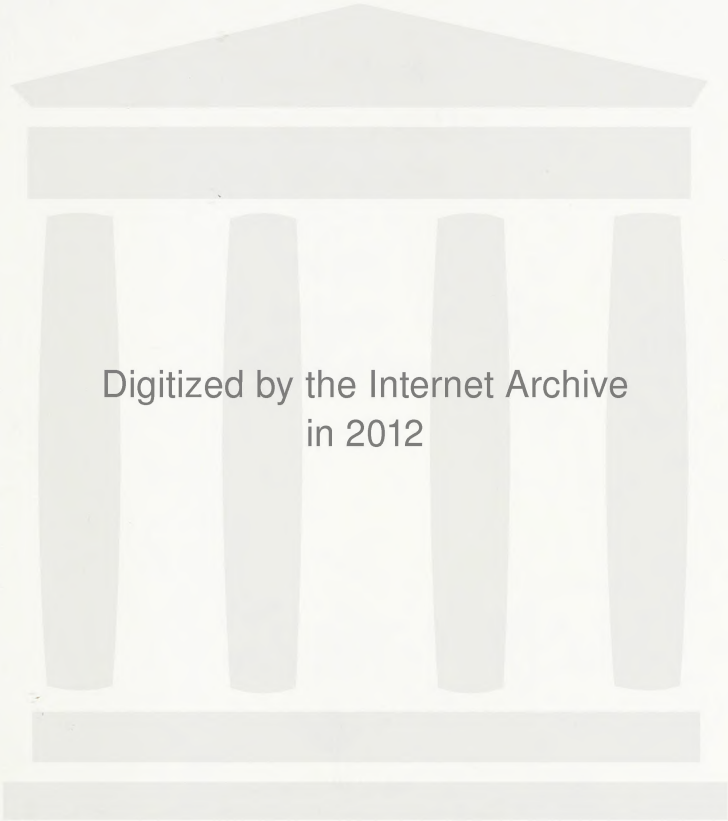


Boston Public Library
REFERENCE


Digitized by the Internet Archive
in 2012

http://archive.org/details/newencyclopaedia26chic

The New Encyclopædia Britannica

Volume 26

MACROPÆDIA

Knowledge in Depth

FOUNDED 1768
15TH EDITION

Encyclopædia Britannica, Inc.
Jacob E. Safra, Chairman of the Board
Ilan Yeshua, Chief Executive Officer

Chicago
London/New Delhi/Paris/Seoul
Sydney/Taipei/Tokyo

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

Eleventh Edition
© 1911
By Encyclopædia Britannica, Inc.

Twelfth Edition
© 1922
By Encyclopædia Britannica, Inc.

Thirteenth Edition
© 1926
By Encyclopædia Britannica, Inc.

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

Fifteenth Edition
© 1974, 1975, 1976, 1977, 1978, 1979, 1980, 1981, 1982, 1983, 1984, 1985, 1986,
1987, 1988, 1989, 1990, 1991, 1992, 1993, 1994, 1995, 1997, 1998, 2002, 2003
By Encyclopædia Britannica, Inc.

© 2003
By Encyclopædia Britannica, Inc.

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

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

Printed in U.S.A.

Library of Congress Control Number: 2002113989
International Standard Book Number: 0-85229-961-3

Britannica may be accessed at http://www.britannica.com on the Internet.

CONTENTS

1 Pre-Columbian Civilizations
45 Prehistoric Peoples and Cultures
71 Printing, Typography, and Photoengraving
112 Printmaking
135 Probability Theory
149 Procedural Law
171 Propaganda
180 Property Law
206 Protestantism
268 Protists
279 Protozoa
289 Psychological Tests and Measurement
294 Public Administration
310 Public Opinion
317 Public Works
415 Publishing
450 Puppetry
458 Radar
471 Radiation
501 Relativity
509 The Study and Classification of Religions
530 Systems of Religious and Spiritual Belief
578 Religious Experience
591 Religious Symbolism and Iconography
601 Rembrandt
609 Reproduction and Reproductive Systems
688 Reptiles
725 Respiration and Respiratory Systems
758 Rhetoric
765 Rio de Janeiro
770 Sacred Rites and Ceremonies
843 Rivers
870 Rock Music
877 Roman Catholicism
914 Rome
933 Franklin D. Roosevelt
938 Jean-Jacques Rousseau
943 Russia
1003 Russian Literature
1012 Rutherford
1014 Sacred Offices and Orders
1037 Saint Petersburg

Pre-Columbian Civilizations

Pre-Columbian civilization refers to the aboriginal American Indian cultures that evolved in Meso-America (part of Mexico and Central America) and the Andean region (western South America) prior to Spanish exploration and conquest in the 16th century. The pre-Columbian civilizations were extraordinary developments in human society and culture, ranking with the early civilizations of Egypt, Mesopotamia, and China. Like the ancient civilizations of the Old World, those in the New World were characterized by kingdoms and empires, great monuments and cities, and refinements in the arts, metallurgy, and writing; the ancient civilizations of the Americas also display in their histories similar cyclical patterns of growth and decline, unity and disunity.

In the New World the roots of civilization lay in a native agricultural way of life. These agricultural beginnings go back several millennia, to early post-Pleistocene times (*c.* 7000 BC) and to the first experimentations by the early Americans with plant cultivation. The domestication of successful food plants proved to be a long, slow process, and it was not until much later that a condition of permanent village farming life was achieved in the tropical latitudes of the two continents.

Sedentary village farming in Meso-America came into being by about 1500 BC. Corn (maize), beans, squashes, chili peppers, and cotton were the most important crops. These early villagers wove cloth, made pottery, and practiced other typical Neolithic skills. It appears that such villages were economically self-contained and politically autonomous, with an egalitarian social order. But rather quickly after this—between about 1200 and 900 BC—the building of large earthen pyramids and platforms and the carving of monumental stone sculptures signaled significant changes in this heretofore simple social and political order. These changes first appeared in the southern Gulf coast region of what is now Mexico; and the sculptures, rendered in a style now called Olmec, are presumed to depict chiefs or rulers. From these and other archaeological indications it has been inferred that a class-structured and politically centralized society developed. There appeared subsequently other large capital towns and cities in neighbouring regions that also displayed a similar Olmec art style. This Olmec horizon (*i.e.*, a cultural diffusion that is contemporaneous at widely scattered sites) represents the first climax, or era of "unification," in the history of Meso-American civilization.

After about 500 BC the Olmec "unification" gave way to an era (consisting of the Late Formative and Classic periods) of separate regional styles and kingdoms. These lasted until *c.* AD 700–900. Among these are the well-known Maya, Zapotec, Totonac, and Teotihuacán civilizations. While sharing a common Olmec heritage, they also displayed many differences. For example, the Maya excelled in the intellectual pursuits of hieroglyphic writing, calendar making, and mathematics, while the Teotihuacán civilization placed its emphasis on political and commercial power. Teotihuacán, in the Valley of Mexico, was an urban centre of some 150,000 people, and the influence of its civilization eventually radiated over much of Meso-America. As such, Teotihuacán constituted a second grand civilizational climax or "unification" (AD 400–600). Teotihuacán power waned after about 600, and a "time of troubles" ensued, during which a number of states and nascent empires competed for supremacy. Among these competitors were the Toltecs of Tula, in central Mexico, who held sway from perhaps 900 to 1200 (the Early Postclassic Period). After their decline (in the Late Postclassic Period), another interregnum of warring states lasted until 1428, when the Aztec defeated the rival city of Azcapotzalco and emerged as the dominant force in central Mexico. This last native Meso-American empire was conquered by Hernán Cortés (or Cortéz) and the Spaniards in 1521.

In the Andean area, the threshold of a successful village agricultural economy can be placed at *c.* 2500 BC, or somewhat earlier than was the case in Meso-America. The oldest primary food crops there were the lima bean and the potato, which had long histories of domestication in the area, although corn appeared soon after the beginnings of settled village life. Indications of a more complex sociopolitical order—huge platform mounds and densely populated centres—occurred very soon after this (*c.* 1800 BC); however, these early Andean civilizations continued for almost a millennium before they participated in a shared stylistic "unification." This has become known as the Chavín horizon, and Chavín sculptural art has been found throughout the northern part of the area.

The Chavín horizon disappeared after about 500 BC, and it was replaced by regional styles and cultures that lasted until about AD 600. This period of regionalization (called the Early Intermediate Period) saw the florescence of a number of large kingdoms both on the Pacific coast and in the Andean highlands; among them were the Mochica, Early Lima, Nazca, Recuay, and Early Tiahuanaco. The period was brought to an end by the Tiahuanaco–Huari horizon (Middle Horizon) (600–1000), which was generated from the highland cities of Tiahuanaco (in modern northern Bolivia) and Huari (in central highland Peru). There is evidence—such as the construction of new centres and cities—that this Tiahuanaco–Huari phenomenon, at least in many regions, was a tightly controlled political empire. The horizon and its influences, as registered in ceramics and textiles, died away rather gradually in the ensuing centuries, and it was replaced by the several regional styles and kingdoms of what has become known as the Late Intermediate Period (1000–1438).

The terminal date of the Late Intermediate Period marked the beginning of the Inca horizon and of the Inca conquests, which spread from the Inca capital, Cuzco, in the southern highlands of modern Peru. By 1533, when Francisco Pizarro and his cohorts took over the empire, it extended from what is now the Ecuador–Colombia border to central Chile.

The synchroneity of horizon unifications and alternating regionalizations in Meso-America and the Andean region is striking and prompts the question of communication between these two areas of pre-Columbian high civilization. Although it is known that there were contacts—with the result that knowledge of food plants, ceramics, and metallurgy was shared between the two areas—it is also highly unlikely that political or religious ideologies were so spread. Rather, the peoples of each of these major cultural areas appear to have responded to their own internally generated stimuli and to have followed essentially separate courses of development. There are fundamental differences between the two cultural traditions. Thus, in Meso-America there was, from early on, a profound interest in hieroglyphic writing and calendar making. Religious ideology, judged from art and iconography, was more highly developed in Meso-America than in the Andean region. In Meso-America the market was a basic institution; it does not appear to have been so in the Andes, where the redistributive economy of the Inca empire—with such features as its government warehouses and a system of highways—must have had deep roots in the past. On the other hand, in the early development and deployment of metallurgy and in governmental institutions and empire-building, the ancient Peruvians were much more efficient than their Meso-American contemporaries. (G.R.W.)

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 951 and 952.

This article is divided into the following sections:

Meso-American civilization 2
Pre-Classic and Classic periods 3
Early hunters (to 6500 BC)
Incipient agriculture (6500–1500 BC)
Early Formative Period (1500–900 BC)
Early village life
Early religious life
The rise of Olmec civilization
Middle Formative Period (900–300 BC)
Horizon markers
Olmec civilization at La Venta
Olmec colonization in the Middle Formative
Early Monte Albán
The Valley of Mexico in the Middle Formative
The Maya in the Middle Formative
Late Formative Period (300 BC–AD 100)
Valley of Mexico
Valley of Oaxaca
Veracruz and Chiapas
Izapan civilization
The earliest Maya civilization of the lowlands
Early Classic Period (AD 100–600)
Definition of the Classic
Teotihuacán
Cholula
Classic Central Veracruz
Southern Veracruz
Classic Monte Albán
The Maya highlands and Pacific coast
Classic civilization in the Maya lowlands: Tzakol phase
Late Classic non-Maya Meso-America (600–900)
Late Classic lowland Maya (600–900)
Settlement pattern
Major sites
Maya art of the Late Classic
The Maya calendar and writing system
Classic Maya religion
Society and political life
The collapse of Classic Maya civilization
Postclassic Period (900–1519) 17
Definition of the Postclassic
The historical annals
The rise of the Aztec
The question of the Toltec
Archaeological remains of Postclassic civilization
Tula
Chichén Itzá
Archaeological unity of the Postclassic
Aztec culture to the time of the Spanish conquest
The nature of the sources
Agriculture
Social and political organization
Tenochtitlán
Aztec religion
Andean civilization 26
The nature of Andean civilization 27
Agricultural adaptation
The cold as a resource
The highlands and the low countries
The pre-Inca periods 29
The Late Preceramic
The Initial Period
The Early Horizon
The Early Intermediate Period
The southern coast
The northern coast
The north highlands
The south highlands
The Middle Horizon
The Late Intermediate Period
The Chimú state
The Chincha
The Inca 35
The origins and expansion of the Inca state
The nature of the sources
Settlement in the Cuzco Valley
The beginnings of external expansion
Pachacuti Inca Yupanqui
Topa Inca Yupanqui
Huayna Cápac
Civil war on the eve of the Spanish conquest
The Spanish conquest
Inca culture at the time of the conquest
Social and political structure
Inca technology and intellectual life
Inca religion
Bibliography 43

MESO-AMERICAN CIVILIZATION

The term Meso-America denotes the part of Mexico and Central America that was civilized in pre-Spanish times. In many respects, the American Indians who inhabited Meso-America were the most advanced native peoples in the Western Hemisphere. The northern border of Meso-America runs west from a point on the Gulf coast of Mexico above the modern port of Tampico, then dips south to exclude much of the central desert of highland Mexico, meeting the Pacific coast opposite the tip of Baja (Lower) California. On the southeast, the boundary extends from northwestern Honduras on the Caribbean across to the Pacific shore in El Salvador. Thus, about half of Mexico, all of Guatemala and Belize, and parts of Honduras and El Salvador are included in Meso-America.

The two major regions

Geographically and culturally, Meso-America consists of two strongly contrasted regions: highland and lowland. The Mexican highlands are formed mainly by the two Sierra Madre ranges that sweep down on the east and west. Lying athwart them is a volcanic cordillera stretching from the Atlantic to the Pacific. The high valleys and landlocked basins of Mexico were important centres of pre-Spanish civilization. In the southeastern part of Meso-America lie the partly volcanic Chiapas–Guatemala highlands. The lowlands are primarily coastal. Particularly important was the littoral plain extending south along the Gulf of Mexico, expanding to include the Petén-Yucatán Peninsula, homeland of the Mayan peoples.

Agriculture in Meso-America was advanced and complex. A great many crops were planted, of which corn, beans, and squashes were the most important. In the highlands, hoe cultivation of more or less permanent fields was the rule, with such intensive forms of agriculture as irrigation and chinampas (the so-called floating gardens reclaimed from lakes or ponds) known in some regions. In contrast, lowland agriculture was frequently of the slash-and-burn variety; a patch of jungle was first selected, felled and burned toward the end of the dry season, and then planted with a digging stick in time for the first rains. After a few years of planting, the field was abandoned to the forest, as competition from weeds and declining soil fertility resulted in diminishing yields. There is good evidence, however, that the slash-and-burn system of cultivation was often supplemented by "raised-field" cultivation in the lowlands; these artificially constructed earthen hillocks built in shallow lakes or marshy areas were not unlike the chinampas of the Mexican highlands. In addition, terraces were constructed and employed for farming in some lowland regions. Nevertheless, the demographic potential for agriculture was probably always greater in the highlands than it was in the lowlands, and this was demonstrated in the more extensive urban developments in the former area.

The extreme diversity of the Meso-American environment produced what has been called symbiosis among its subregions. Interregional exchange of agricultural products, luxury items, and other commodities led to the development of large and well-regulated markets in which cacao beans were used for money. It may have also led to large-scale political unity and even to states and empires. High agricultural productivity resulted in a nonfarming class of artisans who were responsible for an advanced stone architecture, featuring the construction of stepped pyramids, and for highly evolved styles of sculpture, pottery, and painting.



Principal sites of Meso-American civilization.
Adapted from *Grosser Historischer Weltatlas*, vol. II, *Mittelalter* (1970); Bayerischer Schulbuch-Verlag, Munich

The Meso-American system of thought, recorded in folding-screen books of deerskin or bark paper, was perhaps of even greater importance in setting them off from other New World peoples. This system was ultimately based upon a calendar in which a ritual cycle of 260 (13 × 20) days intermeshed with a "vague year" of 365 days (18 × 20 days, plus five "nameless" days), producing a 52-year Calendar Round. The religious life was geared to this cycle, which is unique to them. The Meso-American pantheon was associated with the calendar and featured an old, dual creator god; a god of royal descent and warfare; a sun god and moon goddess; a rain god; a culture hero called the Feathered Serpent; and many other deities. Also characteristic was a layered system of 13 heavens and nine underworlds, each with its presiding god. Much of the system was under the control of a priesthood that also maintained an advanced knowledge of astronomy.

Language groups

As many as 14 language families were found in Meso-America, but most can be grouped into three large "phyla": Uto-Aztecan, Macro-Mayan, and Oto-Manguean. A dominant role was played by Uto-Aztecan, particularly by speakers of the Nahua groups of which Náhuatl, official tongue of the Aztec Empire, was the most important. While Macro-Mayan includes Zoquean and Totonacan, its largest member is Mayan, with a number of mutually unintelligible languages, at least some of which were spoken by the inhabitants of the great Maya ceremonial centres. The modern Mexican state of Oaxaca is now the centre of the heterogeneous Oto-Manguean phylum; but the only linguistic groups that played any great part in Meso-American civilization were the Mixtec and Zapotec, both of which had large, powerful kingdoms at the time of the Spanish conquest. Still a linguistic puzzle are such languages as Tarascan, mother tongue of an "empire" in western Mexico that successfully resisted Aztec encroachments; it has no sure relatives, although some linguistic authorities have linked it with the Quechua language of distant Peru. Huavean and Xinca-Lencan are little-known language groups of southeastern Meso-America.

Pre-Classic and Classic periods

EARLY HUNTERS (TO 6500 BC)

The time of the first peopling of Meso-America remains a puzzle, as it does for that of the New World in general. It is widely accepted that groups of Mongoloid peoples entered the hemisphere from northeastern Siberia, perhaps by a land bridge that then existed, at some time in the Late Pleistocene, or Ice Age. There is abundant evidence that, by 11,000 BC, hunting peoples had occupied most of the New World south of the glacial ice cap covering northern North America. These men hunted such large grazing mammals as mammoth, mastodon, horse, and camel, armed with spears to which were attached finely made, bifacially chipped points of stone. Finds in Meso-America, however, confirm the existence of a "prebifacial-point horizon," a stage known to have existed elsewhere in the Americas, and suggest that it is of very great age. In 1967 archaeologists working at the site of Tlapacoya, southeast of Mexico City, uncovered a well-made blade of obsidian associated with a radiocarbon date of about 21,000 BC. Near Puebla, Mex., excavations in the Valsequillo region revealed cultural remains of human groups that were hunting mammoth and other extinct animals, along with unifacially worked points, scrapers, perforators, burins, and knives. A date of about 21,800 BC has been suggested for the Valsequillo finds.

Tepexpan finds

More substantial information on Late Pleistocene occupations of Meso-America comes from excavations near Tepexpan, northeast of Mexico City. The excavated skeletons of two mammoths showed that these beasts had been killed with spears fitted with lancelike stone points and butchered on the spot. A possible date of about 8000 BC has been suggested for the two mammoth kills. In the same geologic layer as the slaughtered mammoths was found a human skeleton; this Tepexpan "man" has been shown to be female and rather a typical American Indian of modern form. While the association with the mammoths was first questioned, fluorine tests have proved them to be contemporary.

The environment of these earliest Meso-Americans was quite different from that existing today, for volcanoes were then extremely active, covering thousands of square miles with ashes. Temperatures were substantially lower, and local glaciers formed on the highest peaks. Conditions were ideal for the large herds of grazing mammals that roamed Meso-America, especially in the highland valleys, much of which consisted of cool, wet grasslands not unlike the plains of the northern United States. All of this changed around 7000 BC, when worldwide temperatures rose and the great ice sheets of northern latitudes began their final retreat. This brought to an end the successful hunting way

of life that had been followed by Meso-Americans, although man probably also played a role in bringing about the extinction of the large game animals.

INCIPIENT AGRICULTURE (6500–1500 BC)

The most crucial event in the prehistory of Meso-America was man's capture of the food energy contained in plants. This process centred on three plants: Indian corn (maize), beans, and squashes. Since about 90 percent of all food calories in the diet of Meso-Americans eventually came from corn, archaeologists for a long time have sought the origins of this plant—which has no wild forms existing today—in order to throw light on the agricultural basis of Meso-American civilization.

The three major food plants

The search for Meso-American agricultural origins has been carried forward most successfully through excavations in dry caves and rock shelters in the modern southern Mexican states of Puebla and Oaxaca. Sequences from these archaeological sites show a gradual transition from the Early Hunting to the Incipient Cultivation periods. At the Guila Naquitz cave, in Oaxaca, there are indications that the transition began as early as 8900 BC; finds from caves in the Tehuacán valley of Puebla, however, offer more substantial evidence of the beginnings of plant domestication at a somewhat later time. There, the preservation of plant remains is remarkably good, and from these it is evident that shortly after 6500 BC the inhabitants of the valley were selecting and planting seeds of chili peppers, cotton, and one kind of squash. Most importantly, between 5000 and 3500 BC they were beginning to plant mutant forms of corn that already were showing signs of the husks characteristic of domestic corn.

One of the problems complicating this question of the beginnings of early corn cultivation is related to a debate between paleobotanists on wild versus domesticated strains. One school of thought holds that the domesticated races of the plant developed from a wild ancestor. The other opinion is that there was never such a thing as wild corn, that, instead, corn (*Zea mays*) developed from a related grass, teosinte (*Zea mexicana,* or *Euchlaena mexicana*). In any event, by 5000 BC corn was present and being used as a food, and between 2,000 and 3,000 years after that it had developed rapidly as a food plant. It has been estimated that there is more energy present in a single kernel of some modern races than there was in an ear of this ancient Tehuacán corn. Possibly some of this was popped, but a new element in food preparation is seen in the metates (querns) and manos (handstones) that were used to grind the corn into meal or dough.

Beans appeared after 3500 BC, along with a much improved race of corn. This enormous increase in the amount of plant food available was accompanied by a remarkable shift in settlement pattern. In place of the temporary hunting camps and rock shelters, which were occupied only seasonally by small bands, semipermanent villages of pit houses were constructed on the valley floor. Increasing sedentariness is also to be seen in the remarkable bowls and globular jars painstakingly pecked from stone, for pottery was as yet unknown in Meso-America.

In the centuries between 3500 and 1500 BC, plant domestication began in what had been hunting-gathering contexts, as on the Pacific coast of Chiapas and on the Veracruz Gulf coast and in some lacustrine settings in the Valley of Mexico. It seems probable that early domesticated plants from such places as the Tehuacán valley were carried to these new environmental niches. In many cases, this shift of habitat resulted in genetic improvements in the food plants.

Pottery, which is a good index to the degree of permanence of a settlement (because of its fragility it is difficult to transport), was made in the Tehuacán valley by 2300 BC. Fired clay vessels were made as early as 4000 BC in Ecuador and Colombia, and it is probable that the idea of their manufacture gradually diffused north to the increasingly sedentary peoples of Meso-America.

The picture, then, is one of man's growing control over his environment through the domestication of plants; animals played a very minor role in this process, with only the dog being surely domesticated before 1500 BC. At any rate, by 1500 BC the stage was set for the adoption of a fully settled life, with many of the sedentary arts already present. The final step was taken only when native agriculture in certain especially favoured subregions became sufficiently effective to allow year-round settlement of villages.

EARLY FORMATIVE PERIOD (1500–900 BC)

Early village life. It is fairly clear that the Mexican highlands were far too dry during the much warmer interval that prevailed from 5000 to 1500 BC for agriculture to supply more than half of a given population's energy needs. This was not the case along the alluvial lowlands of southern Meso-America, and it is no accident that the best evidence for the earliest permanent villages in Meso-America comes from the Pacific littoral of Chiapas (Mexico) and Guatemala, although comparable settlements also have been reported from both the Maya lowlands (Belize) and the Veracruz Gulf coast.

Barra, Ocós, and Cuadros phases

The Barra (*c.* 1800–1500 BC), Ocós (1500–1200 BC), and Cuadros (1100–900 BC) phases of the Pacific coasts of Chiapas and Guatemala are good examples of early village cultures. The Barra phase appears to have been transitional from earlier preagricultural phases and may not have been primarily dependent upon corn farming; but people of the Ocós and Cuadros phases raised a small-eared corn known as *nal-tel,* which was ground on metates and manos and cooked in globular jars. From the rich lagoons and estuaries in this area, the villagers obtained shellfish, crabs, fish, and turtles. Their villages were small, with perhaps 10 to 12 thatched-roof houses arranged haphazardly.

Ocós pottery is highly developed technically and artistically. Something of the mental life of the times may be seen in the tiny, handmade clay figurines produced by the Ocós villagers. These, as in Formative cultures generally throughout Meso-America, represent nude females and may have had something to do with a fertility cult. The idea of the temple-pyramid may well have taken root by that time, for one Ocós site has produced an earthen mound about 26 feet (eight metres) high that must have supported a perishable building. The implication of the site is that, with increasing prosperity, some differentiation of a ruling class had taken place, for among the later Meso-Americans the ultimate function of a pyramid was as a final resting place for a great leader.

Eventually, effective village farming with nucleated settlements occupied throughout the year appeared in the highlands. But perhaps from the very beginning of Formative life there were different cultural responses directed toward both kinds of environment. In the highlands, divided into a number of mutually contrasting environments no one of which could have provided sufficient resources for the subsistence of a single settlement, villages were presumably linked to each other symbiotically. In the lowlands, particularly in the littoral, one especially favourable environment, such as the lagoon–estuary system, may have been so rich in resources that villages within it would have been entirely self-sufficient. In effect, the former would have resulted in a cultural integration based upon trade, while the latter would have been integrated, if at all, by a unity of likeness. The two kinds of civilization that eventually arose in each region—the highlands definitely urban, the lowlands less so—reflect the same contrast.

Early religious life. Early religious phenomena can only be deduced from archaeological remains. Numerous clay figurines found in tombs afford little evidence of religious beliefs during the agricultural Pre-Classic periods of Zacatenco and Ticomán (roughly 1500 to the 1st century BC). It is possible, however, that terra-cotta statuettes of women were meant to represent an agricultural deity, a goddess of the crops. Two-headed figurines found at Tlatilco, a site of the late Pre-Classic, may portray a supernatural being. Clay idols of a fire god in the form of an old man with an incense burner on his back date from the same period.

The first stone monument on the Mexican plateau is the pyramid of Cuicuilco, near Mexico City. In fact, it is rather a truncated cone, with a stone core; the rest is made of sun-dried brick with a stone facing. It shows the main features of the Mexican pyramids as they were developed in later times. It was doubtless a religious monument,

crowned by a temple built on the terminal platform and surrounded with tombs. The building of such a structure obviously required a protracted and organized effort under the command of the priests.

The final phase of the Pre-Classic cultures of the central highland forms a transition from the village to the city, from rural to urban life. This was a far-reaching social and intellectual revolution, bringing about new religious ideas together with new art forms and theocratic regimes. It is significant that Olmec statuettes have been found at Tlatilco with late Pre-Classic material.

The rise of Olmec civilization. It was once assumed that the Formative stage was characterized only by simple farming villages. It is now realized, however, that coexisting with these peasantlike cultures was a great civilization, the Olmec, that had arisen in the humid lowlands of southern Veracruz and Tabasco, in Mexico.

The Olmec were perhaps the greatest sculptors of ancient Meso-America. Whether carving tiny jade figures or gigantic basalt monuments, they worked with a great artistry that led a number of archaeologists to doubt their considerable antiquity, although radiocarbon dates from the type site of La Venta showed that Olmec civilization was indeed Formative, its beginning dating to at least 1,000 years before the advent of Maya civilization.

The San Lorenzo site

San Lorenzo is now established as the oldest known Olmec centre. In fact, excavation has shown it to have taken on the appearance of an Olmec site by 1150 BC and to have been destroyed, perhaps by invaders, around 900 BC. Thus, the Olmec achieved considerable cultural heights within the Early Formative, at a time when the rest of Meso-America was at best on a Neolithic level. The reasons for its precocious rise must have had something to do with its abundant rainfall and the rich alluvial soil deposited along the broad, natural levees that flank the waterways of the southern Gulf coast. Thus, the ecological potential for corn farmers in this counterpart of Mesopotamia's Fertile Crescent was exceptionally high. The levee lands, however, were not limitless, and increasingly dense populations must inevitably have led to competition for their control. Out of such conflicts would have crystallized a dominant landowning class, perhaps a group of well-armed lineages. It was this elite that created the Olmec civilization of San Lorenzo.

In appearance, the San Lorenzo site is a compact plateau rising about 160 feet above the surrounding plains. Cutting into it are deep ravines that were once thought to be natural but that are now known to be man-made, formed by the construction of long ridges that jut out from the plateau on the northwest, west, and south sides. Excavations have proved that at least the top 25 to 35 feet of the site was built by human labour. There are about 200 small mounds on the surface of the site, each of which once supported a dwelling house of pole and thatch, which indicates that it was both a ceremonial centre, with political and religious functions, and a minuscule town.

San Lorenzo is most noted for its extraordinary stone monuments. Many of these, perhaps most, were deliberately smashed or otherwise mutilated about 900 BC and buried in long lines within the ridges and elsewhere at the site. The monuments weighed as much as 44 tons and were carved from basalt from the Cerro Cintepec, a volcanic flow in the Tuxtla Mountains about 50 air miles to the northwest. It is believed that the stones were somehow dragged down to the nearest navigable stream and from there transported on rafts up the Coatzacoalcos River to the San Lorenzo area. The amount of labour involved must have been enormous and so would have the social controls necessary to see the job through to its completion.

The "colossal heads"

Most striking are the "colossal heads," human portraits on a stupendous scale. Several of these are now known from San Lorenzo, the largest of which is nine feet high. The visages are flat-faced, with thickened lips and staring eyes. Each has a headgear resembling a football helmet, and it is entirely possible that these "helmets" were in fact protective coverings in a rubber-ball game that is known from Olmec figurines to have been played at San Lorenzo.

The central theme of the Olmec religion was a pantheon of deities each of which usually was a hybrid between jaguar and human infant, often crying or snarling with open mouth. This "were-jaguar" is the hallmark of Olmec art, and it was the unity of objects in this style that first suggested to scholars that they were dealing with a new and previously unknown civilization. There is actually a whole spectrum of such were-jaguar forms in Olmec art, ranging from the almost purely feline to the human in which only a trace of jaguar can be seen.

These Olmec monuments were generally carved in the round with great technical prowess, even though the only methods available were pounding and pecking with stone tools. Considerable artistry can also be seen in the pottery figurines of San Lorenzo, which depict nude and sexless individuals with were-jaguar traits.

Olmec trade

Exotic raw materials brought into San Lorenzo from distant regions suggest that the early Olmec controlled a large trading network over much of Meso-America. Obsidian, used for blades, flakes, and dart points, was imported from highland Mexico and Guatemala. Most items were obviously for the luxury trade, such as iron ore for mirrors and various fine stones like serpentine employed in the lapidary industry. One material that is conspicuously absent, however, is jade, which does not appear in Olmec sites until after 900 BC and the fall of San Lorenzo.

There is evidence that the Olmec sent groups from their Gulf coast "heartland" into the Meso-American highlands toward the end of the Early Formative, in all likelihood to guarantee that goods bound for San Lorenzo would reach their destination. San Lorenzo-type Olmec ceramics and figurines have been found in burials at several sites in the Valley of Mexico, such as Tlapacoya, and in the state of Morelos. The Olmec involvement with the rest of Meso-America continued into the Middle Formative and probably reached its peak at that time.

San Lorenzo is not the only Olmec centre known for the Early Formative. Laguna de los Cerros, just south of the Cerro Cintepec in Veracruz, appears to have been a large Olmec site with outstanding sculptures. La Venta, just east of the Tabasco border, was another contemporary site, but it reached its height after San Lorenzo had gone into decline.

MIDDLE FORMATIVE PERIOD (900–300 BC)

Horizon markers. Once ceramics had been adopted in Meso-America, techniques of manufacture and styles of shape and decoration tended to spread rapidly and widely across many cultural frontiers. These rapid diffusions, called horizons, enable archaeologists to link different cultures on the same time level. For the Early Formative, colour zones of red pigment set off by incised lines; complex methods of rocker stamping (a mode of impressing the wet clay with the edge of a stick or shell); the *tecomate,* or globular, neckless jar; and Olmec excised pottery are good horizon markers. The beginning of the Middle Formative over much of Meso-America is marked by the diffusion of a very hard, white pottery, decorated with incised lines, and by solid pottery figurines with large, staring eyes formed by a punch. The people who replaced and probably overthrew the Olmec of San Lorenzo in about 900 BC had such pottery and figurines, the ultimate origins of which are still a puzzle.

During the Middle Formative, cultural regionalism increased, although the Olmec presence can be widely detected. The transition to fully settled life had taken place everywhere, and burgeoning populations occupied hamlets, villages, and perhaps even small towns throughout Meso-America, both highland and lowland.

Olmec civilization at La Venta. La Venta was located on an almost inaccessible island, surrounded at that time by the Tonalá River; the river now divides the states of Veracruz and Tabasco. As San Lorenzo's fortunes fell, La Venta's rose, and between 800 and 400 BC it was the most important site in Meso-America.

At the centre of La Venta is a 100-foot-high mound of earth and clay that may well house the tomb of a great Olmec ruler. Immediately north of the Great Mound is a narrow north–south plaza flanked by a pair of long mounds. Beyond the plaza is a ceremonial enclosure surrounded by a "fence" made entirely of upright shafts of

columnar basalt. A low, round mound on the north side of the ceremonial enclosure contained several tombs, one of which was surrounded and covered by basalt columns. In this tomb were found the bundled remains of two children, accompanied by magnificent ornaments of jade. Offerings were not only placed with the dead but were also deposited as caches in the site, especially along the north–south axis of the ceremonial centre.

Olmec mirrors

Among the most beautiful objects manufactured by the Olmec were the concave mirrors of iron ore, which were pierced to be worn around the neck. These could throw pictures on a flat surface and could probably start fires on hot tinder. Olmec leaders at La Venta, whether they were kings or priests, undoubtedly used them to impress the populace with their seemingly supernatural powers. Olmec sculptors continued to produce the basalt monuments, including colossal heads and "altars," that have been found at La Venta. Significantly, an increasing number of monuments were carved in relief, and some of these were stelae with rather elaborate scenes obviously based upon historical or contemporary events.

Olmec colonization in the Middle Formative. From the Middle Formative there are important Olmec sites located along what appears to have been a highland route to the west to obtain the luxury items that seemed to have been so desperately needed by the Olmec elite; *e.g.*, jade, serpentine, iron ore for mirrors, cinnabar, and so forth. Olmec sites in Puebla, the Valley of Mexico, and Morelos are generally located at the ends of valleys near or on major passes; they were perhaps trading stations garrisoned by Olmec troops. The largest of these sites is Chalcatzingo, Morelos, a cult centre located among three denuded volcanic peaks rising from a plain. On a talus slope at the foot of the middle peak are huge boulders on which have been carved Olmec reliefs in La Venta style. The principal relief shows an Olmec woman, richly garbed, seated within the mouth of a cave; above her, cumulus clouds pour down rain.

Similar Olmec reliefs, usually narrative and often depicting warriors brandishing clubs, have been located on the Pacific plain of Chiapas (Mexico) and Guatemala. Since about 1960, spectacular Olmec cave paintings have been found in Guerrero, offering some idea of what the Olmec artists could do when they worked with a large spectrum of pigments and on flat surfaces.

Olmec culture or civilization did not spread eastward from its Veracruz–Tabasco centres into the Maya lowlands, but occasional Olmec artifacts have been found in Formative Maya contexts, such as at Seibal, in southern Petén, Guatemala. Maya Formative Period occupations, represented by settled farming villages and well-made ceramics, date to *c.* 1000 BC in the lowlands of Guatemala and Belize. It seems reasonably certain, however, that at this early date great ceremonial centres, comparable to those of Olmec San Lorenzo or La Venta, were never constructed in the Maya lowlands.

It was formerly thought that the Olmec worshiped only one god, a rain deity depicted as a were-jaguar, but study has shown that there were at least 10 distinct gods represented in Olmec art. Surely present were several important deities of the later, established Meso-American pantheon, such as the fire god, rain god, corn god, and Feathered Serpent. Other aspects of mental culture are less well-known; some Olmec jades and a monument from La Venta have non-calendrical hieroglyphs, but none of this writing has been deciphered.

To sum up the Olmec achievement, not only was this the first high culture in Meso-America—one that had certainly achieved political statehood—but either it or cultures influenced by it lie at the base of every other Meso-American civilization.

Early Monte Albán. Monte Albán is a prominent series of interconnected hills lying near Oaxaca, Mex. One of these was completely leveled off in Middle Formative times to serve as the base for a site that was to become the Zapotec people's most important capital. Prior to that time, the Early Formative ancestral Zapotec had lived in scattered villages and at least one centre of some importance, San José Mogote. San José Mogote shows evidence of Olmec trade and contacts dating to the time of San Lorenzo.

At Monte Albán, during the earliest, or Monte Albán I, epoch of that site's history, a peculiar group of reliefs was carved on stone slabs and affixed to the front of a rubble-faced platform mound and around a contiguous court. The reliefs are usually called *danzantes*, a name derived from the notion that they represent human figures in dance postures. Actually, almost all of the *danzante* sculptures show Olmecoid men in strange, rubbery postures as though they were swimming in honey. From their open mouths and closed eyes, it is assumed that they are meant to represent dead persons. On many *danzantes* one or more unreadable hieroglyphs appear near the heads of the figures, most likely standing for the names of the sacrificed lords of groups beaten in combat by the Zapotec. Several slabs also bear calendrical notations, and it can be stated that the Middle Formative elite of Monte Albán were the first in Meso-America to develop writing and the calendar (at least in written form).

The *danzantes* figures

The Valley of Mexico in the Middle Formative. The cultures of central Mexico tended to lag behind those of southern Mexico in the development of political and religious complexity. The presence of Olmec figurines and ceramics in Early Formative burials in the Valley of Mexico has been noted, but the local communities of that time were of a modest village sort, as were those of the succeeding Middle Formative. On the western shores of the great lake filling the Valley of Mexico, for instance, remains of several simple villages have been uncovered that must have been not unlike small settlements that can be found in the Mexican hinterland today. The people who lived at El Arbolillo and Zacatenco had simply terraced off village refuse to make platforms on which their pole-and-thatch houses were built. Metates and manos are plentiful; pottery is relatively plain—featuring the abundant hard, white-slipped ware of the Middle Formative—and small female figurines are present by the thousands. Subsistence was based upon corn farming and upon hunting. In some Middle Formative sites, however, such as Tlatilco, there is evidence of Olmec influence, as in the previous Early Formative Period. There are also indications that ceremonial pyramid construction began in the latter part of the Middle Formative at Cuicuilco, a site in the southern part of the valley, which was to become a major centre in the succeeding Late Formative Period.

The Maya in the Middle Formative. In the Maya highlands, the key archaeological region has always been the broad, fertile Valley of Guatemala around present-day Guatemala City. The earliest occupation is known as the Arevalo phase, a village culture of the Early to Middle Formative. It was followed by Las Charcas, a Middle Formative culture known largely from the contents of bottle-shaped pits found dug into the subsoil on the western edge of the modern city. Extremely fine ceramics have been excavated from them, including red-on-white bowls with animal figures, effigy vessels, three-footed cups, and peculiar three-pronged incense burners. Solid female figurines are also present.

Las Charcas

The earliest Middle Formative cultures of the Maya lowlands are called, collectively, the Xe horizon. They apparently developed from antecedent Early Formative cultures of the Maya lowlands that have been discussed above. The problem of the origin of the Mayan-speaking people has not been solved. It may be that they were Olmec people who had been forced out of their homeland to the west by the collapse of San Lorenzo. There were already peoples in the Maya lowlands in Early Formative times, however, and if the early Maya were Olmec, they brought little of their Olmec culture with them. Another hypothesis is that the earliest Maya descended to their lowland homelands from the Guatemalan highlands.

In the Maya lowlands the Mamom cultures developed out of those of Xe times. Mamom shares many similarities with the highland Maya at Las Charcas: pottery is almost entirely monochrome—red, orange, black, and white—and figurines are female with the usual punched and appliquéd embellishments. Toward the end of the Middle Formative, or after about 600 BC, Mamom peoples began

building small ceremonial centres and modest-sized pyramidal platforms.

LATE FORMATIVE PERIOD (300 BC–AD 100)

Probably the most significant features of the Late Formative are: (1) the transformation of Olmec civilization in southeastern Meso-America into something approaching the earliest lowland Maya civilization and (2) the abrupt appearance, toward the end of the Late Formative, of fully urban culture at Teotihuacán in the Valley of Mexico. Most of the distinctive cultures that were to become the great Classic civilizations began to take shape at this time. There was no unifying force in the Late Formative comparable to the earlier Olmec; rather, regionalism and local cultural integration were the rule. There were, however, horizon traits, particularly in pottery, that were almost universal. Ceramics became elaborate in shape, often with composite or recurved outlines, hollow, bulbous feet, and flangelike protrusions encircling the vessel. The use of slips of a number of different colours as pottery decoration at times approached the elaborate polychromes of Classic times.

Horizon traits

The idea of constructing temple-pyramids was probably also a general trait. It was a Meso-American custom to bury a dead person beneath the floor of his own house, which was often then abandoned by the bereaved. As an elite class of noble lineages became distinguished from the mass of the people, the simple house platforms serving as sepulchres might have become transformed into more imposing structures, ending in the huge pyramids of the Late Formative and Classic, which surely had funerary functions. The deceased leader or the gods from which he claimed descent, or both, would then have been worshiped in a "house of god" on the temple summit. These pyramids became the focal point of Meso-American ceremonial life, as well as the centres of settlement.

Valley of Mexico. The Cuicuilco-Ticomán culture succeeded the Middle Formative villages of the valley but retained many of their traits, such as the manufacture of solid handmade figurines. Of considerable interest is the type site of Cuicuilco, located on the southwestern edge of the valley. Lava from a nearby volcano covers all of Cuicuilco, including the lower part of the round "pyramid" for which it is best known. Ceramic analysis and radiocarbon dating have proved that the flow occurred at about the time of Christ. Rising up in four tiers, the Cuicuilco pyramid has a clay-and-rubble core faced with broken lava blocks. The summit was reached by ramps on two sides. Circular temples were traditionally dedicated in Meso-America to Quetzalcóatl, the Feathered Serpent, and he may have been the presiding deity of Cuicuilco.

In the Valley of Teotihuacán, a kind of side pocket on the northeastern margin of the Valley of Mexico, Cuicuilco-Ticomán culture eventually took on a remarkable outline, for there is evidence that by the beginning of the Christian Era a great city had been planned. There is little doubt that by the Proto-Classic stage (AD 100–300) it had become the New World's first urban civilization (see below *Teotihuacán*).

Valley of Oaxaca. Occupation of the Monte Albán site continued uninterrupted, but ceramic evidence for Monte Albán II culture indicates that cultural influences from southeastern Mexico were reaching the Zapotec people. On the southern end of the site's main plaza is a remarkable stone structure called Building J, shaped like an arrow pointing southwest and honeycombed with galleries. Some believe it to have been an astronomical observatory. Incised slabs are fixed to its exterior; these include some older *danzantes* as well as depictions of Zapotec place glyphs from which are suspended the inverted heads of dead chiefs—surely again the vanquished enemies of Monte Albán. Dates are given in the 52-year Calendar Round, with coefficients for days and months expressed by bar-and-dot numerals, a system that is first known for Monte Albán I and that became characteristic of the Classic Maya. Throughout its long Formative and Classic occupation, the dominant ware of Monte Albán is a fine gray pottery, elaborated in Monte Albán II into the usual Late Formative shapes.

Veracruz and Chiapas. La Venta suffered the fate of San Lorenzo, having been destroyed by violence around 400 BC. Olmec civilization subsequently disappeared or was transformed into one or more of the cultures of the southeastern lowlands.

One centre that retained a strong Olmec tradition, however, was Tres Zapotes, near the Tuxtla Mountains in the old Olmec "heartland." Its most famous monument, the fragmentary Stela C, is clearly epi-Olmec on the basis of a jaguar-monster mask carved in relief on its obverse. On the reverse is a column of numerals in the bar-and-dot system, which was read by its discoverer, Matthew W. Stirling, as a date in the Maya calendar corresponding to 31 BC; this is more than a century earlier than any known dated inscription from the Maya area itself. Thus, it is highly probable that this calendrical system, formerly thought to be a Maya invention, was developed in the Late Formative by epi-Olmec peoples living outside the Maya area proper.

Tres Zapotes

Izapan civilization. Izapa, type site of the Izapan civilization, is a huge temple centre near modern Tapachula, Chiapas, on the hot Pacific coast plain. Its approximately 80 pyramidal mounds were built from earth and clay faced with river boulders. A large number of carved stone stelae have been found at Izapa, almost all of which date to the Late Formative and Proto-Classic. Typically, in front of each stela is a round altar, often crudely shaped like a toad.

These stelae are of extraordinary interest, for they contain a wealth of information on Late Formative religious concepts prevalent on the border of the Maya area. Izapan stelae are carved in relief with narrative scenes derived from mythology and legend; among the depictions are warfare and decapitation, ceremonies connected with the sacred world tree, and meetings of what seem to be tribal elders. Many deities are shown, each of which seems derived from an Olmec prototype.

Sites with Izapan-style sculpture are distributed in a broad arc extending from Tres Zapotes in the former Olmec region, across the Isthmus of Tehuantepec into coastal Chiapas and Guatemala, and up into the Guatemalan highlands. Izapan civilization is clearly the intermediary between Olmec and Classic Maya in time and in cultural content, for the following early Maya traits are foreshadowed by it: (1) the stela–altar complex, (2) long-lipped deities, (3) hieroglyphic writing and Long Count dates on some monuments, (4) such iconographic elements as a U-shaped motif, and (5) a cluttered, baroque, and painterly relief style that emphasizes narrative. An important site pertaining to this Izapan culture is Abaj Takalik, on the Pacific slopes of Guatemala, to the east of Izapa. Three sculptural styles are represented there: Olmec or Olmecoid, Izapan, and Classic Maya. Among the latter is one stela with a date read as AD 126, earlier than any monuments discovered in the Maya lowlands.

Perhaps it was not Izapa itself but the great site of Kaminaljuyú, on the western edge of Guatemala City, that transmitted the torch of Izapan civilization to the lowland Maya. This centre once consisted of more than 200 earth and clay mounds, most of which have been destroyed. The major occupation is ascribed to the Miraflores phase, the Late Formative culture of the Valley of Guatemala. Some of these huge Miraflores mounds contained log tombs of incredible richness. In one, the deceased lord was accompanied by sacrificed followers or captives. As many as 340 objects were placed with him, including jade mosaic masks, jade ear spools and necklaces, bowls of chlorite schist, and pottery vessels of great beauty. Also present in the tombs are peculiar "mushroom stones," which may actually have been used in rites connected with hallucinogenic mushrooms.

Kaminaljuyú

The earliest Maya civilization of the lowlands. By the Late Formative, the lowland Maya had begun to shape a civilization that was to become the greatest in the New World. The Petén-Yucatán Peninsula lacks many raw materials and has a relatively low agricultural potential. But what it does have in limitless quantities is readily quarried limestone for building purposes and flint for stonework. Cement and plaster could easily be produced by burning limestone or shells.

The heart of the Maya civilization was always northern Petén, in Guatemala, where the oldest dated Maya stelae are found, although this presents something of a problem in cultural-historical interpretation, since the earliest prototypes for these stelae—as mentioned above—have been found in Pacific-littoral and highland Guatemala. The Late Formative culture of Petén is called Chicanel, evidence of which has been found at many Maya centres. Chicanel pottery includes dishes with wide-everted and grooved rims, bowls with composite silhouette, and vessels resembling ice buckets. Figurines are curiously absent.

Early architecture

Architecture was already quite advanced and had taken a form peculiar to the Maya. Temple platforms were built by facing a cemented-rubble core with thick layers of plaster. At the site of Uaxactún, Structure E-VII-sub affords a good idea of a Chicanel temple-platform. It is a four-sided, stucco-covered, stepped pyramid with pairs of stylized god masks flanking stairways on each side. On its summit was a thatched-roof temple. At Tikal, the giant among Maya ceremonial centres, the so-called Acropolis was begun in Chicanel times, and there was a great use of white-stuccoed platforms and stairways, with flanking polychromed masks as at Uaxactún. Most importantly, there is evidence from Tikal that the Maya architects were already building masonry superstructures with the corbel vault principle; *i.e.*, with archlike structures the sides of which extend progressively inward until they meet at the top. The large sizes of Chicanel populations and the degree of political centralization that existed by this time are further attested to by the discovery in the 20th century of the huge site of El Mirador, in the extreme northern part of Petén. The mass of El Mirador construction dwarfs even that of Tikal, although El Mirador was only substantially occupied through the Chicanel phase.

Chicanel-like civilization is also known in Yucatán, where some temple pyramids of enormous size are datable to the Late Formative. An outstanding site is the cave of Loltún in Yucatán, where a relief figure of a standing leader in pure Izapan style is accompanied by a number of unreadable hieroglyphs as well as a notation in the 260-day count. This inscription raises the question of writing and the calendar among the lowland Maya in the Late Formative. On the whole, the evidence is negative and suggests that several important intellectual innovations considered to be typically Mayan were developed beyond the Maya area proper and appeared there only at the close of the Formative. Izapan civilization appears to have played the crucial role in this evolutionary process.

EARLY CLASSIC PERIOD (AD 100–600)

Definition of the Classic. In the study of the Classic stage, there has been a strong bias in favour of the Maya; this is not surprising in view of the fact that the Maya have been studied far longer than any other people in Meso-America. But the concept of a "Classic" period is a case of the Maya tail wagging the Meso-American dog, since the usual span given to that stage—AD 250–900—is the period during which the Maya were erecting dated stone monuments. This brackets the Maya apogee, but for most areas of non-Maya Meso-America only the first half of the period may be accurately called a "golden age." While the famous and yet mysterious Maya collapse took place at about AD 900, in many other regions this downfall occurred almost three centuries earlier.

Qualitatively, there is little to differentiate the Classic from the Late Formative that preceded it. Various tendencies that were crystallizing in the last centuries before the Christian era reached fulfillment in the Classic. Two cultures stand out beyond all others. One is that of Teotihuacán, which during the Early Classic played a role in Meso-America similar to that which Olmec had performed in the Early Formative. The second is the lowland Maya civilization, which during its six centuries of almost unbroken evolution in the humid forests reached cultural heights never achieved before or since by New World natives. The contrast between the two—one urban and expansionist, the other less urban and non-expansionist—exemplifies well the cultural results of the ecological possibilities offered by highland and lowland Meso-America.

Teotihuacán. Teotihuacán, which was located in the Valley of Teotihuacán, a pocketlike extension of the Valley of Mexico on its northeastern side, was probably the largest city of the New World before the arrival of the Spaniards. At its height, toward the close of the 6th century AD, it covered about eight square miles and may have housed more than 150,000 inhabitants. The city was divided into quarters by two great avenues that crosscut each other at right angles, and the entire city was laid out on a grid plan oriented to these avenues. The Avenue of the Dead, the main north–south artery of the city, is aligned to a point 16° east of true north, which may have had astrological meaning.

Layout of the city

Because irrigation plays some part in the present-day agricultural economy of the Valley of Teotihuacán, it has been suggested that the Early Classic city also was based upon this subsistence system. It is almost inconceivable, however, that a city of such proportions could have relied upon the food production of its own valley or even upon the Valley of Mexico, whether irrigated or not.

Planning and construction of Teotihuacán began, according to radiocarbon dates, about the beginning of the Christian Era, in the Tzacualli phase. At this time, the major avenues were laid out and construction of the major ceremonial structures along the Avenue of the Dead began. Figurines and potsherds extracted from fill inside the 200-foot-high Pyramid of the Sun, the most prominent feature of Teotihuacán, prove that this was erected by the end of the Tzacualli phase. The pyramid rises in four great stages, but there is a fifth and much smaller stage between the third and fourth. An impressive stairway rises dramatically on its west side, facing the Avenue of the Dead. Reexamination suggested the presence of a huge tomb at its base, but this has never been excavated.

On the northern end of the Avenue of the Dead is the Pyramid of the Moon, very similar to that of the Sun, but with an additional platform-temple jutting out on the south. This exhibits the *talud-tablero* architectural motif that is typical of Teotihuacán culture: on each body or tier of a stepped pyramid is a rectangular frontal panel (*tablero*) supported by a sloping batter (*talud*). The *tablero* is surrounded by a kind of projecting frame, and the recessed portion of the panel usually bears a polychrome mural applied to the stuccoed surface.

Near the exact centre of the city and just east of the Avenue of the Dead is the Ciudadela ("Citadel"), a kind of sunken court surrounded on all four sides by platforms supporting temples. In the middle of the sunken plaza is the so-called Temple of Quetzalcóatl, which is dated to the second phase of Teotihuacán, Miccaotli. Along the balustrades of its frontal stairway and undulating along the *talud-tablero* bodies of each stage of this stepped pyramid are sculptured representations of Quetzalcóatl, the Feathered Serpent. Alternating with the Feathered Serpents on the *tableros* are heads of another monster that can be identified with the Fire Serpent—bearer of the Sun on its diurnal journey across the sky.

The Temple of Quetzalcóatl

On either side of the Avenue of the Dead are residential palace compounds (probably occupied by noble families), which also conform to the Teotihuacán master plan. Each is a square, 200 feet on a side, and is surrounded by a wall. The pedestrian would have seen only the high walls facing the streets, pierced by inconspicuous doors. Within the compounds, however, luxury was the rule. Roofs were flat, constructed of large cedar beams overlaid by brush and mortar. Interior walls were plastered and magnificently painted with ritual processions of gods and various mythological narratives. Interconnected apartments were arranged around a large, central, open-air court.

These dwellings were the residences of Teotihuacán's elite. Toward the periphery of Teotihuacán, however, the social situation may have been quite different. One excavation on the eastern side of the city disclosed a mazelike complex of much tinier and shoddier apartments that recall the poorer sections of Middle Eastern cities. It may be guessed that there lay the crowded dwellings of the artisans and other labourers who made the city what it was. There is also evidence that certain peripheral sections were reserved for foreigners.

Manufacturing and trade

Teotihuacán must have been the major manufacturing centre of the Early Classic, for the products of its craftsmen were spread over much of Meso-America. The pottery, particularly during the Xolalpan phase, which represents the culmination of Teotihuacán as a city and empire, is highly distinctive. The hallmark of the city is the cylindrical vessel with three slab legs and cover, often stuccoed and then painted with scenes almost identical to those on the walls of buildings. There are also vessels shaped like modern flower vases and cream pitchers. Thin Orange ware is a special ceramic type produced to Teotihuacán specifications, perhaps in southern Veracruz, and exported by its own traders. Figurines were produced by the tens of thousands in pottery molds.

Among its many commercial specializations, obsidian was probably preeminent, for the Teotihuacanos had gained control of the mines of green obsidian above the present-day city of Pachuca, in Hidalgo. They also had a local but poorer quality source. Millions of obsidian blades, as well as knives, dart points, and scrapers, were turned out by Teotihuacán workshops for export.

The name Teotihuacán meant "City of the Gods" (or, "Where Men Became Gods") in Aztec times, and although the city had been largely deserted since its decline, the Aztec royal house made annual pilgrimages to the site. Teotihuacán culture exerted a profound influence on all contemporary and later Meso-American cultures. Many Aztec gods, such as Tlaloc, his consort Chalchiuhtlicue, and Quetzalcóatl, were worshiped by the Teotihuacanos. Like the Aztec, the Teotihuacanos generally cremated their dead. In fact, there are so many congruences between Teotihuacán practices and those of the later Toltec and Aztec that some authorities believe them to have been speakers of Náhuatl language and the precursors of those people. Some linguistic authorities, however, believe that the Teotihuacanos spoke a Totonac language, similar to what was spoken by the inhabitants of central Veracruz. It is not known whether the people of the city, like the Maya, were literate.

Teotihuacán was the greatest city of Meso-America, indeed, of all pre-Columbian America. Authorities are divided as to whether it was the capital of a great political empire. Some believe that Teotihuacán's expansion was carried by force of arms; others believe its power to have been largely economic and religious. In either case, at its height in the 6th century Teotihuacán was the greatest civilization in Meso-America, with an influence that far outstripped that of the later Aztec empire. For the archaeologist, the universal spread of Teotihuacán ceramic and other traits constitutes an Early Classic horizon.

Cholula. The broad, fertile plains surrounding the colonial city of Puebla, to the southeast of the snowcapped volcanoes that border the Valley of Mexico, were from very ancient times an important centre of pre-Hispanic population. The modern traveler, approaching the city, sees to its west, in the distance, what looks like a sizable hill rising from the plain. This is actually the pyramid of Cholula, the largest single structure in Mexico before the Spanish conquest.

Size and decoration of the Cholula pyramid

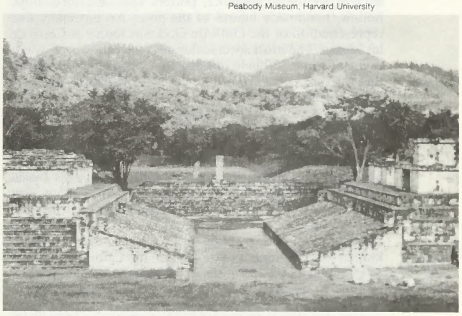
Archaeological exploration of the Cholula pyramid has shown that it was built from adobe in four great construction stages. In its final form, the pyramid measured 1,083 feet by 1,034 feet at the base and was about 82 feet high. By Late Postclassic times the pyramid had been abandoned for so long that the Spaniards who subdued (and massacred) the residents of Cholula considered it a natural prominence. All four superimposed structures within the pyramid were carried out according to strict Teotihuacán architectural ideas. The earliest structure, for instance, has the usual *talud-tablero* motif, with stylized insectlike figures painted in black, yellow, and red appearing in the *tableros.* Similar decoration, also in Teotihuacán style, is to be found in the later structures.

Great quantities of ceramics and pottery figurines have been recovered from the excavations, and these demonstrate a near archaeological identity between Early Classic Teotihuacán and Cholula. Because of the staggering size and importance of its pyramid, it has been suggested that Cholula was some kind of sister city to Teotihuacán. Cholula was surely part of the Teotihuacán cultural sphere and may well have participated in the administration of its empire. Excavations at the base of the pyramid have produced a previously unsuspected cultural element. Several enormous slabs were uncovered, two of which were a kind of altar, while the third was set upright as a stela. All are rectangular but with borders carved in low relief in the complex interlace motif that is the hallmark of the Classic Central Veracruz style.

Classic Central Veracruz. The Meso-American ball game was played throughout the area and still survives in attenuated form in northwestern Mexico. On the eve of the conquest, games took place in a rectangular court bordered on the long sides by walls with both sloping and vertical rebound surfaces. There were two teams, each composed of a small number of players. The ball was of solid rubber, of substantial size, and traveled with considerable speed around the court. It could not be hit or touched with the open hands or with the feet; most times the player tried to strike it with the hip. Consequently, fairly heavy protective padding was necessary to avoid injuries, which in some cases were fatal. Leather padding was worn over the hips, and pads were placed on the elbows and knees. A heavy belt was tied around the waist built up from wood and leather, while in some parts of the Maya region and in Late Formative Oaxaca, gloves and something resembling jousting helmets were worn.

The Meso-American ball game

Peabody Museum, Harvard University



The ball court at Copán, Honduras; Late Classic Maya.

The Classic Central Veracruz style is almost purely devoted to the paraphernalia of the ball game and to the ball courts themselves. At the site of El Tajín, which persisted through the end of the Late Classic, elaborate reliefs on the walls of the courts furnish details on how this equipment was used. *Yugos* ("yokes") were the stone counterparts of the heavy protective belts. During the post-game ceremonies, which may have featured the sacrifice of the captain and other players on the losing side, these U-shaped objects were worn about the waists of the participants. On the front of the *yugo* was placed an upright stone object that may originally have functioned as a ball-court marker and that took two forms: *hachas* ("axes") or thin stone heads, and *palmas* ("palms"). All are carved in an elaborate low-relief style in which life forms are enmeshed in undulating and interlaced scroll designs with raised borders. All of these items, and the style itself, may have evolved out of late Olmec art on the Gulf coast.

Very often the *yugos* represent the marine toad, a huge amphibian with swollen poison glands on the head; in its jaws is a human head. The earliest *hachas,* which were characteristically notched to fit on the *yugos,* were quite thick human heads and may well date to the Late Formative or Proto-Classic. In time, these become very thin and represent human heads wearing animal headgear. *Palmas* are paddle-shaped stone objects with trilobed bases and exhibit a much richer subject matter than either *hachas* or *yugos,* quite often illustrating brutal scenes of sacrifice and death, two concepts that were closely associated with the ball game on the Gulf coast.

Despite the definite presence of the style at Teotihuacán and Cholula, Classic Central Veracruz is focused upon north central Veracruz, where the type site of El Tajín is located, and contiguous parts of Puebla. Today, this region is dominated by speakers of Totonac, a distant relative of Mayan, and the Totonac themselves claim that they built El Tajín. Whether or not Classic Central Veracruz culture was a Totonac achievement, the style persisted through the Classic period and strongly influenced developments in distant regions.

Southern Veracruz. On the southern Gulf coast plain, Olmec traditions seemed to have lasted into the Early Classic and merged with Teotihuacán artistic canons to produce new kinds of art. Cerro de las Mesas, lying in the plains of the Papaloápan River not far from the coast, is one of these hybrid sites. Dozens of earthen mounds are scattered over the surface in a seemingly haphazard manner, and the archaeological sequence is long and complex. The site reached its apogee in the Early Classic, when the stone monuments for which it is best known were carved. Most important are a number of stelae, some of which are carved in a low-relief style recalling Late Formative Tres Zapotes, early lowland Maya, and Cotzumalhuapa (on the Pacific coast of Guatemala).

Cerro de las Mesas pottery

Cerro de las Mesas pottery, deposited in rich burial offerings of the Early Classic, is highly Teotihuacanoid, with slab-legged tripods predominating. At this and other sites in southern Veracruz, potters also fashioned large, hollow, handmade figures of the gods. An especially fine representation of the Old Fire God was found at Cerro de las Mesas. The most spectacular discovery, however, was a cache of some 800 jade objects. Many of the specimens in this treasure trove are of Olmec workmanship, obviously heirlooms from the much earlier Olmec civilization, while some are clearly Early Classic Maya.

The entire coastal plain from Cerro de las Mesas north to the borders of Classic Central Veracruz culture is famed for Remojadas-style pottery figurines, which must have been turned out in incredible quantity for use as burial goods. The Remojadas tradition dates to the Late Formative and lasts until the Early Postclassic. Figurines are hollow and largely mold-made in the Late Classic, while they were fashioned by hand in the Early Classic. The best-known Classic representations are the "smiling figures" of grinning boys and girls wearing loincloths, skirts, or nothing at all. All kinds of genre scenes are represented, including even lovers in swings, as well as more grim activities such as the heart sacrifice of victims tied down in what look like beds.

Classic Monte Albán. The cultural phases designated as Monte Albán III-A and III-B mark the Classic occupation of this major site in the Valley of Oaxaca. There can be little doubt that the people of Monte Albán were Zapotec speakers, who during Classic times had unequaled opportunity to develop their civilization unaffected by the major troubles that disturbed Teotihuacán and the Maya at the close of the Early Classic. Instead of the 18 or 19 sites known for the valley during the Late Formative, there now were more than 200, a testimony to Zapotec prosperity.

Zapotec prosperity

The Monte Albán Classic Period (III-A and III-B) lasted from AD 250 to 700. During the earlier (III-A) part of the period (250–450) the site shows considerable influence from Teotihuacán. The Early Postclassic Period at Monte Albán (IV; 700–1000) was a time of significant cultural change; it is still uncertain, however, whether the Mixtec replaced the Zapotec at that time.

The Classic site of Monte Albán is quite spectacular. Stone-faced platforms are fronted by stairways with flanking balustrades and exhibit a close counterpart of the *talud-tablero* motif of Teotihuacán. The temple superstructures had colonnaded doorways and flat beam-and-mortar roofs. One of the best-preserved ball courts of Meso-America can be seen at Monte Albán, with a ground plan fashioned in the form of a capital I. Spectators watched the game from stone grandstands above the sloping playing surfaces.

Subsurface tombs were dug in many parts of the site as the last resting places of Monte Albán's elite. The finest are actually miniature replicas of the larger temples on the surface, complete with facade and miniature painted rooms. The style of the funerary wall paintings is quite close to Teotihuacán, in which areas of flat colour are contained within very finely painted lines in red or black. Teotihuacán presence can also be seen in the finer pottery of Classic Monte Albán, but the manufacture is local as can be proved from the predominance of the fine gray ware that has always typified Monte Albán.

The tradition of literacy dates to Monte Albán I. By Classic times, inscriptions are abundant, appearing on stelae, lintels, slabs used as doors, and wall paintings. The 52-year Calendar Round was the only form of writing dates. The subject matter of these inscriptions can be related to the scenes that they accompany: quite often it is a bound captive standing on a place-glyph, presumably an enemy leader taken in war—an old Monte Albán preoccupation.

The Zapotec of Monte Albán, like the Maya, never exerted much cultural or other pressure on peoples beyond their lands. They did, however, control lands from the Tehuacán Valley in Puebla as far south as the Pacific shore of Oaxaca. Whether they themselves were also controlled by Teotihuacán has not been demonstrated.

The Maya highlands and Pacific coast. Little is known about the Guatemalan highlands between the demise of the Late Formative Miraflores culture and the onset of the Early Classic. But at the ancient site of Kaminaljuyú, on the western side of Guatemala City, a group of invaders from Teotihuacán built a miniature replica of their capital city. This happened about AD 400, when Teotihuacán was at the height of its power.

This implanted Teotihuacán culture is called Esperanza. Mexican architects must have accompanied the elite, for Kaminaljuyú structures copy the older prototypes down to the last detail, including the support of the lower moldings around *tableros* with slate slabs. The abundant volcanic building stone, however, so freely used at Teotihuacán, was not present, so that Esperanza temple platforms are built from clay instead.

Each temple platform was rebuilt several times, the later structures being raised over the earlier. Within the stairways fronting each successive platform a great leader was buried. The rich burial furniture in the tombs is informative, for it included three classes of goods: (1) items such as Thin Orange pottery manufactured in Teotihuacán or in one of its satellite areas, (2) hybrid Teotihuacán-Maya pottery and other objects, probably made in Kaminaljuyú, and (3) pottery imported from Petén and of Early Classic Maya manufacture. Also discovered in one tomb was a slate mirror carved in Classic Central Veracruz style. Jade objects occur in abundance in the Esperanza tombs, and in one structure an enormous boulder was recovered; it had been imported from the Maya source along the Motagua River in the southeastern lowlands. The Esperanza elite were enormously wealthy.

What were they doing in the Maya highlands in the first place? Were they an army of imperial conquest? Or were their interests more in the realm of trade? Or both? It is not possible to be definite in these interpretations; but it is known that among the Aztec of the Late Postclassic there was an institution called the *pochteca*, a hereditary guild of armed merchants who traveled into distant lands looking for luxury goods to bring back to the royal house. Quite often the *pochteca* would seize lands of hostile peoples through which they passed, or they would provoke incidents that led to the intervention of the regular Aztec army.

The *pochteca*

It has been suggested that the Teotihuacanos in Kaminaljuyú were also *pochteca*. They had clear access to the Petén-Yucatán Peninsula and may have exercised political control over it. Kaminaljuyú may have been one of their principal bases of operations in the inclusion of the Maya, both highland and lowland, within the Teotihuacán state.

Within a zone only 75 miles long and 30 miles wide, on the Pacific coast plain of Guatemala, is a cluster of nine compactly built ceremonial centres that together form the Cotzumalhuapa civilization. It forms a puzzle, for there are strong affiliations with most other contemporary civilizations in Meso-America. Stylistic influence from the lowland Maya, Classic Central Veracruz, and Teotihuacán can be detected among others. While Cotzumalhuapa

took form by the Early Classic, it continued into the Late Classic; but there are great problems in dating individual sculptures.

The problem of Cotzumalhuapa has been linked with that of the Pipil, a shadowy people living in the same region on the eve of the Spanish conquest, who spoke Nahua rather than Maya. It is possible that these Classic sites were actually Pipil capitals, but the case cannot be proved. There is some hieroglyphic writing on Cotzumalhuapa sculptures, mainly dates within what seems to be a 52-year Calendar Round, the glyphs for days being Mexican rather than Maya. There are no real texts, then, to help with the problem.

Classic civilization in the Maya lowlands: Tzakol phase. Archaeologists have divided the entire area occupied by speakers of Mayan languages into three subregions: (1) the Southern Subregion, essentially the highlands and Pacific Coast of Guatemala, (2) the Central Subregion, which includes the department of Petén in northern Guatemala and the immediately adjacent lowlands to the east and west, and (3) the Northern Subregion, consisting of the Yucatán Peninsula north of Petén proper. Between 250 and 900 the most brilliant civilization ever seen in the New World flourished in the forested lowlands of the Central and Northern subregions.

Lowland Maya civilization falls into two chronological phases or cultures: Tzakol culture, which is Early Classic and began shortly before AD 250, and the Late Classic Tepeu culture, which saw the full florescence of Maya achievements. Tepeu culture began about 600 and ended with the final downfall and abandonment of the Central Subregion about 900. (These dates, based on the correlation of the Long Count system of the Maya calendar with the Gregorian calendar, are the most generally accepted; but there is a slight chance that a rival correlation espoused by the American archaeologist Herbert J. Spinden may be correct, which would make these dates 260 years earlier.)

One of the earliest objects inscribed with the fully developed Maya calendar is the Leiden Plate, a jade plaque, now housed in the National Museum of Ethnology, Leiden, Neth., depicting a richly arrayed Maya lord trampling a captive underfoot. On its reverse side is a Long Count date corresponding to 320. Although it was found in a very late site on the Caribbean coast, stylistic evidence suggests that the Leiden Plate was made at Tikal, in the heart of northern Petén. In the mid-20th century the University of Pennsylvania's ambitious field program at the Tikal site produced Stela 29, erected 28 years before, in 292. Both objects and, in fact, almost all early Tzakol monuments draw heavily upon a heritage from the older Izapan civilization of the Late Formative, with its highly baroque, narrative stylistic content.

Because of the Maya penchant for covering older structures with later ones, Tzakol remains in the Central Subregion have to be laboriously dug out from their towering Late Classic overburdens. Nevertheless, it is clear that at sites like Tikal, Uaxactún, and Holmul, Maya civilization had reached something close to its final form. Enormous ceremonial centres were crowded with masonry temples and "palaces" facing onto spacious plazas covered with white stucco. The use of the corbel vault for spanning rooms—a trait unique to the lowland Maya—was by this time universal. Stelae and altars (a legacy from Izapa) are carved with dates and embellished with the figures of men and perhaps gods. Polychrome pottery, the finest examples of which were sealed in the tombs of honoured personages, emphasizes stylized designs of cranes, flying parrots, gods, and men. These often occur on bowls with a kind of apron or basal flange encircling the lower vessel. Along with these purely Maya ceramics are vessels that show the imprint of distant Teotihuacán: the cylindrical vase supported by three slab legs, the "cream pitcher," and the *florero* ("flower vase").

Ceremonial centres

Wall painting had already reached a high degree of perfection in the Central Subregion, as attested by an extremely fine mural at Uaxactún depicting a palace scene in which two important lords confer with each other. This mural art is quite different from that of Teotihuacán, being very naturalistic instead of formal and including a definite interest in portraiture. Nonetheless, excavations in Petén sites have shown that Teotihuacán influence was quite pervasive. From Tikal, for example, comes Stela 31, depicting a richly garbed Maya lord, festooned with jade ornaments, standing between two warriors from Teotihuacán. These foreigners carry shields that bear the visage of the Teotihuacán rain god, Tlaloc. It is certain that there was a three-way trading relationship between Tikal, Kaminaljuyú, and Teotihuacán in Early Classic times.

Thus, the Teotihuacán involvement with Tikal and the Central Subregion may have taken, as at Kaminaljuyú, the form of *pochteca* trading colonies that exerted some control over the lowland Maya. The lord on Stela 31 may have been a puppet ruler manipulated by tough merchant-warriors. Teotihuacán as a city and capital of an empire began to weaken toward the close of the 6th century. It could therefore be expected that the disruptions that effectively ended the life of the great Mexican capital would be reflected in the Maya area. This is exactly the case. In the Guatemalan highlands, Kaminaljuyú declined rapidly after AD 600, and the entire Southern Subregion was to play little part in Maya culture until the Late Postclassic. The lowland Maya suffered some temporary reverses; few stelae were erected between 534 and 692, and there is evidence that existing monuments were mutilated.

LATE CLASSIC NON-MAYA MESO-AMERICA (600–900)

The cultural situation in Late Classic Meso-America is the reverse of that prevailing in the Early Classic: Central Mexico now played only a minor role, while the lowland Maya reached their intellectual and artistic heights. In contrast to the old Teotihuacanos, however, the Maya were not expansionistic. It is true that Maya cultural influence has been detected along the Gulf coast and in the states of Morelos and Tlaxcala—as in the painted murals of Cacaxtla in the latter state—but it is unlikely that this was the result of a military takeover. The outcome of this state of affairs, with no one people powerful enough or sufficiently interested in dominating others, was a political and cultural fragmentation of Meso-America after 600. It was not until the great Toltec invasions of the Early Postclassic that anything approaching an empire was to be seen again.

The decline in fortunes of the Valley of Mexico, and especially of Teotihuacán, cannot now be explained. Climatic deterioration, resulting in drier conditions and thus a diminished subsistence potential, may have been a factor.

Decline of Teotihuacán

Nevertheless, Teotihuacán was never completely abandoned, even though its great palaces had been burned to the ground and its major temples abandoned. People continued to live in some sections, but their houses were mere hovels compared to the dwellings of the Early Classic. In general, the Valley of Mexico was a cultural and political vacuum in Late Classic times.

One of the very few centres of the Late Classic in central Mexico that amounted to much was Xochicalco, in Morelos. Strategically located on top of a hill that was completely reworked with artificial terraces and ramparts, Xochicalco was obviously highly defensible, an indication of the unsettled times then prevailing in central Mexico. The site shows a bewildering variety of cultural influences, particularly Maya. The principal structure of Xochicalco is a temple substructure of masonry that is completely carved in relief with undulating Feathered Serpents, indicating that it was dedicated to the cult of Quetzalcóatl. All indications are that Xochicalco was a cosmopolitan and very powerful centre, perhaps the most influential west of Veracruz and northwest of the Maya area. It was literate and civilized at a time when most other parts of central Mexico were in cultural eclipse.

The Late Classic occupation of Oaxaca, especially of the Valley of Oaxaca, is designated as Monte Albán III-B (450–700). The Mixtec invasions of the valley probably began in earnest around 900. The Mixtec occupied the hilly, northern part of Oaxaca; their records, which extend to the 7th century, show them to have been organized into a series of petty states headed by aggressive, warlike kings. By the Postclassic, they had become the dominant force throughout Oaxaca and in part of Puebla.

The tendencies in central Veracruz art and architecture that began in the Late Formative culminated in the Late Classic at the great centre of El Tajín, placed among jungle-covered hills in a region occupied by the Totonac Indians, whose capital this may well have been. Its most imposing structure is the Pyramid of the Niches, named for the approximately 365 recesses on its four sides. In this and other buildings at El Tajín, the dominant architectural motif is the step-and-fret. There are a number of other temple pyramids at the site, as well as palacelike buildings with flat, concrete roofs, a tour de force of Meso-American engineering knowledge. El Tajín's three major ball courts are remarkably important for the reliefs carved on their vertical playing surfaces, for these give valuable information on the religious connotations of the sacred game. Like Xochicalco, El Tajín was in some way linked to the destiny of the lowland Maya, and the collapse of Maya civilization around 900 may have been reflected in the demise of the Veracruz centre.

The Pyramid of the Niches

Further down the Gulf coast plain, the Remojadas tradition of hollow pottery figurines continued to be active in the Late Classic, with a particularly large production of the mysterious smiling figures of dancing boys and girls, which were intended as funerary offerings. But in addition, there was a great deal of pottery and figurines that were fashioned under very strong Maya influence. In fact, much of southern Veracruz at this time was a cultural extension of the lowland Maya. There is no indication, however, that these peoples had any acquaintance with Maya literacy or with Maya building techniques.

LATE CLASSIC LOWLAND MAYA (600–900)

Settlement pattern. There is still controversy over whether the Late Classic sites built by the lowland Maya were actually cities or whether they were relatively empty ceremonial centres staffed only by rulers and their entourages.

The common people built their simple pole-and-thatch dwellings on low earthen mounds to keep them dry during the summer rains. Thus, total mapping of a particular site should always include not only masonry structures but also house mounds as well. So far, only a few Maya sites have been so mapped. The mightiest Maya centre of all, Tikal in northern Petén, has a total of about 3,000 structures ranging from the tiny mounds up to gigantic temple pyramids; these are contained, however, within an area of six square miles. The Tikal population has been estimated from this survey to be 10,000–11,000 people, but perhaps as many as 75,000 within an even wider area could have belonged to Tikal.

Tikal

This sounds very much like a city, but the evidence actually can be differently interpreted. First, at the time of the conquest the Maya generally buried their dead beneath the floors of houses, which were then abandoned. Thus, an increase in number of house mounds could just as easily indicate a declining population in which the death rate exceeded the birth rate. Second, the appearance of even such a tremendous centre as Tikal is quite different from that of such true cities as Teotihuacán. An ordinary Maya family typically occupied two or three houses arranged around a rectangular open space. These were grouped into unplanned hamlets near good water and rich, well-drained soils. A survey of Petén has shown that for every 50 to 100 dwellings there was a minor ceremonial centre; this unit has been called a zone. Several zones formed a district for which a major centre like Tikal acted as the ceremonial and political nucleus. Neither Tikal nor any other such centre shows signs of town planning or neatly laid out streets.

There are also ecological factors that must have set certain limits upon the potential for urban life in the Maya lowlands. Slash-and-burn cultivation would have made for widely settled populations; and, as has been argued, the uniformity of the lowland Maya environment would have worked against the growth of strong interregional trade, always a factor in urban development. Yet these statements must be qualified. It is known that raised-field, or chinampa-type, farming was used in many places and at many times in the Maya lowlands. This would have allowed for greater population concentration. It is also known that there was a brisk trade in some commodities from one lowland Maya region to another.

Ecological factors

What, then, can be concluded about lowland Maya urbanism? Clearly, the urban form, even at a metropolis such as Tikal, was not as large or as formally developed as it was at highland Teotihuacán. At the same time, a centre whose rulers could draw upon the coordinated efforts of 75,000 people must inevitably have had some of the functions of a true city—in governance, religion, and trade, as well as in the development of the arts and intellectual life.

Major sites. While there are some important differences between the architecture of the Central and Northern subregions during the Late Classic, there are many features shared between them. A major Maya site generally includes several types of masonry buildings, usually constructed by facing a cement-and-rubble core with blocks or thin slabs of limestone. Temple pyramids are the most impressive, rising in a series of great platforms to the temple superstructure above the forests. The rooms, coated with white stucco, are often little more than narrow slots because of the confining nature of the corbeled vaults, but this was probably intentional, to keep esoteric ceremonies from the public.

The so-called palaces of Maya sites differ only from the temple pyramids in that they are lower and contain a great many rooms. Their purpose still eludes discovery; many scholars doubt that they really served as palaces, for the rooms are damp and uncomfortable, and there is little or no evidence of permanent occupation. The temples and palaces are generally arranged around courts, often with inscribed stelae and altars arranged in rows before them. Leading from the central plazas are great stone causeways, the function of which was probably largely ceremonial. Other features of lowland sites (but not universal) are sweathouses, ball courts, and probably marketplaces.

There are more than 50 known sites that deserve to be called major. Most are in the Central Subregion, with probably the greatest concentration in northern Petén, where Maya civilization had its deepest roots. Tikal is the largest and best-known Classic site of the Central Subregion. It is dominated by six lofty temple pyramids, one of which is some 230 feet high, the tallest structure ever raised by the Meso-American Indians. Lintels of sapodilla wood still span the doorways of the temple superstructures and are carved with reliefs of Maya lords enthroned amid scenes of great splendour. Some extraordinary Late Classic tombs have been discovered at Tikal, the most important of which produced a collection of bone tubes and strips delicately incised with scenes of gods and men. Ten large reservoirs, partly or entirely artificial, supplied the scarce drinking water for the residents of Tikal.

Other important sites of northern Petén include Uaxactún, Naranjo, Nakum, and Holmul, of which only the first has been adequately excavated. To the southeast of Petén are two Maya centres—Copán and Quiriguá—that show notable differences with the Petén sites. Copán is located above a tributary of the Motagua River in western Honduras in a region now rich in tobacco. Its architects and sculptors had a ready supply of a greenish volcanic tuff far superior to the Petén limestone. Thus, Copán architecture is embellished with gloriously baroque figures of gods, and its stelae and other monuments are carved with an extraordinary virtuosity. Copán also has one of the most perfectly preserved ball courts in Meso-America. Quiriguá is a much smaller site 30 miles north of Copán. While its architectural remains are on a minor scale, it is noted for its gigantic stelae and altars carved from sandstone.

The principal watercourse on the western side of the Central Subregion is the Usumacinta River, originating in the Guatemalan highlands and emptying into the Gulf of Mexico. For much of its course the Usumacinta is lined with such great Maya ceremonial centres as Piedras Negras and Yaxchilán. Even more renowned is Bonampak, a satellite of Yaxchilán located on a tributary of the Usumacinta. The discovery in 1946 of the magnificent murals embellishing the rooms of an otherwise modest structure astounded the archaeological world. From floors to vault capstones, its stuccoed walls were covered with

The Bonampak murals



A prisoner pleading for mercy before his captors, detail of a mural at Bonampak, Chiapas state, Mexico; original c. AD 800, Late Classic Maya; watercolour copy by Antonio Tejeda.
Peabody Museum, Harvard University; photograph by Hillel Burger

highly realistic polychrome scenes of a jungle battle, the arraignment of prisoners, and victory ceremonies. These shed an entirely new light on the nature of Maya society, which up until then had been considered peaceful.

In the hills just above the floodplain of the Usumacinta lies Palenque, the most beautiful of Maya sites. The architects of Palenque designed graceful temple pyramids and "palaces" with mansard-type roofs, embellished with delicate stucco reliefs of rulers, gods, and ceremonies. The principal structure is the Palace, a veritable labyrinth of galleries with interior courts; over it looms a four-story square tower that may have served as both lookout and observatory. A small stream flowing through the site was carried underneath the Palace by a long, corbel-vaulted tunnel. The temples of the Cross, Foliated Cross, and Sun were all built on the same plan, the back room of each temple having a kind of sanctuary designed like the temple of which it was a part. It can be supposed that all three temples served the same cult. The most extraordinary feature of Palenque, however, was the great funerary crypt discovered in 1952 deep within the Temple of the Inscriptions. Within a sarcophagus in the crypt were the remains of an unusually tall ruler, accompanied by the richest offering of jade ever seen in a Maya tomb. Over his face had been fitted a mask of jade mosaic, while a treasure trove of jade adorned his body.

Northward from the Central Subregion, in the drier and flatter environment of the Yucatán Peninsula, the character of lowland Maya civilization changes. Just north of Petén is the Río Bec zone, as yet little explored but noted for temple pyramids and palaces with flanking false towers fronted by unclimbable "stairways" reaching dummy "rooms" with blank entrances. Río Bec structures are carved with fantastic serpents in deep relief, a feature that becomes even more pronounced in the Chenes country to the northwest, in the modern state of Campeche. At Chenes sites, Maya architects constructed frontal portals surrounded by the jaws of sky serpents and faced entire buildings with a riot of baroquely carved grotesques and spirals.

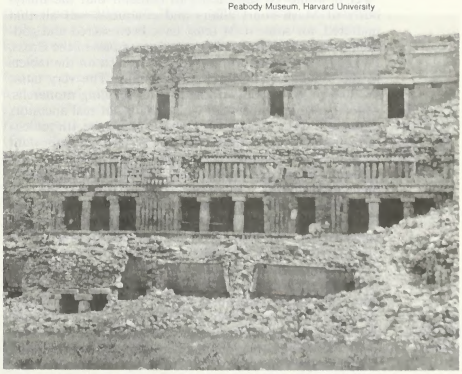
This elaborate ornamentation of buildings is far more restrained and orderly in the style called Puuc, so named from a string of low hills extending up from western Campeche into the state of Yucatán. The Puuc sites were for the Northern Subregion what the Petén sites were for the Central, for they are very numerous and clearly were the focal point for Maya artistic and intellectual culture. Uxmal is the most important Puuc ceremonial centre and an architectural masterpiece. It has all of the characteristics of the Puuc style: facings of thin squares of limestone veneer over a cement-and-rubble core; bootshaped vault stones; decorated cornices around columns in doorways; engaged or half-columns repeated in long rows; and lavish use of stone mosaics in upper facades, emphasizing sky-serpent faces with long, hook-shaped noses, as well as frets and latticelike designs of crisscrossed elements. Uxmal

The nearby centre of Kabah, connected to Uxmal by a ceremonial causeway, has an extraordinary palace completely faced with masks of the Sky Serpent. Other major Puuc sites are Sayil, with a multistoried palace, and Labná. The Puuc style reaches east across the Yucatán Peninsula, for at Chichén Itzá, a great site that was to occupy centre stage during the Toltec occupation of the Northern Subregion, there are several buildings strongly Puuc in character.

Puuc sites may be said to represent a lowland Maya "New Empire" in the sense that their apogee occurred in the 9th and 10th centuries, a time during which the great Petén, or Central Subregion, centres were in decline or had collapsed. Just how late Puuc sites remained active, with major constructions being dedicated, remains something of a question. In about 1000 a major change took place in northern Yucatán. It was marked by the construction of a number of Toltec-style temples and palaces at Chichén Itzá, a site that also has many Puuc-style edifices. It is not known if Toltec Chichén Itzá existed contemporaneously with such Puuc sites as Uxmal and Labná, and if so, for how long. Eventually, Chichén Itzá appears to have dominated northern Yucatán, lasting well into the Postclassic Period (about 1250). Questions also surround the bringers of Toltec-style architecture to Chichén Itzá. They may have been either central Mexican Toltecs or Gulf coast peoples who probably were Maya-speakers and who had adopted central-Mexican ways. In this connection, it should be noted that Puuc sites were under several influences from Gulf-coast Mexico, particularly from central Veracruz.

Maya art of the Late Classic. Maya art, at the height of its development, was fundamentally unlike any other in Meso-America, for it was highly narrative, baroque, and often extremely cluttered, unlike the more austere styles found elsewhere. It is essentially a painterly rather than sculptural tradition, and it is quite likely that even stone reliefs were first designed by painters. Much of this art has disappeared for all time because of the ravages of the wet, tropical environment on such perishable materials as wood, painted gourds, feathers, bark, and other substances. There must have been thousands of bark-paper codices, not one of which has survived from Classic times.

Peabody Museum, Harvard University



Ruins of the palace at Sayil, a Puuc site, Yucatán state, Mexico; Late Classic Maya.

Following the downfall of Teotihuacán, Maya artists were free to go their own way. Magnificently carved stelae and accompanying altars are found at most major sites, the greatest achievement in this line being found at Copán, where something approaching three-dimensional carving was the rule. Palenque and Yaxchilán specialized in graceful bas-reliefs placed as tablets or lintels in temple pyramids and palaces. In the Northern Subregion, however, the sculptor's art was definitely inferior in scope and quality and shows strong influence from alien, non-Maya cultures.

Wood carvings

A few wooden objects have somehow survived. Particularly noteworthy are the massive wooden lintels of Tikal, with scenes of lords and their guardian deities, accompanied by lengthy hieroglyphic texts. In ancient times, wood carvings must have been vastly more common than sculptures. The wet climate has also destroyed innumerable examples of mural art.

Maya pottery can be divided into two groups: (1) the pots and pans of everyday life, usually undecorated but sometimes with geometric designs, and (2) grave offerings. Vessels meant to accompany the honoured dead were usually painted or carved with naturalistic and often macabre scenes. To achieve polychrome effects of great brilliance, the Maya potters painted in semitranslucent slips over a light background, then fired the vessels at a very low temperature. Relief carving was carried out when the vessels were leather-hard, just before firing.

The most precious substance of all to the Maya was jade, to which their craftsmen devoted great artistry. Jade was mainly fashioned into thin plaques, carved in relief, or into beads. In the absence of metal tools, jade was worked by applying abrasives and water with cane or perhaps other pieces of jade.

The Maya calendar and writing system. It is their intellectual life that established the cultural superiority of the Maya over all other American Indians. Much of this was based upon a calendrical system that was partly shared with other Meso-American groups but that they perfected into a tool capable of recording important historical and astronomical information. Most Maya inscriptions that have been interpreted are calendrical inscriptions. Since the late 1950s, it has been learned that the content of Classic Maya inscriptions was far more secular than had been supposed. For many years specialists believed that the inscriptions recorded little more than the passage of time and that, in fact, the Maya were time worshipers; but it has been shown that certain inscriptions recorded the birth, accession, marriage, and military victories of ruling dynasties. One very significant advance in following dynastic histories and plotting political territoriality was the discovery in 1958 of "emblem glyphs," symbols standing for royal lineages and their domains.

Yet it would be misleading to contend that the hurly-burly of Maya court affairs and conquests was all that mattered, for some texts must have been sacred and god-oriented. At Palenque, in the similar temples of the Cross, Foliated Cross, and Sun, the dates inscribed on the tablets in the sanctuaries fall into three groups. The very latest seem to refer to events in the lives of reigning monarchs. An earlier group must deal with distant but real ancestors of those kings, while the very earliest fall in the 4th millennium BC and apparently describe the birth of important gods to whom the respective temples were dedicated and who may have been regarded as the progenitors of Palenque's royal house.

The meaning of many non-calendrical signs and even of complete clauses is not known, but there is a difference between this and assigning an actual Maya word to an ancient glyph or a sentence to a glyphic clause. While it is certain that the language of the Classic inscriptions was Mayan, it is also certain that it was more archaic than any of the Mayan languages spoken at the time of the conquest, six centuries after the Classic downfall.

The Maya codices

The three extant Maya codices, none dating earlier than 1100, contain a strong phonetic component, in fact a kind of syllabary, which can be successfully read as Yucatec-Maya, but the Classic peoples of the Central Subregion more likely spoke an ancestor of the Cholan branch of Maya. Furthermore, Maya hieroglyphic writing covers the entire span from about AD 250 to the conquest, during which time both the language or languages and the writing system itself must have undergone extensive evolution.

In writing systems in general, there is usually a development from pictographic signs, in which a picture stands for a word or concept, through logographic systems, in which words are still the basic unit but phoneticism is employed to reduce ambiguities (as in Chinese), to phonetic syllabaries, and finally to alphabets. Probably most Classic Maya hieroglyphs are logograms with a mainly ideographic orientation, and it seems that there was a considerable degree of flexibility in how the words and sentences could be written. By the Postclassic, this had been codified into a much more rigid system closely resembling that of Japanese, in which a well-defined syllabary can supplement or even replace logograms. There are approximately 300 to 500 logograms in Classic Maya (the number varies according to how one separates affixes from so-called main signs), but it will probably be many years before the majority of these are satisfactorily deciphered. Great progress, however, may be expected in unraveling their meaning in specific contexts.

(M.D.C./G.R.W.)

Maya mathematics included two outstanding developments: positional numeration and a zero. These may rightly be deemed among the most brilliant achievements of the human mind. The same may also be said of ancient Maya astronomy. The duration of the solar year had been calculated with amazing accuracy, as well as the synodical revolution of Venus. The Dresden Codex contains very precise Venusian and lunar tables and a method of predicting solar eclipses.

Maya chronology consisted of three main elements: a 260-day sacred year (*tzolkin*) formed by the combination of 13 numbers (1 to 13) and 20 day names; a solar year (*haab*), divided into 18 months of 20 days numbered from 0 to 19, followed by a five-day unlucky period (Uayeb); and a series of cycles—*uinal* (20 *kins,* or days), *tun* (360 days), *katun* (7,200 days), *baktun* (144,000 days), with the highest cycle being the *alautun* of 23,040,000,000 days. All Middle American civilizations used the two first counts, which permitted officials accurately to determine a date within a period defined as the least common multiple of 260 and 365: 18,980 days, or 52 years.

The Long Count

The Classic Maya Long Count inscriptions enumerate the cycles that have elapsed since a zero date in 3114 BC. Thus, "9.6.0.0.0," a *katun*-ending date, means that nine *baktuns* and six *katuns* have elapsed from the zero date to the day 2 Ahau 13 Tzec (May 9, AD 751). To those Initial Series were added the Supplementary Series (information about the lunar month) and the Secondary Series, a calendar-correction formula that brought the conventional date in harmony with the true position of the day in the solar year.

Both Classic and recent Maya held the *tzolkin* as the most sacred means of divination, enabling the priests to detect the favourable or evil influences attached to every day according to the esoteric significance of the numbers and the day-signs.

(Ja.S.)

Classic Maya religion. It has been denied that there was any such thing as a pantheon of deities in Classic times, the idea being that the worship of images was introduced by the Toltec or Itzá invaders, or both, in the Postclassic. Several gods who played significant roles in the Postclassic codices, however, can be identified on earlier Maya monuments. The most important of these is Itzamná, the supreme Maya deity, who functioned as the original creator god, as well as lord of the fire and therefore of the hearth. In his serpent form he appears on the ceremonial bar held in the arms of Maya rulers on Classic stelae. Another ophidian deity recognizable in Classic reliefs is the Feathered Serpent, known to the Maya as Kukulcán (and to the Toltecs and Aztecs as Quetzalcóatl). Probably the most ubiquitous of all is the being known as Bolon Tzacab (first called God K by archaeologists), a deity with a baroquely branching nose who is thought to have functioned as a god of royal descent; he is often held as a kind of sceptre in rulers' hands.

The Classic Maya lavished great attention on their royal dead, who almost surely were thought of as descended from the gods and partaking of their divine essence. Many reliefs and all of the pictorial pottery found in tombs deal with the underworld and the dangerous voyage of the soul through that land. Classic Maya funerary ceramics show that this dark land was ruled by a number of gods, including several sinister old men often embellished with jaguar emblems, the jaguar being associated with the night and the nether regions.

Human sacrifice

The Classic, as well as the Postclassic, Maya practiced human sacrifice, although not on the scale of the Aztecs. The victims were probably captives, including defeated rulers and nobles. Self-sacrifice or self-mutilation was also common; blood drawn by jabbing spines through the ear or penis, or by drawing a thorn-studded cord through the tongue, was spattered on paper or otherwise collected as an offering to the gods. (M.D.C./G.R.W.)

The four main categories of documents that provide knowledge of the Maya civilization and its religion are: archaeological remains; native books in hieroglyphic writing; books in native languages written in Latin script by learned Indians; and early accounts written in Spanish by conquerors or priests.

From surviving temples, tombs, sculpture, wall paintings, pottery, and carved jades, shells, and bone, a significant amount of valuable information can be gained; *e.g.,* representations of godheads and ritual scenes. Perhaps the most important archaeological source, however, is the hieroglyphic texts carved on stone monuments or stone or bone artifacts and painted on pottery. These, insofar as they can be translated, provide descriptions of ceremonies and beliefs.

Three native hieroglyphic books of pre-Columbian date survived the Spanish conquest: the Dresden, Madrid, and Paris codices, named for the cities in which they are now housed. Written on bark paper, they deal with astronomical calculations, divination, and ritual. They appear to be Postclassic copies of earlier Classic originals.

After the Spanish conquest, books were written by learned Indians who transcribed or summarized hieroglyphic records. Such is the case of the *Books of Chilam Balam,* in Yucatec Maya, and of the *Popol Vuh,* in Quiché, a highland Maya language. The former consist of historical chronicles mixed with myth, divination, and prophecy, and the latter (which shows definite central Mexican influences) embodies the mythology and cosmology of the Postclassic Guatemalan Maya. The *Ritual of the Bacabs* covers religious symbolism, medical incantations, and similar matters.

The most important of the early accounts written by the Spanish themselves is Diego de Landa's *Relación de las cosas de Yucatán* ("On the Things of Yucatán"), which dates to about 1566. It describes Postclassic rather than Classic religion, but given the deeply conservative nature of Maya religion, it is highly probable that much of this description is pertinent for the earlier period. Landa's account is also an excellent description of other aspects of Maya life in 16th-century Yucatán.

To these archaeological, ethnohistorical, and historical sources may be added the observations of modern ethnologists about the present-day Maya. Thus, in the Guatemalan highlands, the 260-day calendar still survives, as do ancient prayers to and information about Maya gods.

It is likely that a simpler religion of nature worship prevailed in Early Formative times. This probably began to undergo modification during the Middle Formative, as astronomical knowledge became more precise. Certainly by the Late Formative (300 BC, if not earlier), with the appearance of major centres and pyramid and temple constructions, an elaborate worldview had evolved. Deified heavenly bodies and time periods were added to the earlier-conceived corn and rain gods. Concepts derived from priestly speculation were imposed upon the simpler religious beginnings. Religion became increasingly esoteric, with a complex mythology interpreted by a closely organized priesthood.

Creation. The Maya, like other Middle American Indians, believed that several worlds had been successively created and destroyed before the present universe had come into being. The Dresden Codex holds that the end of a world will come about by deluge: although the evidence derived from Landa's *Relación* and from the Quiché *Popol Vuh* is not clear, it is likely that four worlds preceded the present one. People were made successively of earth (who, being mindless, were destroyed), then of wood (who, lacking souls and intelligence and being ungrateful to the gods, were punished by being drowned in a flood or devoured by demons), and finally of a corn gruel (the ancestors of the Maya). The Yucatec Maya worshiped a creator deity called Hunab Ku, "One-God." Itzamná ("Iguana House"), head of the Maya pantheon of the ruling class, was his son, whose wife was Ix Chebel Yax, patroness of weaving.

Myths about the creation of man

Four Itzamnas, one assigned to each direction of the universe, were represented by celestial monsters or two-headed, dragonlike iguanas. Four gods, the Bacabs, sustained the sky. Each world direction was associated with a Bacab, a sacred ceiba, or silk cotton tree, a bird, and a colour according to the following scheme: east–red, north–white, west–black, and south–yellow. Green was the colour of the centre.

The main act of creation, as stated in the *Popol Vuh,* was the dawn: the world and humanity were in darkness, but the gods created the Sun and the Moon. According to other traditions, the Sun (male) was the patron of hunting and music, and the Moon (female) was the goddess of weaving and childbirth. Both the Sun and the Moon inhabited the earth originally, but they were translated to the heaven as a result of the Moon's sexual license. Lunar light is less bright than that of the Sun because, it was said, one of her eyes was pulled out by the Sun in punishment for her infidelity.

Because the Maya priests had reached advanced knowledge of astronomical phenomena and a sophisticated concept of time, it appears that their esoteric doctrines differed widely from the popular myths.

Cosmology. The Maya believed that 13 heavens were arranged in layers above the earth, which itself rested on the back of a huge crocodile or reptilian monster floating on the ocean. Under the earth were nine underworlds, also arranged in layers. Thirteen gods, the Oxlahuntiku, presided over the heavens; nine gods, the Bolontiku, ruled the subterranean worlds. These concepts are closely akin to those of the Postclassic Aztec, but archaeological evidence, such as the nine deities sculptured on the walls of a 7th-century crypt at Palenque, shows that they were part of the Classic Maya cosmology.

Time was an all-important element of Maya cosmology. The priest-astronomers viewed time as a majestic succession of cycles without beginning or end. All the time periods were considered as gods; time itself was believed to be divine.

The gods. Among the several deities represented by statues and sculptured panels of the Classic period are such gods as the young corn god, whose gracious statue is to be seen at Copán, the sun god shown at Palenque under the form of the solar disk engraved with anthropomorphic features, the nine gods of darkness (also at Palenque), and a snake god especially prominent at Yaxchilán. Another symbol of the corn god is a foliated cross or life tree represented in two Palenque sanctuaries. The rain god (Chac) has a mask with characteristic protruding fangs, large round eyes, and a proboscis-like nose. Such masks are a common element in Puuc architecture.

Agricultural and astral deities

The three hieroglyphic manuscripts, especially the Dresden Codex, depict a number of deities whose names are known only through Postclassic documents. Itzamná, lord of the heavens, who ruled over the pantheon, was closely associated with Kinich Ahau, the sun god, and with the moon goddess Ix Chel. Though Itzamná was considered an entirely benevolent god, Ix Chel, often depicted as an evil old woman, had definitely unfavourable aspects.

The Chacs, the rain gods of the peasants, were believed to pour rain by emptying their gourds and to hurl stone axes upon the earth (the lightning). Their companions were frogs (*uo*), whose croakings announced the rains. Earth gods were worshiped in the highlands, and wind gods were of minor importance in Maya territory.



The corn god (left) and the rain god, Chac. Drawing from the Madrid Codex (Codex Tro-Cortesianus), one of the Mayan sacred books. In the Museo de América, Madrid.
By courtesy of the Museo de America, Madrid

The corn god, a youthful deity with an ear of corn in his headdress, also ruled over vegetation in general. His name is Ah Mun, and he is sometimes shown in combat with the death god, Ah Puch, a skeleton-like being, patron of the sixth day-sign Cimi ("Death") and lord of the ninth hell. Several other deities were associated with death; *e.g.,* Ek Chuah, a war god and god of merchants and cacao growers, and Ixtab, patron goddess of the suicides.

In Postclassic times, central Mexican influences were introduced; *e.g.,* the Toltec Feathered Serpent (Quetzalcóatl), called Kukulcán in Yucatán and Gucumatz in the Guatemalan highlands.

The ancient Maya's attitude toward the gods was one of humble supplication, since the gods could bestow health, good crops, and plentiful game or send illness and hunger. Prayers and offerings of food, drink, and incense (*pom*) were used to placate the gods. A strong sense of sin and a belief in predestination pervaded the Maya consciousness. Man had to submit to the forces of the universe. The priests, because of their astronomical and divinatory knowledge, determined favourable days for such undertakings as building houses and hunting.

Death. As was noted above, the Classic Maya buried the dead under the floors of their houses. High priests or powerful lords were laid to rest in elaborate underground vaults. The dead were believed to descend to the nine underworlds, called Mitnal in Yucatán and Xibalba by the Quiché. There is no evidence of a belief among the Maya in a heavenly paradise, such as that which prevailed in central Mexico. The modern Lacandón, however, believe that the dead live forever without work or worry in a land of plenty located somewhere above the earth.

The end of the world

Eschatology. The present world, the Maya believed, is doomed to end in cataclysms as the other worlds have done previously. According to the priestly concept of time, cycles repeat themselves. Therefore, prediction was made possible by probing first into the past and then into the future: hence the calculations, bearing on many millennia, carved on temples and stelae. Evil influences were held to mark most of the *katun* endings. The *Chilam Balam* books are full of predictions of a markedly direful character. The priests probably believed that the present world would come to a sudden end, but that a new world would be created so that the eternal succession of cycles should remain unbroken.

Forms of sacrifice

Sacrifice. Sacrifices made in return for divine favour were numerous: animals, birds, insects, fish, agricultural products, flowers, rubber, jade, and blood drawn from the tongue, ears, arms, legs, and genitals. Evidence of human sacrifice in Classic times includes two Piedras Negras stelae, an incised drawing at Tikal, the murals at Bonampak, various painted ceramic vessels, and some scenes in native manuscripts. Only in the Postclassic era did this practice become as frequent as in central Mexico. Toltec-Maya art shows many instances of human sacrifice: removal of the heart, arrow shooting, or beheading. At Chichén Itzá, in order to obtain rain, victims were hurled into a deep natural well (cenote) together with copper, gold, and jade offerings. Prayers for material benefits (which were usually recited in a squatting or standing position), fasting and continence (often for 260 days), and the drawing of blood from one's body often preceded important ceremonies and sacrifices.

These practices had become so deeply rooted that, even after the Spanish conquest, Christian–pagan ceremonies occasionally took place in which humans were sacrificed by heart removal or crucifixion. The last recorded case occurred in 1868 among the Chamula of Chiapas.

The priesthood. Bejeweled, feather-adorned priests are often represented in Classic sculpture. The high priests of each province taught in priestly schools such subjects as history, divination, and glyph writing. The priesthood, as described by Landa, was hereditary. *Ahkin,* "he of the sun," was the priests' general title. Specialized functions were performed by the *nacom*s, who split open the victims' breasts, the *chac*s who held their arms and legs, the *chilan*s who interpreted the sacred books and predicted the future. Some priests used hallucinatory drugs in their roles as prophets and diviners.

Rites. Ritual activities, held on selected favourable days, were complex and intense. Performers submitted to preliminary fasting and sexual abstinence. Features common to most rites were: offerings of incense (*pom*), of balche (an intoxicating drink brewed from honey and a tree bark), bloodletting from ears and tongues, sacrifices of animals (human sacrifices in later times), and dances. Special ceremonies took place on New Year's Day, 0 Pop, in honour of the "Year-Bearer"; *i.e.,* the *tzolkin* sign of that day. Pottery, clothes, and other belongings were renewed. The second month, Uo, was devoted to Itzamná, Tzec (fifth month) to the Bacabs, Xul (sixth) to Kukulcán, Yax (10th) to the planet Venus, Mac (13th) to the rain gods, and Muan (15th) to the cocoa-tree god. New idols were made during the eighth and ninth months, Mol and Ch'en, respectively.

The ritual ball game

Both the Classic and Postclassic Maya practiced a typically Middle American ritual ball game, as evidenced by numerous grandiose ball courts at Tikal, Copán, and Chichén Itzá. No court, however, has been found at Mayapán, and Landa does not mention that game. It appears, therefore, that the Yucatec had ceased to play it, while it remained of the utmost importance in central Mexico.

Archaeological remains at Uxmal and Chichén Itzá point to phallic rites, doubtless imported into the Yucatán from the Gulf coast. The *Chilam Balam* books strongly condemn the Mexican immigrants' sexual practices, which were quite alien to Maya tradition.

Sorcery. *Ahmen,* "he who knows," was the name given to sorcerers and medicine men, who were both prophets and inflicters or healers of disease. They made use of a mixture of magic formulas, chants, and prayers and of traditional healing methods, such as administering medicinal herbs or bleeding. Belief in witchcraft is widespread among present-day Maya Indians, as it most probably was in pre-Columbian times.

The evolution of Maya religion parallels that of Mexican religions from the Classic to the Postclassic era, with the sun worship and human sacrifice complex gaining importance as it did in Mexico proper.

The profoundly original feature of Maya religious thought, in comparison with that of other pre-Columbian civilizations, is the extraordinary refinement of mathematical and astronomical knowledge, inextricably mixed with mythological concepts. Even the most learned Aztec priests never reached the intellectual level of their Maya counterparts of the 1st millennium, nor did they conceive of the eternity of time and of its "bearers," the divinized time periods. The ancient Maya may be said to have been among the very few people in history (along with the Zurvanites of Iran) who worshiped time.

The simple, naturalistic religion of the corn-growing peasants, however, subsisted apart from the priesthood's abstract speculations and has partly survived to this day among the Christianized Maya Indians or the unevangelized Lacandón. (Ja.S./G.R.W.)

Society and political life. There is a vast gap between the lavishly stocked tombs of the Maya elite who ran the ceremonial centres and the simple graves of the peasantry. Careful measurements of the skeletons found in tombs and graves have also revealed that persons of the Maya ruling class were much taller than the tillers of the soil who provided them tribute. It is likely that this gulf was unspannable, for throughout Meso-America the rulers and nobility were believed to have been created separately from commoners.

The most revealing testimony to this royal cult is the temple pyramid itself, for almost every one explored has a great tomb hidden in its base. On death, each ruler might have been the object of ancestor worship by members of his lineage, the departed leader having become one with the god from whom he claimed descent. Ancestor worship, in fact, seems to be at the heart of ancient and modern society and religion among the Maya.

The ordinary folk may have participated in the ceremonies of even the greatest Maya centres. The modern highland Maya have a complex ceremonial life in which a man advances through a series of *cargos,* or "burdens," each one of which brings him greater prestige, costs him a great deal of money, and requires that he reside in the otherwise nearly empty centre for a year at a time carrying out his religious duties. The same may have prevailed in Classic times, though all activities were then under the direction of a hereditary and divine elite class, long since destroyed by the Spaniards.

Warfare

Warfare apparently was a continuing preoccupation of the Maya lords. Translations of hieroglyphic inscriptions show that in some cases such warfare led to territorial aggrandizement and the domination of one centre or polity by another; however, the principal purpose of war appears to have been to gain captives for slavery and sacrifice.

It has often been said that the Maya realm was a theocracy, with all power in the hands of the priests. That this is a misconception is apparent from the monuments themselves, which show kings, queens, heirs, and war prisoners, but no figures surely identifiable as priests. In 16th-century Yucatán, the priesthood was hereditary, and it is reported that younger sons of lords often took on that vocation. Quite probably such a class was also to be found among the Late Classic Maya, but neither for the Maya nor for any other Classic civilization of Meso-America can the term theocracy be justified.

The collapse of Classic Maya civilization. In the last century of the Classic period, Maya civilization went into a decline from which it never recovered. Beginning about 790 in the western edge of the Central Subregion, such ceremonial activity as the erection of stelae virtually came to a standstill. During the next 40 years this cultural paralysis spread gradually eastward, by which time the great Classic civilization of the Maya had all but atrophied. A date in the Maya calendar corresponding to 889 is inscribed on the last dated monuments in the Central Subregion; soon after the close of the 9th century it is clear that almost all of this region was abandoned.

For this event, which must have been one of the greatest human tragedies of all time, there are few convincing explanations. It now seems that the Classic Maya civilization in the region of its greatest development went out "not with a bang but a whimper." Massive foreign invasions can be discounted as a factor, but non-Maya elements did appear in the west at the same time as ceremonial activity terminated. These became the inheritors of whatever was left of the old civilization of the Central Subregion after AD 900, having established trading colonies and even a few minor ceremonial centres on its peripheries.

Whatever incursions did take place from the west were piecemeal and probably the result of the general decline, rather than its cause. Similarly, there is little reason to believe that there were peasant revolts on a general scale. The only real fact is that most of the inhabitants of the Central Subregion went elsewhere. Probably some were absorbed by such still flourishing ceremonial centres of the Northern Subregion as Uxmal and Kabah, while others might have migrated up into the congenial highlands of Chiapas and Guatemala. Although a population explosion and severe ecological abuse of the land must have played their role in the tragedy, the full story of the decline and fall of this brilliant aboriginal civilization remains to be told. (M.D.C./G.R.W.)

Postclassic Period (900–1519)

DEFINITION OF THE POSTCLASSIC

The final period of pre-Columbian Meso-American history is referred to as the Postclassic. Its beginning is usually placed at 900, and it terminates with the arrival of the Spanish conquistador Hernán Cortés in 1519 or with his conquest of the Aztec in 1521. The 900 date is based on two considerations: first, the 10th century was the period of the catastrophic collapse of the lowland Maya civilization and the cessation of the custom of erecting monuments dated by the Long Count; second, 900 was also the approximate date of the founding of the city of Tula in central Mexico and the rise of a people called the Toltec, who, according to the historical annals, built the first great empire in Meso-America. At one time it was thought that the date marked the collapse of all of the regional Classic civilizations of the area as the result of massive population dislocation. But it now appears that some Classic civilizations declined as early as 750, whereas others persisted until as late as 1200. The period is usually divided into two phases: Early Postclassic (900–1200) and Late Postclassic (1200–1519), the former equivalent with the period of the Toltec, the latter with that of the Aztec. The Postclassic civilizations of Meso-America came to an abrupt end with the coming of the Spanish in the early 16th century. For an account of the Spanish conquest, see the article LATIN AMERICA, THE HISTORY OF.

The two Postclassic phases

The Postclassic Period as a whole has also been distinguished from the Classic on the basis of assumed major changes in Meso-American political, economic, and social institutions. It has been asserted, for example, that the Classic period was one of relatively peaceful contact between polities, of the absence of large imperialistic states and empires (and of the militaristic élan and organization that accompanies such states). The Classic has been further characterized by the absence of true cities, by theocratic rather than secular government, and by an overall superiority of arts and crafts, with the exception of metallurgy, which appears for the first time in the Postclassic Period. In contrast, the Postclassic was characterized as a period of intense warfare and highly organized military organization, of empires and cities, of secular government, and of overall artistic decline.

Subsequent research, however, has cast considerable doubt on these conclusions. Many of the contrasts were drawn from events in the lowland Maya area and applied to the entire culture area; others were concluded essentially by a comparison of the Classic Maya of the lowland tropical forest of northern Guatemala and the Yucatán Peninsula with the Postclassic Aztec living in central Mexico in a dry mountain basin 7,000 feet above sea level. The differences, in part, are the product of separate culture evolution, conditioned by ecological factors. Cities and large states comparable to those built by the Toltec and Aztec were present in Early Classic times at Teotihuacán in central Mexico and probably at Monte Albán in Oaxaca. Militarism was at least significant enough to be a major artistic theme throughout the Classic period, even among the lowland Maya. One could also question the criterion of artistic decline, since a number of Postclassic crafts were highly developed, such as Aztec sculpture, Mixtec ceramics and metallurgy, and Zapotec architecture.

The separation between Postclassic and Classic is therefore little more than a convenient way of splitting up the long chronicle of Meso-American cultural development into manageable units for discussion and analysis. The Postclassic is a period also in which historical traditions combine with archaeological data, whereas the Classic

either lacks a written history or, in the case of the lowland Maya, provides little more than cryptic biographies of kings. Perhaps this is the best rationale for definition of the period.

At the time of the Spanish conquest, Meso-America was occupied by a number of peoples speaking languages as distinct from each other as English is from Chinese. On the central Gulf coast and adjacent escarpment were the Totonac; in Oaxaca and adjacent portions of Puebla and Guerrero two major ethnic groups, the Mixtec and the Zapotec, shared the western and eastern portions of the area, respectively; and in Michoacán lived the Tarascan. Various peoples of the Maya linguistic family occupied most of Guatemala, the Yucatán Peninsula, eastern Tabasco, and highland Chiapas; a detached group, the Huastec (Huaxtec), occupied the north Gulf coast. An equally widespread

The Nahua peoples

family, the Nahua (to which the Aztec belonged) occupied most of the Central Plateau, a huge area in the northwest frontier, portions of Guerrero, the Pacific coast of Chiapas and Guatemala (where they were known as the Pipil), and the Gulf coast. Some detached groups had spread beyond the frontier of Meso-America into Nicaragua and Panama. The linguistic family to which the Nahua belong (the Uto-Aztecan) is the only Meso-American family with affinities to languages north of the Rio Grande, including those of such western U.S. Indians as the Hopi, Paiute, and Shoshoni.

One of the Nahua-speaking nations, the Mexica, or Tenochca (or the Aztec, as they are commonly called), were the dominant people in Meso-America in 1519, having created by conquest an empire estimated as covering some 80,000 square miles (207,000 square kilometres) and having a population of 5,000,000 to 6,000,000 people.

All of these diverse ethnic groups shared a common cultural tradition, but separate historical origins and environmental factors had also produced a substantial degree of regional differentiation. Most of the cultural characteristics of the area go back at least to the beginning of the Postclassic, and many appeared in Classic times. The various regional cultures and languages have great time depths and undoubtedly were present during the Classic period. Common institutional characteristics included organization into centralized polities, including populations minimally in the tens of thousands, with a formal government, supported by a highly organized taxation system; stratification into social classes (including slave and serf classes); occupational specialization—in some areas full time with a guildlike organization; highly organized local and interregional trade involving professional merchants and regularly meeting markets; and a professional priesthood.

Post-classic technology

The technological base of this elaborate institutional structure seems weak by western European standards, since the primary technology (*i.e.*, the tools used to manufacture other technology) was based on chipped and ground stone, metal being reserved primarily for ornaments. Since draft animals were absent, all power was based on human energy. The economic base of the civilization was a highly productive agriculture, but the basic tools were primitive—stone axes for clearing vegetation and a number of wooden digging tools for working the soil. The crop complex was rich, with corn (maize) serving as the staple food and beans an important source of protein. But the list of secondary crops was large: chili peppers, tomatoes, squashes, sweet potatoes, cassava (manioc), cotton, tobacco, cacao, pineapples, papayas, maguey, nopals (prickly pears), sapotes (zapotes), peanuts (groundnuts), avocados, amates (paper figs), and many others.

Many crops were limited to particular environmental zones, thus acting as a major stimulus to trade. In many areas, particularly the tropical lowlands, the slash-and-burn, or swidden, system of farming was employed: forests were cleared, planted for up to three years, and rested for longer periods to restore fertility and eliminate the more difficult weeds. This regular rotation of fields resulted in high production per capita but had low demographic potential because in any given year most of the land lay fallow. In some lowland areas permanent grain and orchard cropping were practiced. In the drier highlands a number of specialized techniques were used, and agriculture generally was more intensive. Particularly important were terracing, irrigation, and swamp reclamation. The per capita productivity of highland agriculture was probably less (because of the higher labour input), but the demographic capacity was considerably greater than that in the lowlands. As a result of these highly effective approaches to farming, the population was dense when compared to western Europe in the 16th century. Population estimates for the conquest period have varied from 3,000,000 to 30,000,000; a reasonable estimate is between 12,000,000 and 15,000,000.

The diet of the average Meso-American was relatively uniform throughout the area. Dried corn was boiled in lime-impregnated water to soften the hull, ground into a dough on milling stones (manos and metates), and then either made into tortillas or mixed with water and drunk as a gruel called posol. The tortillas were eaten with sauces prepared from chili peppers and tomatoes, along with boiled beans. This was essentially the diet of the peasant, with the addition of pulque, the fermented sap of the maguey, at higher altitudes. To this were added the other crops in minor quantities and combinations depending on the specific local environment. Luxury foods included cocoa drinks, meats (from game or from the only two domestic animals of significance, the hairless dog and the turkey), and fish. The diet of the peasant, as is the case even today, was low in animal protein; but apparently the quantity of vegetable protein ingested made up for this deficiency.

Post-classic arts and crafts

Major Postclassic Meso-American crafts were weaving of cotton and maguey fibre; ceramics for pottery vessels, figurines, and musical instruments; stone sculpture; featherwork used for personal and architectural ornament; lapidary work (jadeite, jade, serpentine, and turquoise); metalwork (using gold, copper, and, more rarely, silver) for ornaments and a few tools; woodworking, the products including large dugout canoes, sculpture, magnificently made drums, stools, and a great variety of household items; baskets for containers and mats; painting; and, most particularly, stone and lime concrete masonry architecture.

On the intellectual, ideological, and religious levels, although some diversity and certain elaborations occurred in some regions, there was a fundamental unity to the Meso-American area, the product of centuries of political and economic ties. The religion was polytheistic, with numerous gods specialized along the lines of human activities. There were gods for basic activities such as war, reproduction, and agriculture; cosmogenic gods who created the universe and invented human culture; and gods of craft groups, social classes, political systems, and their subdivisions. Gods were all-powerful and had to be constantly propitiated with offerings and sacrifices, a concept reaching its peak in personal bloodletting and human sacrifice. Certain gods, such as the god of rain (called Tlaloc in central Mexico), were found throughout the area. A fundamental concept was that of a quadripartite multilevel universe that, by 1519, had gone through five creations and four destructions. Meso-American religion heavily emphasized the astral bodies, particularly the Sun, the Moon, and Venus, and the observations of the movement of these bodies by the astronomer-priests were extraordinarily detailed and accurate. The major purpose of these observations was astrological, and the Meso-American priests had developed a number of time counts, or calendrical rounds, based in part on these observations. Two basic calendars, a 260-day divinatory calendar and one based on the solar year of 365 days, were found throughout the area.

One of the great intellectual achievements of Meso-American civilization was writing; in Postclassic times books were made from the inner bark of the paper-fig tree and used to record calendars, astronomical tables, dynastic history, taxes, and court records.

Religion was a pervasive force in Meso-American life, as the art demonstrates; and considerable surplus energy was devoted to it (*e.g.*, temple construction, support of a numerous professional priesthood). Many writers have stated that the major focus of Meso-American culture was in this sphere. In fact, the contrast between Postclassic and

Classic was in part based on the presumed even greater emphasis on religion in the art and architecture of the latter period.

THE HISTORICAL ANNALS

The rise of the Aztec. A major characteristic of the Postclassic, in contrast to the Classic, is the abundant historical documentation. The Aztec record is particularly rich, and much of it is undoubtedly genuine, although there is always the possibility that records were rewritten or tampered with for political reasons. One of the functions of Meso-American writing was to record the succession and achievements of dynastic lines, and consequently it served as a validation of power. Virtually all of the dynasties of the local states recorded their history. A problem in the utilization of these documents, other than the low number of survivals, is the fact that many of them have strong mythological overtones. The Aztec themselves, for example, as creators of a great empire, explained their rise in part to the fact that they were the chosen people of the war god Huitzilopochtli and were the sustainers of the sun god Tonatiuh. They started their history as a poor, nomadic tribe from the north, who entered the Basin of Mexico, led by a magician-priest, and ultimately settled on the lake islands because of a series of astrological predictions and signs. They lived for a while as a subject people and then embarked on their destined role as conquerors and priests of the sun god. Virtually all historical traditions of local groups begin with a migration, a period of trials, and ultimate success—and some records even claim that the people were hunters and gatherers during the early part of their history.

The Chichimeca peoples

On the northern frontier of Meso-America, in the arid Mexican Plateau, true hunters and gatherers, referred to as the "Chichimeca" by the civilized peoples, did actually reside in 1519. The name Chichimeca was frequently applied to the migrant groups. It is difficult to see how hunting and gathering bands could successfully invade areas of dense civilized populations; but agricultural groups, during periods of dynastic weakness, undoubtedly could. In fact, the term Chichimeca was also applied to agricultural but less civilized peoples (such as the Otomí in central Mexico) and thus connoted a lack of polish or a rustic life-style. Since the northwestern portion of Meso-America was occupied by such people and since they were Nahua in speech, the legends of periodic north–south migrations of invaders, though they may have a factual basis, probably refer to movements of agricultural rather than hunting and gathering peoples.

The histories of these invading groups take on a more convincing historical character after the legends of migration. In the Aztec case they record the founding of Tenochtitlán in 1325. By 1376 the Aztec had increased in numbers and prestige sufficiently to obtain a member of the ruling family of Culhuacan, a neighbouring state, to rule as their *tlatoani*, or king. His name was Acamapichtli. The Aztec at this time were paying tribute to another state, Azcapotzalco, on the lake shore; and they remained under this obligation through the reigns of his two successors, Huitzilhuitl (*c.* 1390–1415) and Chimalpopoca (1415–26). During the reign of Chimalpopoca, Maxtla, the ruler of Azcapotzalco, attempted to secure tighter control over subject states by replacing their *tlatoanis* with his own men. He succeeded in arranging the assassination of Chimalpopoca and the exile of Nezahualcóyotl, ruler of Texcoco, a state on the east shore of Lake Texcoco. In response to these acts, a coalition was formed between Nezahualcóyotl, Itzcóatl (Chimalpopoca's successor), and another small state (Tlacopan), and the power of Azcapotzalco was broken.

The triple alliance

A triple alliance was then formed between Tenochtitlán, Texcoco, and Tlacopan, which by 1519 resulted in the dominance of Aztec Tenochtitlán. Under the Aztec rulers Itzcóatl (1428–40), Montezuma I (1440–69), Axayacatl (1469–81), Tizoc (1481–86), Ahuitzotl (1486–1502), and Montezuma II (1502–20), and the two Texcocan rulers—Nezahualcóyotl (1431–72) and Nezahualpilli (1472–1516)—the triple alliance succeeded in conquering the vast domain described above. Tlacopan seems to have been relegated to an inferior political role early in the history. The records of the Aztec and neighbouring states in the Basin of Mexico between 1300 and 1519 are relatively free from mythological tales and have sufficient cross-referencing to present a reasonably clear picture of military events, dynastic succession, institutional changes, and economic development. The period from 1200 to 1300 is essentially one of migration legends of the dynasties of the various states, the historical traditions of which are discussed below.

The question of the Toltec. The historical traditions also state that these migrations were responsible, along with a series of natural disasters, for the collapse of a great empire ruled by a people called the Toltec from their capital of Tollan, or Tula. Many dynasties of the conquest period, not only in central Mexico but even as far afield as highland Guatemala and the Yucatán Peninsula, claimed descent from the Toltec, apparently as a result of their dispersion after the fall of Tula.

The traditions describe the Toltec as the first civilizers, the first city builders, and the originators of craft skills and astrological knowledge. The major questions are: Did the Toltec really exist as a people? Where was Tula? Did these people actually play the extraordinary political and cultural role ascribed to them? To begin with, the annals themselves are in fundamental disagreement with respect to dates and the lists of Toltec kings. There are at least three major chronologies of the Toltec Empire (see Table 1). The dates by Ixtlilxóchitl, a learned mestizo of the post-conquest period, place the Toltec well within the Classic period of Meso-American archaeology, but the others correlate them with the early portion of the Postclassic. Most writers favour the later dates, but this would mean that the Toltec were not the first civilized peoples in central Mexico, as they claim.

Table 1: Chronologies of the Toltec Empire

Ixtlilxóchitl		*Anales de Cuauhtitlán*		Codex Ramírez	
Chalchiuhtlanetzin	510–562	Huetzin	896–?		
Ixtlilcuechahauac	562–614	Totepeuh	?–887	Mixcoatl	900–947
Huetzin	614–666	Ihuitimal	887–923		
Totepeuh	666–718	Topiltzin	923–947		980–999
Nacoxoc	718–770	Matlacxochitl	947–983		1000–34
Mitl-tlacomihua	770–829	Nauhyotzin	983–997		1034–49
Xihuiquenitzan	829–833	Matlaccoatzin	997–1025		1049–77
Iztaccaltzin	833–885	Tlilcoatzin	1025–46		1077–98
Topiltzin	885–959	Huemac	1047–1122		1098–1168

Mythic evidence

Adding further doubt to the veracity of the Toltec history is the admixture of myth and magic in the annals, not only at the beginning (which, like the histories of later dynasties, begins with a migration under a magician priest) but throughout the narrative. The ruler Topiltzin, for example, is also called Quetzalcóatl (the Nahua name for the Feathered Serpent god); he is opposed by Tezcatlipoca (also an Aztec god) and is driven out of Tula. He flees with his followers to the Gulf of Mexico and embarks on a raft of serpents. The story sounds like a duplicate of the cosmic myth or conflict between the two gods (see below *Cosmogony and eschatology*). Notably, the Maya in Yucatán had a tradition of a landing on the west coast made by foreigners, under a leader named Kukulcán (which is the Maya word for Feathered Serpent), who founded a city at Chichén Itzá and ruled over the Maya.

In spite of all the objections, the traditions of a great empire and of the city of Tula are so persistent that they must refer to some historical event and, indeed, have some archaeological support.

ARCHAEOLOGICAL REMAINS OF POSTCLASSIC CIVILIZATION

The early Postclassic period (900–1200) in central Mexico is associated with three major sites, all of which began in Classic times: Cholula in Puebla, Xochicalco in Morelos, and Tula in Hidalgo. Cholula was a major centre as far back as Early Classic times, probably as a political dependency of Teotihuacán. It reached its maximum growth in Late Classic times, following the collapse of Teotihuacán, when the largest structure ever built by Meso-Americans was erected (see above *Cholula*).



Temple pyramid known as Structure B at Tula, Hidalgo state, Mexico.
© Robert Frerck—Odyssey Productions

In Postclassic times Cholula continued as a major religious and cultural centre. Xochicalco probably was of minor significance in Early Classic times; but it went through a phase of explosive growth in the Late Classic and was probably abandoned by 1200, possibly earlier. Tula, on the other hand, a small centre in the Late Classic, went through a rapid growth during the period 900–1200 and then declined to a provincial centre in the Late Postclassic. There is a strong suggestion that the demise of Classic Teotihuacán was in part related to the emergence of one or all of these major centres.

Tula. The location of the Toltec capital of Tollan, or Tula, is not certain. The archaeological site located on a low ridge near the modern town of Tula has been the persistent choice of all historians since the conquest, in part because of the coincidence of place-names. There is further support for this identification in that the annals provide a great number of place-names near the modern Tula that have persisted since the conquest. There is also support for the identification in that the glyph Ce Acatl, the birthday and birth name of the great Toltec leader Topiltzin, has been found carved on a hill near Tula. Moreover, the sculpture from the site is heavily loaded with symbolism that relates to the Quetzalcóatl cosmology and cosmogony. It clearly was the city of the god Quetzalcóatl. The confusion between the god and the ruler can be ascribed to the fact that the name Quetzalcóatl may have served as a title of office carried by all Toltec rulers. The archaeological dates are in agreement with the *Anales de Cuauhtitlán* and the Codex Ramírez (see below *The nature of the sources*).

The question of Tula's existence

The major factors that have made some researchers reluctant to accept this identification lie in the claim that Tula was the capital of a great pan-Meso-American empire and that the Toltec were the first civilizers in central Mexico. Archaeologically, it is quite clear that Tula was preceded by the great Classic centre of Teotihuacán. Tula as a site does not really approach the earlier Teotihuacán or the later Tenochtitlán in size, in the number of public buildings, or in estimated population, although studies indicate that Tula had a population of between 30,000 and 60,000. Furthermore, although some basic stylistic elements of the art and architecture of Tula are widespread, the style, in an integrated specific sense, is limited (with one notable exception) to a small area in central Mexico. These facts make it difficult to accept Tula as the capital of a great empire. But archaeological evidence of even the Aztec empire is skimpy. In both cases, this may mean that the expansion was a rapid, explosive one that failed to last long enough to register these effects. But at least in the case of Tenochtitlán it did result in the rapid growth of a truly gigantic urban centre.

Because of these objections and because Teotihuacán fits better the description of the Toltec as the builders of the first truly civilized society in central Mexico, that site must still be considered a possible candidate.

The art and architecture of Tula shows a striking similarity to the later art and architecture of Tenochtitlán, and the themes represented in the art indicate a close approximation in religious ideology and behaviour. The symbols of sun sacrifice and the marching predators represented in sculpture both suggest that the concept that the Aztec had of themselves as the warrior-priests of the sun god was directly borrowed from the people of Tula.

On the basis of the symbolism represented in the carvings on a temple pyramid at Tula called Structure B, it has been concluded that the pyramid was dedicated to the god Quetzalcóatl, lending further support to the identification of the site as the Toltec capital.

Chichén Itzá. Also in support of the identification of Tula as the Toltec capital are the architectural characteristics and stylistic features of the sculpture of a large site in northern Yucatán called Chichén Itzá. The resemblance between the two sites is extraordinarily close. At Chichén are found flat beam and masonry roofs (contrasting sharply with the typical Maya corbeled vault), serpent columns, colonnaded halls attached to the bases of temples, altars with Atlantean figures, sculptured representations of skulls and crossbones, marching felines, canines and raptorial birds devouring human hearts, and depictions of warriors with typical Toltec accoutrements. Furthermore, there are even scenes showing Toltec and Maya warriors in combat. The Temple of the Warriors at Chichén Itzá looks like an attempt to duplicate Structure B at Tula.

Tula and Chichén Itzá compared

One of the puzzling aspects of the relationship between the two sites is that the public architecture of Chichén Itzá is actually more monumental than that at Tula, leading at least one Meso-American specialist to believe that Tula's style was derived from Chichén. Many of the stylistic features themselves, however, have prototypes in Classic Teotihuacán, whereas there is little in Classic Maya culture that could be considered as the source. What is more probable—and this agrees with the Toltec version of the relationship—is that the Toltec state in Yucatán was politically independent from Tula and was larger in area and population. The presence of rival states in central Mexico such as Xochicalco and Cholula may have kept the core of the Toltec polity relatively restricted in space. The much larger area and population controlled by the Toltec state at Chichén would explain the differences in the scale of architecture. The superior military organization and equipment of the Toltec perhaps explains their apparent success in Yucatán.

Archaeological unity of the Postclassic. The Postclassic period of Meso-American archaeology generally is a period characterized by considerable regionalism combined with a certain degree of uniformity. To a great extent, the latter was the product of the large states and extensive trade networks centred in the central plateau region. The Early Postclassic in some areas may be described as a continuation of the Late Classic; on the Gulf coast, for example, sites like El Tajín continued to be occupied, while in the Valley of Oaxaca (although Monte Albán was abandoned) the Zapotec tradition continued with the new centre at Mitla. In other areas, new styles either began or reached their climactic development, such as the Mixteca–Puebla style in painting, ceramics, and metallurgy, which evolved either in western Oaxaca or, more probably, at Cholula in Puebla. On the Guatemalan Pacific piedmont and in Tabasco, two specialized ceramic traditions (both of which

Ceramic traditions

began in Late Classic times) evolved: (1) plumbate (so called because of its slip, which had an unusually high iron content in the natural clay that fired to a lead-colour glaze); and (2) Fine Orange (so called because of its fine-grained, temperless paste). Wares of these two styles were widely traded.

The unity of the Postclassic consisted primarily of the diffusion of religious ideology, particularly the sun god–warfare–sacrificial complex and of the related institutional development such as the military orders (the latter probably originated at Classic Teotihuacán). This ideology clearly originated in central Mexico, at either Cholula or Tula or both. The specific artistic style of representation of the themes in painting and sculpture spread as well. Along with this was diffused a specific style of representation of the social calendar and writing generally and much greater emphasis on the 52-year cycle. The specific style most probably originated at Cholula.

In the highland areas of Meso-America the Late Postclassic was a period of maximum population growth. The Early Postclassic was, however, the period of maximum expansion of sedentary peoples on the northern frontier, probably the product of minor changes in climate as a result of increased rainfall. This frontier retracted substantially in Late Postclassic times, possibly because the rainfall decreased. This was perhaps the major factor in the precipitate arrival of barbarous tribes into the plateau, as the annals state.

The Postclassic, over large areas of the lowlands, on the other hand, was strikingly different. One of the most intriguing problems of Meso-American archaeology is the peculiar sequence of events in the lowland Maya area. At the time of European contact much of the northern portion of Yucatán was well settled. A narrow band of densely settled country also extended along the east coast south to modern Belize City and along the entire length of the west coast (where it joined another area of substantial settlement in the south Gulf coast). Most of the heart of the peninsula, the department of Petén in Guatemala, and large portions of the states of Campeche and Quintana Roo in Mexico (the most densely settled portion of the Classic Maya territory) were virtually abandoned.

Reasons for the decline in population

One of the major problems of Meso-American archaeology is the explanation of this massive population decline. The immediate causes are clear: it must have been the product of migrations out of the area or a set of internal factors that caused a decline in situ or both. Various hypotheses as to processes and causes have been suggested. These may be grouped in the following categories: natural disasters (earthquakes, famines, epidemics, and hurricanes have all been suggested); ecological processes (primarily the deterioration of the natural environment by overintensification of land use in response to population pressure); and sociopolitical processes (internal warfare, invasion from outside, peasant revolts, breakdown of critical trade networks). Some of these hypotheses are clearly derivations from others or are not explanations but rather are descriptions of events that were produced by other processes. It seems certain that the causes were multiple and in some way related. Of great interest is the fact that at least one other lowland area, the Pacific Coastal Plain of Guatemala, experienced a comparable Postclassic decline. (W.T.Sa.)

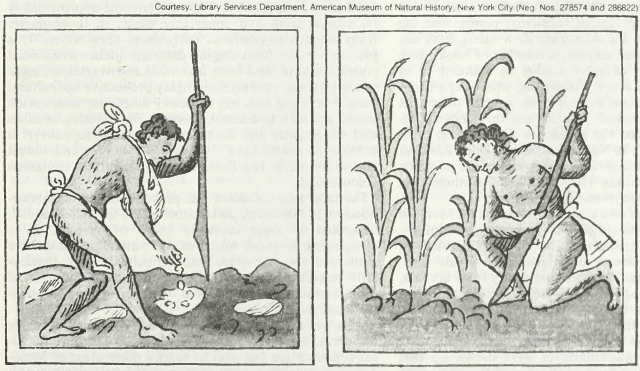
AZTEC CULTURE TO THE TIME OF THE SPANISH CONQUEST

The nature of the sources. At the time of the Spanish conquest the dominant people of Meso-America were the Aztec. This description is based primarily on written documents from the 16th century but also includes some archaeological data. The literature, both published and unpublished, of the 16th century is enormous and takes in all aspects of Aztec culture. Much of it covers the period within a few decades after the conquest, and it is uncertain how much change had occurred because of the introduction of Spanish culture. Some Aztec institutions, such as the military orders, were immediately abolished by the Spaniards; and the sources, therefore, give only the barest outline of their organization. This information, however, combined with archaeological data, gives a fairly detailed picture of Aztec culture at the time of the Spanish conquest. The sources can be classified by content and purpose into five categories, each of which is described below.

Accounts written by the conquistadores. Eyewitness accounts of Aztec culture on the eve of the conquest are, of course, the most directly pertinent sources because they describe Aztec culture before it became transformed by the Spanish conquest. Important among these are the *Cartas de Relación* ("Letters of Information"), sent by Hernán Cortés to the Holy Roman emperor Charles V, and the *Historia verdadera de la conquista de la Nueva España* (1632; *The True History of the Conquest of New Spain*) by Bernal Díaz del Castillo. Religious rites and ceremonies, temples, and paraphernalia of the cults are often described in these accounts. Their value, however, is lessened by the writers' ignorance of Náhuatl (the Aztec language at the time of the conquest), their lack of understanding of the Indian way of thinking, and their deep hostility to the native religion, which they considered to be inspired by the devil. These documents, therefore, have been interpreted with utmost caution.

Roman Catholic missionaries also wrote accounts of the Aztec. Paradoxically, the priests generally showed more understanding and tolerance than did the laymen. Thanks to their training and theological knowledge, they were able to analyze the Indian mind and to gain insight into the

Courtesy, Library Services Department, American Museum of Natural History, New York City (Neg. Nos. 278574 and 286822)



Aztec farmers (left) planting and (right) cultivating corn with the assistance of a wooden digging tool; illustrations from the Florentine Codex, a version, in Náhuatl, of the *Historia general de las cosas de Nueva España* by Bernardino de Sahagún, 16th century.

meaning of the myths and ritual. The missionaries, as a rule, learned the native languages, especially Náhuatl.

Postconquest histories of the Aztec written in Spanish. Within a few decades of the conquest, a series of histories had been written in the Spanish language, based in part on Aztec books and in part on information supplied by the upper class. Among the most detailed of these is the three-part *Historia de las Indias de la Nueva España e Islas de Tierra Firma* ("History of the Indies of New Spain"), written in about 1580 by the Dominican friar Diego Durán.

Postconquest ethnographic accounts written in Spanish and Náhuatl. These works are comparable in methodology and subject matter to the kinds of studies of native peoples conducted by present-day anthropologists. Probably the finest of them was written by Bernardino de Sahagún. Sahagún was a Franciscan priest who arrived in Mexico very early (1529), learned the Náhuatl tongue, and spent his life building a wonderful monument, a real encyclopaedia called the *Historia general de las cosas de Nueva España* ("General History of the Things of New Spain"). His work covers virtually all aspects of Aztec culture. It contains particularly detailed accounts of religion, ethnobotany, folk medicine, and economics, dictated to him in Náhuatl by Aztec noblemen and priests. As a source, it has the added value of being written in both Náhuatl and Spanish. One of the most complete versions of this work, written in Náhuatl, is called the Florentine Codex.

The work of Sahagún

The codices. Aztec sacred books and works, which were kept in the temples, and other native books have become known in Western scholarship as codices. Sacred books were written (or rather, painted) on deerskin or agave-fibre paper by scribes (*tlacuiloanime*), who used a combination of pictography, ideograms, and phonetic symbols and dealt with the ritual calendar, divination, ceremonies, and speculations on the gods and the universe. Because of their religious content only a small fraction of these escaped destruction by the Spaniards; the few specimens that have survived—such as the Codex Borbonicus, the Codex Borgia, the Codex Fejérváry-Mayer, and the Codex Cospi—usually come accompanied by Spanish notations. These sources are limited in scope and subject matter but nevertheless are valuable documents. Their interpretation is far from easy. Only a few of them, such as the Borbonicus, are truly Aztec, while others, such as the Borgia, seem to emanate from the priestly colleges of the "Mexica-Puebla" area, between the central highlands and the Oaxaca Mountains.

The codices

Other native books, either pre-Cortesian or post-Cortesian, also afford valuable material. Examples include such manuscripts as the Codex Telleriano-Remensis, the Azcatitlan, and the Codex of 1576, which describe the history of the Aztec tribe and state and occasionally depict religious scenes and events; the Codex Badianus, an herbal with magnificent drawings of medicinal plants; and the Codex Mendoza and the *Matrícula de tributos,* both tax documents of the Aztec empire. A number of books were written in the Latin alphabet—either in Náhuatl or in Spanish—by learned Aztec chroniclers, who used ancient pictographic manuscripts as their basis. Among those that were prepared in central Mexico are the Codex Chimalpopoca (also called the *Anales de Cuauhtitlán*; "Annals of Cuauhtitlán"), in Náhuatl, and the Codex Ramírez (also called the *Historia de los mexicanos por sus pinturas*; "History of the Mexicans Through Their Paintings"), in Spanish; both are anonymous compilations.

Official ecclesiastical and government records. Much of this literature is unpublished. Its purpose was administrative rather than intellectual, but it has provided an extraordinarily rich source of information for all 16th-century ethnic groups. The documents vary from tax lists, censuses, and marriage and baptismal records to broad geographic-economic surveys. Among the most valuable of the last type are the *Relaciones geográficas* of 1579–85, a series of surveys ordered by Philip II of his overseas possessions. Formal questionnaires were drawn up that demanded information from each town in the empire on virtually all aspects of Meso-American life: questions on the natural environment and resources, crops, population history, settlement patterns, taxes paid, markets and trade, the language, native history and customs, and progress of the missionization program. (W.T.Sa./Ja.S.)

Agriculture. The homeland of the Aztec, from which they ruled their vast domain, was a large (about 3,000 square miles), mountain-rimmed basin with a floor at approximately 7,000 feet above sea level. The surrounding ranges reached a maximum elevation of 18,000 feet in the volcano of Popocatépetl. The annual rainfall varied from 20 to 35 inches (500 to 900 millimetres) in the valley floor to a maximum of 50 inches on the southern escarpment. Approximately 80 percent of the rain fell between May 1 and October 1. Because of the high elevation, the area suffered from severe winter frosts that normally began in mid-October and lasted until the end of March. Normally, the rainfall was adequate for corn, even in the drier portions of the basin, but a major problem was the timing of the rains and the frosts. A delay of the rainfall to mid- or late June, accompanied by early autumn frosts, could produce crop disasters.

Factors limiting cultivation

Another major problem for the pre-Hispanic cultivator was the paucity of level land. Much of the land surface is sloping, and the problem of soil erosion was acute. Furthermore, of the 1,600 square miles of relatively level land, 400 square miles were occupied by a chain of lakes; and much of the immediate lakeshore plain was waterlogged.

Because of the effect of elevation on the growing season, the areas above 8,300 feet were also unsuitable for cultivation, removing an additional 400 square miles from the agricultural resource. Even within zones of cultivation, the presence of steep slopes and thin soil further reduced the area of cultivation. It is doubtful that more than 50 percent of the basin was suitable for labour-intensive methods of cultivation. Yet in 1519 it supported a population of 1,000,000 to 1,500,000; *i.e.,* a density of 500 people per square mile (200 per square kilometre), the densest population in Meso-American history. This was achieved by an extraordinarily intensive system of farming that involved a number of specialized techniques. Soil fertility was maintained by plant and animal fertilizers, by short-cycle fallowing, and by irrigation. In gently sloping terrain, erosion was controlled by earth and maguey terraces, in steeper areas by stone terracing. The problem of humidity was solved by canal irrigation of both the floodwater and permanent type. Much of the irrigation was done just before planting in April and May in order to give crops a head start and hence avoid the autumn frosts. Terracing functioned also as a method of conserving moisture. There is also evidence that dry-farming techniques were applied to store moisture in the soil. The most significant achievement of Aztec agriculture, however, was that of swamp reclamation, even including colonization of the lakes. This system of farming, called chinampa, was first applied to Lake Chalco. The lake covered approximately 60 square miles and apparently varied in its character from swamps to ponds of fairly deep, open water. By a process varying from digging drainage ditches to artificial construction of land from lake mud and vegetation, most of the lake was converted to highly productive agricultural land. A series of masonry causeway dikes were constructed across the lake to control flooding. By a system of dikes and sluice gates the Aztec even managed to convert a portion of saline Lake Texcoco, the largest and lowest lake in the basin, to a freshwater bay for further chinampa colonization.

The total area colonized was probably in the neighbourhood of 30,000 acres, and Tenochtitlán, the Aztec capital, depended on these lands for much of its food. By a comparable method, much of the waterlogged lakeshore plain was also converted into agricultural land. Particularly notable is the fact that all of these techniques of food production were achieved by human power and simple hand tools.

Aside from agriculture, the basin had a number of major resources, some of which were exploited not only for local consumption but also to supply other areas of Meso-America. Obsidian, natural glass of volcanic origin, was a superb material for a great variety of stone tools; and the northeastern ranges of the basin contained one of Meso-

Other resources of the basin area

America's major deposits. Basalt for manos and metates (milling stones) was also abundant. Lake Texcoco was a major source of salt, and the lakes generally provided waterfowl, fish, and other aquatic foodstuffs. The great pine forests above the limits of agriculture were a major source of lumber. On the other hand, the basin, because of its high elevation, was unsuitable for a great variety of tropical products, including cotton, paper, tropical roots and fruits, tobacco, copal incense, rubber, cacao, honey, precious feathers and skins, and such prized goods as metal, jade, and turquoise. The major motivation of Aztec conquest was to obtain control of these resources.

Social and political organization. Aztec technology differed little from that of other Meso-American groups. One of its distinctive aspects was differentiation by status levels. The use of most of the extra-local resources noted above was limited to a small upper and middle class; and there were striking differences in dress, housing, and diet by social class. Commoners, for example, wore clothing woven from maguey fibre, while the upper classes wore cotton garments. The use of imported foods, at least on a regular basis, was limited to the upper and middle classes. Commoners lived in small adobe or stone and mud huts, the upper and middle class in large multiroomed palatial houses of cut stone, lime plaster, and concrete.

The nuclear family

Aztec social and political organization can be divided into a number of levels of increasing size and complexity of organization. The nuclear family—that is, a pair of cohabiting adults and their unmarried children—formed the lowest level of organization. The nuclear family functioned in procreation, education of children, and as a unit of food preparation and consumption, with a well-defined division of labour between husband and wife. Among the Aztec, however, a number of nuclear families usually resided together in a single cooperating household, or extended family. Such a family usually consisted of a man, his married sons or brothers, and their families. The average peasant household of this type was small. Up to three nuclear families occupied a small multiroom house divided into apartments for each family. The houses were usually placed within a courtyard fenced with organ cactus or adobe walls, forming a compound. The extended family household probably functioned as a unit of land use and food production. In the towns, however, some households could be considerably larger, and the household of Montezuma II included several thousand people.

A number of households, varying from a few score to several hundred, were organized into an internally complex corporate group referred to as a calpulli by the Aztec and translated as barrio ("ward") by the Spaniards. Questions about the structure and function of this level of Aztec organization have caused a great deal of debate among Meso-American specialists. It is clear, however, that it was a physical and territorial unit as well as a socially organized one. It was a unit of land tenure. Calpulli lands were owned communally but were distributed among various households. The household retained the right of usufruct, but only the calpulli as a whole could sell or rent lands.

The calpulli rural communities varied considerably in physical appearance. Some were isolated, tightly nucleated physical settlements surrounded by their agricultural land, whereas in others houses were dispersed through the land holdings. In a few cases, they were physically attached as wards to one or more other calpulli. These differences corresponded to ecological, economic, and political factors. Rural, dispersed settlements were found on terraced hillsides in which houses were tightly integrated with the terrace; in the chinampa area, each house was placed on its chinampa holding. On the other hand, nucleated, isolated calpulli were found in areas of level land, and the ward type was usually found in the towns and cities. In the latter case, many lost their agricultural character and became units of craft specialization. The calpulli was a unit of political administration within the larger unit that will be referred to here as the state. It was ruled by a council of household heads presided over by a chief selected by the council from within a particular lineage. The calpulli functioned as a unit of taxation to the central government, as a unit of corvée labour, and as a military regiment.

Structure of the calpulli

The structure of the calpulli is open to question. Some sources call it a kin group, "a lineage" with a common ancestor; and as a result some anthropologists have referred to it as a clan, or sib. There is no evidence, however, of either exogamy or unilineal descent; in fact, marriage records from the post-conquest period show a strong tendency toward endogamy. There is some evidence of internal ranking and significant status differentiation, another non-clanlike feature. The sources also mention smaller territorial subdivisions, referred to as *barrios pequeños,* or "little wards." If these are descent lines, then the calpulli resembled quite closely a type of kin group called by anthropologists a ramage, or a conical clan. This is a group with a myth of common descent, divided into ranked senior and junior lineages based on the seniority of older versus younger brother in the group genealogy. In support of this reconstruction is the statement that the calpulli god was a deified ancestor.

The calpulli also functioned as a unit of education, for each possessed a school for young men—the telpuchcalli—primarily for military and moral instruction.

Above the level of the calpulli was the state. With the exception of those historical periods when larger polities, such as the Aztec empire, emerged, such states in Meso-America, including the Basin of Mexico, were small. Just prior to the Aztec expansion there were 50 or 60 such states in the basin, with an average size of about 50 to 60 square miles. In 1519 these once independent domains had an average population of 25,000 to 30,000 people. In less densely settled areas, the territories were larger and populations smaller. The range of size was from a few thousand up to 100,000.

The average small state included a central town with a population of several thousand, the balance of the population consisting of the rural calpulli. The central town was divided into wards that corresponded in size and to a certain degree in structure to the rural calpulli but were clearly different in function; they in turn were divided into *barrios pequeños.* At the head of the state was an official called the *tlatoani,* to whom all household heads owed allegiance, respect, and tax obligations. The *tlatoani*'s position was fixed within a particular lineage, the particular choice varying from state to state. In some areas, succession passed from father to son; in others, the succession went through a series of brothers and then passed to the eldest son of the eldest brother. In still other states, the office was elective, but the choice was limited to sons or brothers of the deceased ruler. The office was accompanied by all of the trappings and sumptuary behaviour typical of despotic states. The ruler resided in a large, multiroom masonry palace inhabited by a great number of wives, servants, and professional craftsmen. He was carried in a sedan chair in public and treated with exaggerated respect by his subordinates. The *tlatoani* held considerable power: he appointed all lesser bureaucrats, promoted men to higher military status, organized military campaigns, and was the distributor of booty and tribute; he collected taxes in labour, military service, and goods from his supporters; he owned private estates manned by serfs; he was the final court of appeal in judicial cases; and he was titular head of the religious cult and head of the town market.

Functions of the *tlatoani*

Many of these functions were delegated to a large staff of professional administrators: priests, market supervisors, military leaders, judges, tax collectors, and accountants. The tax collectors, or *calpixque,* were especially important administrators because they acted as the rulers' agents in collecting goods and services from the calpulli chiefs.

Most of these positions were appointed and selected from two classes—the *pipiltin* (plural of *pilli*), and the professional warriors. Society was divided into three well-defined castes. At the top were the *pipiltin,* nobles by birth and members of the royal lineage. Below them was the *macehual* class, the commoners who made up the bulk of the population. At the base of the social structure were the *mayeques,* or serfs, attached to private or state-owned rural estates. Within these three castes, a number of social classes could be differentiated, according to wealth, occupation, and political office. The Aztec system made a distinction between ascribed and achieved status. By a

The three social castes

system of promotions, usually as a reward for military deeds, commoners were appointed to such political offices as *calpixque* and judges. Many *pipiltin* held no political office and, unless they had inherited private estates, were forced to live off the largess of the ruler. Commoners who had captured four enemy warriors in combat were promoted to the rank of *tecuhtli,* entered one of the military orders, were assigned a private estate with serfs for their maintenance, and acted as an elite professional army. The children of both *pipiltin* and *tecuhtli* could enroll in the religious college, or calmecac, where they could be trained as priests or political administrators. The calmecac apparently was also open to certain other commoners, such as wealthy and influential merchants and craftsmen.

Aside from the commoner-warriors, the *macehual* class was further differentiated into class levels. Certain occupations were accorded higher prestige than others (merchants, lapidarians, goldsmiths, and featherworkers are mentioned, and the list probably included stone sculptors); and all urban occupations were assigned higher status as compared with rural farming. Since occupations were restricted to calpulli membership and since the calpulli were kin groups, it follows that crafts tended to be hereditary. In small towns the craft specializing group would have to be the *barrio pequeño.* In the cities it was definitely the larger unit, but in either case crafts would be found within hereditary corporate groups.

The system of social stratification emphasized ascribed status but also permitted considerable vertical mobility. The land-tenure system was an important aspect in maintaining both processes, as could be expected in a basically agrarian society. Although most of the land was held in common by the calpulli, private estates with serfs helped to maintain the prestige of the *pilli* class and similar estates assigned to political office; and the *tecuhtli* positions freed able commoners from the necessity of subsistence procurement.

The taxation system also helped to maintain the social system. All heads of households owed military service to the *tlatoani.* For the *pipiltin* and *tecuhtli,* this was the only tribute demanded. Urban craftsmen also paid tribute in their craft products but were exempt from corvée labour. That obligation, plus taxes in agricultural products, were the burdens of the rural peasants, and the *mayeque*s owed their labour and agricultural produce to their overlord.

Pawns and slaves

Two other elements in the Aztec social system were pawns and slaves. The former were poor men who could sell themselves or members of their household for a specified period of time. Their rights were carefully defended by Aztec law, and they were not slaves but more like indentured servants. True slaves did exist and in some parts of Meso-America were used as workers or servants. Among the Aztec, the *mayeque*s were their counterpart. Slaves were bought in lowland markets and used primarily for human sacrifice.

The high development of craft specialization—much of it full-time—in Aztec towns has been noted above. But many rural communities also had part-time specialities, a feature due in part to the heterogeneity of the highland environment, with its highly local distribution of resources. Foreign goods were brought into the Aztec homeland by great caravans of professional merchants called *pochteca,* who frequently undertook journeys exceeding a year in length. As a group the merchants enjoyed very high prestige and even had their own tribunals. Various merchant wards of a great number of towns and cities in central Mexico were organized into one great trading guild that had its centre at Tenochtitlán. They also organized and administered the town markets, another highly evolved aspect of Aztec institutions. These markets were held in great open plazas—in smaller towns every fifth day, in larger towns and cities daily, although in the latter case the market population reached a peak every fifth day.

The centres and the political organization of large states such as the Aztec empire were fundamentally similar in character to small ones; but the vast differences in size (Tenochtitlán, the Aztec capital, may have had 140,000 to 200,000 inhabitants in 1519) demanded some changes. Generally, when one central Mexican state conquered another, the ruler of the conquering town extorted an annual tribute, but there was little attempt at political integration. In the case of the Aztec, this policy was generally maintained, but many conquered states were given Aztec governors. Furthermore, conquest was usually accompanied by an exchange of women from the two ruling lineages (conqueror and conquered), and successors to the throne of the conquered states were through these women, from the royal lineage of Tenochtitlán. As a result, the ruling class gradually tended toward a single kin group. Because of the great number of states conquered by the Aztec (400 to 500), some form of intermediate-level territorial and administrative organization became imperative. The states conquered by the Aztec were grouped into 38 provinces. One town in each province served as capital, and an Aztec tax collector-governor was placed there to supervise the collection, storage, and disposition of the tribute. In many provinces, the Aztec established garrisons. These consisted of warriors and their families culled from all of the towns of the Valley of Mexico, and they were assigned lands in the conquered province. Since they supported themselves, they were colonists as well as troops. The planting of colonists, combined with such factors as the merchant guild and royal family intermarriage, suggests that the Aztec elite were attempting to integrate more closely the population of the Valley of Mexico as a kind of core nationality for the empire. Other indications that the Aztec were in the process of achieving further political integration are statements in several *relaciones* that the tax collectors served as courts of appeals in serious judicial cases and also that the Aztec introduced the cult of their national god Huitzilopochtli to conquered provinces.

Organization of the provinces

Tenochtitlán. Tenochtitlán itself was a huge metropolis covering more than five square miles. It was originally located on two small islands in Lake Texcoco; but it gradually spread into the surrounding lake by a process, first of chinampa construction, then of consolidation. It was connected to the mainland by several causeway dikes that terminated in smaller lakeside urban communities. The lake around the city was also partly covered with chinampas with numerous rural settlements. Together, the complex of settlements—the city, the chinampa villages, and the settlements along the lakeshore plain—must have appeared from the air as one gigantic settlement. The population in 1519 was in the neighbourhood of 400,000 people, the largest and densest concentration in Meso-American history.

The majority of the people in the city were non-food-producing specialists; *i.e.,* craftsmen, merchants, priests, warriors, and administrators. In Tenochtitlán, as in other larger towns, the larger calpulli formed craft guilds. Guild organization was internally complex, an economic development related to the higher level of political integration and the greatly expanded trade and tax base that accrued from it. The great market in the barrio of Tlatelolco was reported by the Spaniards to have had 60,000 buyers and sellers on the main market day. The Spaniards also described the enormous canoe traffic on the lake moving goods to the market. There is even evidence that many chinampa cultivators, in response to the expanded market, were shifting from the production of staple crops to truck gardening.

The Aztec capital was originally two separate cities, Tlatelolco and Tenochtitlán, which merged into one through the conquest of Tlatelolco. The division was maintained for administrative purposes, however, and with further growth it became necessary to divide Tenochtitlán into four great wards (also referred to as calpulli). Each ward contained 12 to 15 calpulli, some 50 to 60 in all. Tlatelolco must have had 10 to 20 calpulli as well, bringing the total up to perhaps 80.

With this enormously expanded tax base, the central government became internally complex. The Spaniards described the palace of Montezuma II as containing 300 rooms grouped around three courts. Land titles dating from after the conquest give it an area of 10 acres. Aside from the private apartments of the king, the palace included libraries, storehouses, workshops for royal craftsmen, great halls for justice and other councils, and offices

Montezuma's palace

for an army of accountants. The sources even describe a royal zoo and aviary and a number of country retreats. The internal organization of the taxation, military, and judicial departments must have been far more complex than in small states; but precise data is lacking.

Within the city there were literally hundreds of temples and related religious structures. There were at least two large complexes, religious centres of the dual cities of Tenochtitlán and Tlatelolco. Each of the four great wards of Tenochtitlán, as well as each calpulli, had smaller temple complexes, so that the total number must have run into the hundreds. The great temple complex of Tenochtitlán consisted of three large pyramid temples (the principal temple platform, dedicated to Huitzilopochtli and Tlaloc, was 100 feet high and measured 300 feet on a side at its base). There were also six small pyramid temples, three calmecac buildings (dormitories and colleges for priests), a ball court, a great wooden rack for the skulls of sacrificed victims, a sacred pool, a sacred grove, and several large open courts. All of these structures were placed within a vast walled enclosure, 1,200 feet on a side. The temple complex at Tlatelolco was at least half as large.

Aztec religion. Perhaps the most highly elaborated aspect of Aztec culture was the religious system. The Aztec derived much of their religious ideology from the earlier cultures of Meso-America or from their contemporaries. This was particularly true during the final phase of their history, when their foreign contacts broadened. Indeed, much confusion about Aztec religious ideology stems, in part, from the fact that Aztec civilization was still in a process of assimilation and reorganization of these varied religious traditions. Moreover, as the empire expanded and Tenochtitlán evolved into a heterogeneous community, the religious needs correspondingly changed from those of a simple agrarian society. The ruling class, particularly, demanded a more intellectual and philosophical ideology.

The great ceremonial rites

The Aztec approach to contact with the supernatural was through a complex calendar of great ceremonies, which were held at the temples and were performed by a professional priesthood that acted as the intermediary between the gods and human beings. Many of these were public in the sense that the populace played the role of spectators. Elements in all the ceremonies were very similar and included ritual ablutions to prepare the priests for the contact; offerings and sacrifices to gain the gods' favour; and theatrical dramas of myths by masked performers in the form of dances, songs, and processionals. Each god had his special ceremony that, considering the richness of the pantheon, must have filled the calendar. These ceremonies must have played a significant recreative function, as do ceremonies held in honour of patron saints in present-day Mexico.

Aztec religion heavily emphasized sacrifice and ascetic behaviour as the necessary preconditions for approaching the supernatural. Priests were celibate and were required to live a simple, spartan life. They performed constant self-sacrifice in the form of bloodletting as penitence (by passing barbed cords through the tongue and ears). This pattern of worship reached its climax in the practice of human sacrifice; it was in this aspect of Aztec culture that religion, war, and politics became closely related. Ideologically at least, Aztec warfare was waged for the purpose of obtaining sacrificial victims. The tribute lists, of course, demonstrate that there was a more mundane purpose as well, and it would be a serious mistake to think of Aztec warfare as functioning primarily in the religious sphere.

The priesthood

The cult of the gods required a large professional priesthood. Spanish documents indicate that the priesthood was one of the most elaborate of Aztec institutions. Each temple and god had its attendant priestly order. At Tenochtitlán the high priests of Tlaloc and Huitzilopochtli served as heads of the entire priestly organization. Within the orders were priests in charge of ceremonies, of the education of novices, of astrology, and of the temple lands. (These consisted of specific rural communities assigned by the state to particular temples.) Furthermore, there were several grades of priests. As noted above, the priests maintained a number of schools, or calmecacs, where sons of the nobility and certain commoners were given instruction. Most of the novices ultimately left the priesthood and carried out economic and political functions; others remained, joined the priesthood on a permanent basis, and lived at the calmecac.

Much of Aztec religion probably was practiced at home at special household altars. Common archaeological artifacts are small baked-clay idols or figurines, representing specific gods apparently used in these household ceremonies, along with incense burners. (W.T.Sa.)

Cosmogony and eschatology. The Aztec believed that four worlds had existed before the present universe. Those worlds, or "suns," had been destroyed by catastrophes. Humankind had been entirely wiped out at the end of each sun. The present world was the fifth sun, and the Aztec thought of themselves as "the People of the Sun." Their divine duty was to wage cosmic war in order to provide the sun with his *tlaxcaltiliztli* ("nourishment"). Without it the sun would disappear from the heavens. Thus the welfare and the very survival of the universe depended upon the offerings of blood and hearts to the sun, a notion that the Aztec extended to all the deities of their pantheon.

The myth of the "suns"

The first sun was called Nahui-Ocelotl, "Four-Jaguar," a date of the ritual calendar. Humankind was first destroyed by jaguars. The animal was considered by the Aztec as the *nahualli* ("animal disguise") of the creator god Tezcatlipoca.

At the end of the second sun, Nahui-Ehécatl, "Four-Wind," a magical hurricane transformed all people into monkeys. That disaster was caused by Quetzalcóatl (the Feathered Serpent) in the form of Ehécatl, the wind god.

A rain of fire had put an end to the third sun, Nahuiquiahuitl, "Four-Rain." Tlaloc as the god of thunder and lightning presided over that period.

The fourth sun, Nahui-Atl, "Four-Water," ended in a gigantic flood that lasted for 52 years. Only one man and one woman survived, sheltered in a huge cypress. But they were changed into dogs by Tezcatlipoca, whose orders they had disobeyed.

Present humanity was created by Quetzalcóatl. The Feathered Serpent, with the help of his twin, Xólotl, the dog-headed god, succeeded in reviving the dried bones of the old dead by sprinkling them with his own blood. The present sun was called Nahui-Ollin, "Four-Earthquake," and was doomed to disappear in a tremendous earthquake. The skeleton-like monsters of the west, the *tzitzimime,* would then appear and kill all people.

Two deeply rooted concepts are revealed by these myths. One was the belief that the universe was unstable, that death and destruction continually threatened it. The other emphasized the necessity of the sacrifice of the gods. Thanks to Quetzalcóatl's self-sacrifice, the ancient bones of Mictlan, "the Place of Death," gave birth to men. In the same way, the sun and moon were created: the gods, assembled in the darkness at Teotihuacán, built a huge fire; two of them, Nanahuatzin, a small deity covered with ulcers, and Tecciztécatl, a richly bejeweled god, threw themselves into the flames, from which the former emerged as the sun and the latter as the moon. Then the sun refused to move unless the other gods gave him their blood; they were compelled to sacrifice themselves to feed the sun.

Cosmology. According to the Aztec cosmological ideas, the earth had the general shape of a great disk divided into four sections oriented to the four cardinal directions. To each of the four world directions were attached five of the 20 day-signs, one of them being a Year-Bearer (east, *acatl,* "reed"; west, *calli,* "house"; north, *tecpatl,* "flint knife"; south, *tochtli,* "rabbit"), a colour (east, red or green; west, white; north, black; south, blue), and certain gods. The fifth cardinal point, the centre, was attributed to the fire god Huehuetéotl, because the hearth stood at the centre of the house.

Above the earth, which was surrounded by the "heavenly water" (*ilhuicáatl*) of the ocean, were 13 heavens, the uppermost of which, "where the air is delicate and frozen," was the abode of the Supreme Couple. Under the "divine earth," *teotlalli,* were the nine hells of Mictlan, with nine rivers that the souls of the dead had to cross. Thirteen was

considered a favourable number, nine extremely unlucky.

All of the heavenly bodies and constellations were divinized, such as the Great Bear (Tezcatlipoca), Venus (Quetzalcóatl), the stars of the north (Centzon Mimixcoa, "the 400 Cloud-Serpents"), the stars of the south (Centzon Huitznáua, "the 400 Southerners"). The solar disk, Tonatiuh, was supposed to be borne on a litter from the east to the zenith, surrounded by the souls of dead warriors, and from the zenith to the west among a retinue of divinized women, the Cihuateteo. When the night began on the earth, day dawned in Mictlan, the abode of the dead.

Vegetation and fertility gods

Deities. The ancient tribes of central Mexico had worshiped fertility gods for many centuries when the Aztec invaded the valley. The cult of these gods remained extremely important in Aztec religion. Tlaloc, the giver of rain but also the wrathful deity of lightning, was the leader of a group of rain gods, the Tlaloques, who dwelt on mountaintops. Chalchiuhtlicue ("One Who Wears a Jade Skirt") presided over fresh waters, Huixtocíhuatl over salt waters and the sea. Numerous earth goddesses were associated with the fertility of the soil and with the fecundity of women, as Teteoinnan ("Mother of the Gods"), Coatlicue ("One Who Wears a Snake Skirt"), Cihuacóatl ("Serpent-Woman"), and Itzpapálotl ("Obsidian-Butterfly"). Their significance was twofold: as fertility deities, they gave birth to the young gods of corn, Centéotl, and of flowers, Xochipilli; as symbols of the earth that devoured the bodies and drank the blood, they appeared as warlike godheads. Tlazoltéotl, a Huastec goddess, presided over carnal love and over the confession of sins.

Xipe Totec, borrowed from the faraway Yopi people, was a god of the spring, of the renewal of vegetation, and at the same time the god of the corporation of goldsmiths. Human victims were killed and flayed to honour him.

The concept of a supreme couple played an important role in the religion of the old sedentary peoples such as the Otomí. Among the Aztec it took the form of Intonan, Intota ("Our Mother, Our Father"), the earth and the sun. But the fire god Huehuetéotl was also associated with the earth. In addition, Ometecuhtli ("Lord of the Duality") and Omecihuatl ("Lady of the Duality") were held to abide in the 13th heaven: they decided on which date a human being would be born, thus determining his destiny.

Among the fertility gods are to be counted the "400 Rabbits" (Centzon Totochtin), little gods of the crops, among which are Ometochtli, the god of *octli* (a fermented drink), and Tepoztécatl, the god of drunkenness.

The Aztec brought with them the cult of their sun and war god, Huitzilopochtli, "the Hummingbird of the Left," who was considered "the reincarnated Warrior of the South," the conquering sun of midday. According to a legend probably borrowed from the Toltec, he was born near Tula. His mother, the earth goddess Coatlicue, had already given birth to the 400 Southerners and to the night goddess Coyolxauhqui, whom the newborn god exterminated with his *xiuhcoatl* ("turquoise serpent").

Tezcatlipoca, god of the night sky, was the protector of the young warriors. Quetzalcóatl, the ancient Teotihuacán deity of vegetation and fertility, had been "astralized" and transformed into a god of the morning star. He was also revered as a wind god and as the ancient priest-king of the Toltec golden age: the discoveries of writing, the calendar, and the arts were attributed to him.

Mythology of death and afterlife. The beliefs of the Aztec concerning the other world and life after death showed the same syncretism. The old paradise of the rain god Tlaloc, depicted in the Teotihuacán frescoes, opened its gardens to those who died by drowning, lightning, or as a result of leprosy, dropsy, gout, or lung diseases. He was supposed to have caused their death and to have sent their souls to paradise.

Quauhteca and Cihuateteo

Two categories of dead persons went up to the heavens as companions of the sun: the Quauhteca ("Eagle People"), who comprised the warriors who died on the battlefield or on the sacrificial stone, and the merchants who were killed while traveling in faraway places; and the women who died while giving birth to their first child and thus became Cihuateteo, "Divine Women."

All the other dead went down to Mictlan, under the northern deserts, the abode of Mictlantecuhtli, the skeleton-masked god of death. There they traveled for four years until they arrived at the ninth hell, where they disappeared altogether.

Offerings were made to the dead 80 days after the funeral, then one year, two, three, and four years later. Then all link between the dead and the living was severed. But the warriors who crossed the heavens in the retinue of the sun were thought to come back to earth after four years as hummingbirds. The Cihuateteo were said to appear at night at the crossroads and strike the passersby with palsy.

Worldview. The world vision of the Aztec conceded only a small part to man in the scheme of things. His destiny was submitted to the all-powerful *tonalpohualli* (the calendrical round); his life in the other world did not result from any moral judgment. His duty was to fight and die for the gods and for the preservation of the world order. Moreover, witchcraft, omens, and portents dominated everyday life. That such a pessimistic outlook should have coexisted with the wonderful dynamism of Aztec civilization is in itself a remarkable achievement.

Aztec ritual calendar. Tonalpohualli, an Aztec term meaning "the count of days," was the name of the ritual calendar of 260 days. It ran parallel to the solar calendar of 365 days, which was divided into 18 months of 20 days and five supplementary unlucky days. The word *tonalli* means both "day" and "destiny": the 260-day calendar was mainly used for purposes of divination. The days were named by the combination of 20 signs—natural phenomena such as wind and earthquake, animals like rabbit and jaguar, plants such as reeds, and objects like flint knife and house—with the numbers 1 to 13. Thus the calendrical round included 20 series of 13 days.

Astrology and divination

Specialized priests called *tonalpouhque* interpreted the signs and numbers on such occasions as birth, marriage, departure of traders to faraway lands, and election of rulers. Each day and each 13-day series were deemed lucky, unlucky, or indifferent according to the deities presiding over them. Thus Ce-Coatl ("One-Snake") was held as favourable to the traders, Chicome-Xochitl ("Seven-Flower") to the scribes and the weavers, and Nahui-Ehécatl ("Four-Wind") to the magicians. The men who were born during the Ce-Ocelotl ("One-Jaguar") series would die on the sacrificial stone, those whose birth took place on the day Ometochtli ("Two-Rabbit") would be drunkards, and so on. The *tonalpohualli* dominated every aspect of public and private life. (Ja.S.)

ANDEAN CIVILIZATION

For several thousand years before the Spanish invasion of Peru in 1532 a wide variety of high mountain and desert coastal kingdoms developed in western South America. The extraordinary artistic and technological achievements of these people, along with their historical continuity across centuries, have encouraged modern observers to refer to them as a single, Andean civilization.

A look at a modern map reveals that no single South American state encompasses all of the territories controlled by the Inca (Inka) before the coming of the Spanish; rather these territories were spread over parts of Ecuador, Peru, Bolivia, Chile, and Argentina, and in 1532 they were all part of a single Inca state called Tawantinsuyu, the "Realm of the Four Parts." Earlier, local hegemonies—some coastal, others centred in the mountains, and still others bridging these geographic barriers—had risen, expanded, and eventually collapsed.

The Inca of Cuzco (Cusco) were themselves newcomers to most of the regions that they came to dominate. Such rapid expansion did not allow for complete consolidation; and the Spanish were able to take advantage of what had been a recent incorporation of numerous regional ethnic



Principal sites of Andean civilization.

groups and the resentments that the Inca victory had created among the ethnic lords. Some of these, like Don Francisco Cusichaq, lord of Xauxa, the earliest colonial capital, lived long enough after 1532 to testify before a Spanish court of inquiry that he regretted having opened the country to the Europeans. For 30 years his bookkeepers had recorded on their knotted quipu (*khipu*) accounts not only everything the Spanish had received from Xauxa warehouses but also, on separate knot-strings, everything that had been considered stolen.

The outsider visiting the Andes perceives two overwhelming geographic realities: the Pacific coastal desert stretching for thousands of miles and the high Andes rising parallel to the coast. These contrasting regions—utter desert on the coast and high, looming mountains to the east (where the bulk of the pre-Columbian population lived above 10,000 feet [3,000 metres])—could, and at several times in Andean history did, coalesce into a single political entity. Thus it is possible to speak of a single Andean civilization, even if at times, early and late, there was no political integration. One indicator of this social unity is extant even now: Quechua, one of the Andean languages, is still spoken by some 10,000,000 people from northern Ecuador to northern Argentina, a distance of thousands of miles.

The nature of Andean civilization

The coastal desert was inhabited for millennia by fishermen, and many of their settlements have been studied by archaeologists. The people in these communities were familiar with the sea and depended heavily on its products, but from very early times they also used and possibly cultivated native varieties of cotton. Textiles have been the major art form in the Andes for thousands of years. It is known that these textiles—found preserved in the coastal sands—have woven into them a wealth of information on Andean peoples; and, while the information in the textiles still cannot be read, it is believed that they will eventually be as revealing as have been the Meso-American codices.

The coastal desert

In modern Peru irrigation eventually may permit the cultivation of the lower reaches of most rivers. Still, it is useful to note that of some 50 rivers descending from the Andean glaciers to the Peruvian coast, only three have water flowing through them year-round. Such an ambitious irrigation scheme would be most productive only if the waters were tapped quite high on the western slope and if several rivers were connected through canals high in the Andes, thus allowing the scarce waters of three or four valleys to be pooled into a single one as needed. Rumours of such a project reached the first Spaniards in Peru: in the final decade before the invasion, the Inca were said to be planning to bore through a mountain in what today is northern Chile, so that water from the Amazonic watershed would flow westward to the deserts and thus alter the continental divide.

Archaeologists, particularly non-Peruvian scholars, have concentrated on the study of coastal peoples: they have found that sites are more accessible along the Pan-American Highway; that the hot and dry climate presents none of

the challenges of the high altitudes; and that the remains, mummified in the desert sands, are immediately rewarding. Pottery finds have portrayed such things as fishing or warfare, diseases, weapons, cultivated plants, and differences in rank and in sexual habits among the Andeans. Usually this evidence has been recovered by professional grave looters but sometimes also by archaeologists themselves. One of the most remarkable of the latter type of finds is the grave of a Moche leader that was discovered near the village of Sipan on the northern coast of Peru in the mid-1980s. Since the mid-20th century architectural studies of ceremonial and political centres have allowed researchers to follow changes in the location and the architectural features of important Andean cities. Distance from the sea and the degree of dependence on maritime products, the proximity to irrigation waters from the highlands, and the repeated efforts to control militarily more than a single irrigated valley have all received attention from archaeologists.

A major question remains: did these coastal polities extend upward to the Andean highlands to control areas beyond the slopes where the irrigation works tapped the rivers? The Peruvian historian María Rostworowski has pointed to similarities, found in colonial administrative papers, between coastal places-names and personal names in the Cajamarca Highlands, an area due east and above the coastal political entities. The colonial papers have not explained the presence of such distant colonies, but they have introduced a topic fundamental to understanding Andean success: given the apparently inhospitable environments of both the desert coast and the nearby high Andes, how could so many separate societies have fed such enormous populations and constructed highways, palaces, and temples in what were clearly urban centres for so many centuries?

AGRICULTURAL ADAPTATION

One answer to this question was suggested in the 1930s by the German geographer Carl Troll. His solution took into account a unique aspect of Andean ecology: the greatest population concentration (more than 1,000,000 people) and the highest agricultural productivity occurred around Lake Titicaca, which is some 12,500 feet above sea level. Nowhere else in the world—not even in Tibet or Nepal—has cultivation been so successful at such a high altitude. The effort to understand the ramifications of this paradox is far from complete, but Troll's insights have proved fertile: (1) The fields and terraces clustered around the lake were located just a few degrees south of the Equator, where daytime temperatures are truly tropical. (2) At this altitude climatic contrasts are not so much seasonal as diurnal, *i.e.*, summer by day and winter by night. Contrasts of 55° to 70° F (30° to 40° C) within a single 24-hour period are not uncommon, and nearly 300 nights of frost per year have been recorded on the high, windy plateau (puna) surrounding the lake. (3) Populations settled in such circumstances seem to have endured as others have survived in the Arctic, the Kalahari, and the Gobi, but it is clear that in the Andes a far denser population fared much better than have groups in other environmentally harsh regions, acquiring with time an intimate familiarity with the agricultural and pastoral possibilities of high altitude.

Tuber cultivation

These peoples cultivated many varieties of tubers, of which only the potato has achieved widespread use in the world. But since the soils at this altitude were easily exhausted, "second- and third-year" tubers had to be domesticated to take advantage of the nutrients left unused in the soil. Then, as now, it was usual to allow the ground to rest—for six, eight, or even 10 years—after which some of the "rested" acreage was returned to cultivation annually, a rotation pattern that is still familiar to the local people.

The upper elevation limit of cultivation has varied throughout the centuries, as the climate has fluctuated. Thus, considerable effort was invested in the development of ever more frost-resistant varieties of tubers. Modern observers have noted that tubers grown close to and above about 13,000 feet were mostly of the pentaploid varieties, bitter hybrids resulting from selection and crossing by the grower. Although they usually required additional nurture and processing that were beyond the procedures familiar today, the bitter varieties represented a gain in total productivity.

Terracing

A significant improvement in agriculture was the construction of massive terraces, which not only extended the cultivated area but also created protected microclimates where particular varieties could flourish. It has been suggested that an "amphitheatre" found in the Cuzco region was actually an experimental field where the concentric terraces reproduced tiny variations in the upland environment. When the use of highland irrigation and raised-ridged fields are taken into account, it becomes clear that these upland populations were highly familiar with, and respectful of, the potential for high-altitude agriculture and were intent on gaining additional acreage in circumstances that elsewhere would not have seemed worth the effort.

Another incentive for settlement at high altitudes was the presence of glacier-fed pastures for alpaca herds. The llama—it and the alpaca were the two camelids domesticated by the Andean peoples—could live at altitudes ranging from sea-level to those in the high mountains. The alpaca's habitat, however, was much narrower; it did best above 13,000 feet, and its preference for a swampy range was catered to by pastoralists. It has been found that even today alpaca-herding is a full-time occupation, almost impossible to combine with agriculture. While Andean herders did belong to wider ethnic groups, they tended to be specialists, relying for their food staples on their kinsmen closer to Lake Titicaca.

Present-day distribution and use of these animals (known collectively as camelids) tends to mask their importance in pre-Columbian times. A European inspector, reporting in the 1560s on the camelid wealth of a single Aymara chiefdom near Lake Titicaca, claimed, "I have heard of an Indian who is not even a lord, one don Juan Alanoca of Chucuito, who has more than 50,000 head." Such control of vast herds, combined with the hundreds of varieties of high-altitude tubers and grains, helps to explain the density of Andean populations.

THE COLD AS A RESOURCE

Beyond such skilled manipulation of the natural geography there lay an awareness of frost. As noted above, in the high Andes frost can occur almost every night of the year. Elsewhere people have endured the cold; in the Andes the cold was transformed into a positive and even creative factor.

Food preservation and storage

It is not known when this step was taken. For at least 1,000 years people in the Andes have been aware that the sharp alternation between tropical noon and arctic midnight can be utilized. Any animal or vegetable tissue exposed to this daily contrast can be processed into nutritive products that keep for decades, and the process can be achieved either at the household or the state level.

Chuño is the name popularly used for processed tubers, but a rich vocabulary for tubers exists in the Quechuan (Andean) languages: there is a separate term for each plant and for each mode of preparation. Chuño cannot be made where a diurnal temperature extreme is absent; thus, north of modern Cajamarca in northern Peru no chuño is prepared, since nocturnal frosts are rare or absent. Animal tissues also can be handled in this manner. After 1532 European meats were added to those of local birds, fish, and camelids. The name for these preserved meats is charqui, or jerky (*ch'arki* in Quechua), the one Andean word that has made its way into common English usage.

Such food reserves allowed both the peasants and the state to compensate for natural and man-made calamities. They filled thousands of warehouses—many of which are still extant—that were built in ways and places so as to use the tiny differences of exposure to the sun, winds, and humidity. Those built by the state or by the ethnic lords along the more than 15,500 miles of roads provided food for both human and camelid porters, for the armies, and for priests traveling to the many shrines.

The presence of such large stores made possible the incredible forays of Spaniards like Diego de Almagro, who reached Chile from Cuzco across thousands of miles of deserts and snow-covered mountains. As late as 1547, 15

years after the Spanish invasion, one Spaniard, Polo de Ondegardo, reported that he had fed 2,000 soldiers for seven weeks with the food still stored above Xauxa, which had been the first European capital. A detailed archaeological study of an Inca storage system was made by the American anthropologist Craig Morris, who found almost 500 warehouses at Huánuco Pampa. There were some 1,000 warehouses at Xauxa and many more near Cuzco, the Inca capital.

THE HIGHLANDS AND THE LOW COUNTRIES

The cultivators of high-altitude tubers and lowland crops—the plants of which seem botanically far apart at first glance—were actually in continuous contact. This point was stressed by the pioneer Peruvian archaeologist Julio C. Tello and was later verified by foreign scholars. The inhabitants all along the Andean highlands were aware of the diverse populations and climates of the Pacific coastal deserts to the west and of the Amazon lowlands to the east. The Chilean researcher Lautaro Núñez has traced the several societies who inhabited a single valley: products and settlement patterns changed through the centuries, but at all times each successive ethnic group accumulated resources from diverse ecological niches into a single system.

Dispersed settlement pattern

By adding written Spanish sources to the information provided by archaeologists, it is possible to explain further the density of the Andean population and its great productivity. Throughout the Andes, south of Cajamarca, political units large and small were characterized by a dispersed settlement pattern. The preferred location of the seat of power frequently was at very high altitudes, almost at the upper limit of cultivation, and kinsmen of these highlanders were settled permanently at three, five, or even 10 days' walk from the political centres. The German anthropologist Jürgen Golte has stressed that the agricultural calendar permitted such absences, since crops matured at different dates according to altitude; but many outliers were too far away from the political and demographic nucleus to permit seasonal migrations. The outlier communities could be large or small and could be established on the dry Pacific shore or in wet Amazonic enclaves. The Lupaca (Lupaqa), an Aymara-speaking polity whose political centre was located on the puna on the shores of Lake Titicaca, controlled outliers on both slopes.

Other ethnic groups reached in only one direction. For example, the two lords of the Karanga (Caranga), on what today is the highest part of the Bolivian High Plateau, do not seem to have controlled any outliers of their own on the Amazonic slope. Their main puna farms and most of their subjects lived above 12,000 feet, and their camelid herds were pastured even higher. The Karanqa also controlled corn (maize) fields at less lofty altitudes in what today is Chilean territory, several days' walk away. Farther west and closer to the coast were their fruit and coca-leaf gardens. Finally, even farther north, across the Atacama Desert near the modern city of Arica, the Karanqa had their "own" fishers.

One unexpected feature of such outliers is that they were usually multiethnic: several political centres shared settlements of salt miners, fishers and seaweed gatherers, cultivators of hot peppers and coca leaves, and timber cutters and honey gatherers. The political mechanisms by which conflicting groups could reach truces, even if temporary, or the means by which caravans moved with safety when connecting the central settlements with their multiple outliers are still not known.

This diverse pattern of settlement and political control and of pooling dispersed resources and populations has been named "Andean ecological complementarity," or the "vertical archipelago." Such complementarity went beyond the efficient control of the nocturnal cold and of the high altitude. Even if many details of how it worked still escape understanding, it is obvious that each ethnic group was able to diversify the risks that would have existed if each had been concentrated in any of the separate Andean ecological tiers. Beyond defensive strategies, in ecological complementarity it is possible to detect new opportunities that would permit massive storage of a wide range of foods going beyond those grown locally. Eventually there emerged dense populations and large polities like the Inca. It is notable that the foci of Andean civilizations across the centuries—Chavín, Huari (Wari), Tiahuanaco (Tiawanaku), Cuzco—were all located on the high puna.

The pre-Inca periods

Archaeological periods

The names the several prehistoric populations called themselves are not known, and archaeologists have come to distinguish the various peoples and civilizations by descriptive terms—the Late Preceramic, the Initial (or Lower Formative) Period, the Early Horizon, the Early Intermediate Period, the Middle Horizon, the Late Intermediate Period, and the Late Horizon (also called the Upper Formative, or Inca, Period). Each of the periods lasted for centuries, some for millennia. These designations stress the differences between the peoples who inhabited the coast and those who lived in the highlands, although contacts between the two areas were frequent at all times in prehistory. What have been termed "horizons" in Andean studies were much shorter periods of time when wide areas of the central Andes were united culturally and politically with one another and with sections of the Pacific coast.

THE LATE PRECERAMIC

There is ample evidence of human occupation by 3500 BC, at which time there was already considerable diversity along the Pacific. In the central and northern coastal areas lived people who cultivated beans, squash, cotton, and chili peppers and who exploited the sea, catching fish with cotton nets and shell or composite hooks, collecting shellfish, and hunting sea mammals. One group at Chilca, south of modern Lima, built conical huts of cane thatched with sedge. The dead were buried wrapped in twined-sedge mats and the skins of the guanaco, a wild camelid. Some people camped in winter on the lomas, patches of vegetation outside the valleys that were watered at that season by fogs. In summer, when the lomas dried up, they built camps along the shore. The lomas provided wild seeds, tubers, and large snails; deer, camelids (probably guanaco), owls, and foxes were hunted. The lomas had long been shrinking, and the winter camps were abandoned (*c.* 2500 BC) in favour of permanent fishing villages. Nowhere are the deposits thick enough to show stratification, but they have been arranged in chronological order by comparing the implement types and noting their distribution within the shrinking patches of vegetation. Some small patches still survive.

In the far south, the lomas were and still are more extensive than in the centre, and projectile points are abundant in them and in caves in the valleys. Deer can still be seen on the lomas, and it appears that hunting of them and of guanaco was the main activity in Late Preceramic times.

In the far north, in the Talara region and extending north into Ecuador, are stone tools and mangrove-dwelling mollusks, left by people who enjoyed a wetter climate than that now prevailing, and one inland site at El Estero, provisionally dated somewhat earlier (*c.* 5000 BC), has well-made polished stone axes and mortars that indicate the exploitation of forests and grasslands yielding seeds.

Much longer periods of occupation have been postulated for the highlands: the American scholar Richard S. MacNeish has suggested a human presence as early as 15,000 BC in the Ayacucho Basin, which would correspond to the traditional "first wave" of immigrants into the New World. Since there has been much less research in the highlands than on the coast, little is known of the highland Late Preceramic. The caves at Lauricocha at about 13,000 feet in the central Andes, which had been occupied by deer and camelid hunters since nearly 8000 BC, were still used, at least as summer camps, by hunters who employed small leaf-shaped points. Gourds, squash, cotton, and lucuma, with seed plants such as quinoa and amaranth, were cultivated in the Ayacucho Basin before 3000 BC; corn and beans came within the next millennium. There were also ground stone implements for milling seeds. It has been claimed that llamas and guinea pigs long had been domesticated.

After about 2500 BC came a great increase in the speed

of development, which is best known on the coast. Population increased, and stable settlements were established in many places. By 2000 BC there were perhaps 100 villages on the coast with populations of 50 to 500 people, with a few of up to 1,000, indicating a total population of about 50,000. This was a far cry from the thinly scattered bands and occasional villages of about 1,000 years before. Considerable variation has been observed from place to place, but most sites have shown a predominance of seafood, including fish, shellfish, sea lions, and sea birds.

Population growth

On the north central coast, the stretch between the Casma and Huarmey rivers was heavily populated. One site, at Culebras, was a large village on a terraced hillside, with semi-subterranean houses whose underground parts were lined with basalt blocks and whose upper parts were built of lighter materials such as adobe blocks. They originally had hard clay floors, and some had guinea-pig hutches consisting of stone-lined tunnels connecting two rooms at floor level. The guinea pig, normally vegetarian, appears to have been taught to feed on small fish. A site at Huarmey has provided the earliest known instance of corn on the coast, and it also occurred in the top Preceramic levels at Culebras.

Burials at Culebras were tightly flexed, wrapped in twined mats and cotton cloth, and accompanied by gourd vessels and beads and pendants of stone, shell, or bone. The skulls of these people were deformed by having been bound to cradleboards in infancy. There was a cemetery, but many burials were under house floors. No ceremonial buildings are known in this area.

Farther north, at the mouth of the Chicama River, is Huaca Prieta, which was the first Preceramic site to be excavated. A thick midden, it contains some subterranean houses lined with cobblestones and roofed with earth supported by whalebones and wooden beams. The twined textiles found there were the vehicle for a peculiar art style, showing highly stylized crabs, double-headed snakes, birds, and human beings, expressed by warp manipulation designed to bring groups of warps of one colour to one face. The dyes have faded, and the only way to recover the designs is by examination under a microscope. Such textiles were not confined to this area, but they have been more fully studied there. Woven textiles were rare, and weaving was combined with twining in a way that shows that a loom was not used.

Ceremonial architecture

Unlike the area farther north, sites along the central coast had ceremonial buildings, of which the most remarkable is El Paraíso in the Chillón Valley. This is an imposing stone-built structure on an artificial mound, with a central stairway leading up to a group of rectangular rooms. The central block, which occupied a commanding position in a side valley, has been partly reconstructed, but there were extensive wings that may have been residential, though they now appear as little more than piles of stones. Floodwater farming may have been practiced there, but definite signs of it have been obliterated by modern cultivation. At Río Seco, a few miles to the north, are two pyramids, constructed by filling a group of preexisting rooms with boulders, building adobe-walled rooms on top of them, and finally filling these up also.

Apart from one site, Kotosh, near modern Huánuco in the central Andes, little is known of the highland final Preceramic. A Japanese research team has found structures of undressed stone chosen to present flat wall surfaces, set in mud, covering an area of at least 200 by 100 yards (180 by 90 metres), in some parts of which was a succession of buildings piled up to a considerable height. Among these were two superimposed temples, the lower being a rectangular structure on a stepped platform about 26 feet high. The floor was surrounded by a broad, low bench, and each outside wall had two or three rectangular niches. The walls and floor were covered with two coats of mud plaster, and beneath the central niche at one end was a pair of crossed forearms modeled in the same material. This temple was later buried in boulders surrounded by a retaining wall and covered by a new floor on which a second temple was built, of which little remains. The burial of the first temple to act as a raised foundation for the second recalls the construction at Río Seco.

THE INITIAL PERIOD

The next epoch, called the Initial Period by the American scholar John H. Rowe, and the Lower Formative by the Peruvian archaeologist Luis G. Lumbreras, began with the introduction of pottery. The earliest ceramics have yielded radiocarbon dates of about 1800 BC, although Rowe has suggested that even a date of 2100 BC is plausible. Ceramics from this period have been found on the central coast between Las Haldas, in the Casma-Huarmey region, and Lima. These are considerably later than the earliest pottery finds at Puerto Hormiga on the northern coast of Colombia near Panama (before 3000 BC) and Valdivia in Ecuador (*c.* 2700 BC). The period ends with the spread of the Chavín cult (also called the Early Horizon; see below).

Lumbreras has stressed agriculture as a more telling indicator: while no single starting date has been cited for this achievement, beans may have been cultivated for centuries if not millennia before the date of the earliest pottery. Bottle gourds and squashes were other cultivated species. Potatoes and other tubers, so important in later times, did not keep well in highland circumstances; but some researchers believe that Andean peoples were reliant on wild tubers and rhizomes 10,000 years ago, although these groups had not yet domesticated them. It has been demonstrated that on the coast virtually all the crops that were important in 1532 (with the notable exception of corn) were already known and in daily use during the Initial Period.

The introduction of pottery at first made little difference to the general pattern of life; cooking continued to be done by roasting on hot stones. On the coast, there was a gradual increase in the consumption of cultivated plants, grown mainly in the lower reaches of the valleys; but the basic reliance on seafood continued. An important innovation was the development, or possibly the introduction, of the heddle loom, but, if it was introduced, its origin is not known. At first it seems only to have been used for making plain-weave cotton cloth. Village life and temple buildings spread over the country, except to the far south, where conditions favoured the continuance of hunting and gathering. Corn spread from the centre over most of the coast, and cassava, or manioc (an edible root), and peanuts (groundnuts) appeared there for the first time, their ultimate source being the Amazonian lowlands.

Technological advances in the Initial Period

New ceremonial centres showed considerable diversity. Examples include La Florida, a huge pyramid in Lima that formed the nucleus of a yet-unmapped building complex. The Tank site at Ancón consists of a series of stone-faced platforms on a hill. Las Haldas has a platform and three plazas; two smaller similar sites are also known. The old centres at El Paraíso and Río Seco had been abandoned, but, in the highlands, Kotosh continued to be occupied. Any constructions at Yarinacocha in a wet, stoneless area would have been of wood or other perishable materials.

The variety of the pottery suggests that it was derived from several different sources. In the Lima area, the earliest examples are neckless jars and incurved bowls with thickened rims and rounded bottoms, very uneven in shape. Some later types are pebble polished and the jars thinner. Other later types include bottles with straight spouts, which may have simple incised or applied decoration, and open bowls. Finally, as the period drew to a close, tan-coloured decorated wares, with punctate or red-painted areas outlined by incised lines, as well as orange ware with black stripes, were produced. A type found on the south coast is a small, double-spouted bottle with simple negative-painted decoration, the first appearance of a form long-lived in that area and of a decorative technique that later spread widely over the country.

In the highlands, the earliest pottery at Kotosh consists predominantly of simple bowls with somewhat constricted mouths, and bowls with gently rounded bases meeting the vertical to outsloping concave walls at a sharp angle. There are rare double-spout-and-bridge bottles, closed vessels with two tubular spouts connected by a solid bridge. The ware is mainly dark gray, black, or dark brown to sombre red, and it may have a red slip. The decoration—which was either applied to a broad zone covering most of the walls or, on the neckless jars, formed a ring around

the mouth—consists of linear incision, hatching, stamped circles, punctation, or excision. Postfired painting in red, yellow, and white frequently covers excised, hatched, and stamped areas. Despite the fact that Kotosh was on the eastern side of the Andean watershed, its pottery had little in common with that of Yarinacocha, save some similar decorations and the double-spout-and-bridge bottle.

The first known pottery of Yarinacocha is far from primitive. It consists mainly of bowls, mostly with complex outlines. Large open bowls with a broad labial flange, concave sides, and in some cases a second flange where side meets base, could have been cooking pots. Small bowls with inward-sloping sides meeting the rounded base at a sharp angle could have served for drinking; and a shallow bowl, with rounded base meeting the low, slightly outsloping concave sides at a lesser angle, may have been a plate for solid food. There are shards from large urns that may have served for brewing cassava beer. Decoration of finely hatched or cross-hatched geometrical areas, outlined by broad incised lines, occurs on most vessels, and one has a similarly executed feline face. In spite of severe weathering, postfired red paint, later so characteristic of the south coast, is found on some vessels.

THE EARLY HORIZON

The Early Horizon emerged after the appearance and rapid spread of the Chavín art style, ending the regional isolation of the Initial Period. The Chavín art style derives its name from the ruined temple complex of Chavín de Huántar in the Andean highlands of central Peru. The dates suggested for the emergence of the style beyond the environs of the temple, however, vary among scholars. Rowe dated it from 1400 BC, while Lumbreras suggested 850 BC; and the very designation of Chavín as a horizon has been challenged. But even those who have most favoured dropping the concept of horizon for this period have noted that in about 1000 BC there was an invasion of highlanders into the coastal Casma Valley who brought with them radically different architectural styles, ceramics, and food plants and animals that supplanted those in the valley; such a penetration was clearly a unification of the coast and the highlands into a single polity.

Beginnings of a unified art style

Chavín came to cover most of the north and centre of Peru, and its influence affected a good part of the south coast, excluding only the southern highlands. The art style, which is regarded as the expression of a cult, is expressed in painted textiles (of which few have survived), in pottery, and chiefly in stone carvings. Archaeologists at one time generally agreed that the chief object of worship was a cat, probably the jaguar, but this has been questioned, although many natural bird, animal, and human forms had feline mouths and other attributes. Feline representations were widespread, whereas some unquestioned deities were confined to the immediate neighbourhood of Chavín.

Most Chavín temples seem to have been ceremonial centres without people living around them, although the complex at Chavín itself seems to have been accompanied by a considerable town. The remainder appear to have been focuses for scattered settlements. The most elaborate

The temple at Chavín

temple known is that at Chavín, which lies at an elevation of 10,530 feet on a tributary of the Marañón River, east of the Callejón de Huaylas district of the Santa River. The temple consists of a group of stone platforms formed of rubble faced by walls of coursed masonry in which two thin courses alternate with one thick one. They are honeycombed with galleries running parallel to the walls at different levels and ventilated by shafts. The oldest part of the temple is a U-shaped structure, with the open top of the U facing east; the rectangular central arm contains a cruciform gallery, at the crossing of which stands a remarkable prismatic shaft of white granite, some 15 feet high, carved in low relief to represent a standing human figure with snakes typifying the hair and a pair of great fangs in the upper jaw.

This figure, which has variously been called El Lanzón, the Great Image, and the Smiling God, is thought to have been the chief object of worship in the original temple. The southern arm of the temple was subsequently twice widened by rectangular additions, into which some of the original galleries were prolonged. After the second addition, the two were joined by a freestanding facade having a central portal with a lintel supported on two cylindrical columns. The lintel bears 14 eagles in low relief, supplied with feline jaws with prominent fangs behind the beak, and each column is entirely covered by a mythical bird bristling with feline fangs and faces. These have been interpreted as attendants of the god worshiped in that part of the temple, who had perhaps superseded the Smiling God and could have been the god shown on the Raimondi Stone, now in Lima. The stone shows the Staff God, a standing semihuman figure having claws, a feline face with crossed fangs, and a staff in each hand. Above his head, occupying two-thirds of the stone, is a towering, pillarlike structure fringed with snakes and emerging from a double-fanged face, which Rowe interpreted as a symbolic treatment of his hair as a tongue coming out of a mouth. Unlike the Smiling God, this figure has been found in areas as far from Chavín as the northern and southern coasts of Peru. Except for the columns, which are of black slate, the stones of the facade are light in colour on the south side and dark on the north. East of the facade is a small sunken court of the same period, which contained a number of slabs with carvings in low relief, and to the east of this is a much larger court surrounded by platforms. Within this court is a square, slightly sunken area, in which was found the Tello obelisk, a rectangular pillar carved in low relief to represent a caiman and covered with Chavín symbolic carvings, such as bands of teeth and animal heads. This is considered to be an object of worship like the Smiling God and Staff God. Carvings found on and around the temple include a cornice of projecting slabs, on the underside of which are carved jaguars, eagles, and snakes, and a number of tenoned heads of men and the Smiling God; they are thought to be decorations or the attendants of gods rather than objects of worship.

On the coast, the temples were built mostly of adobe. In the Nepeña Valley, two temples—Cerro Blanco and Punkurí—differ so much that they must also differ in age, but it is not known which is the earlier. Cerro Blanco is a massive platform of conical adobes and stones, supporting rooms with walls bearing Chavín decoration, including eyes and feline fangs, modeled in mud plaster in low relief and painted red and greenish yellow. Punkurí has a low, terraced platform with a wide stairway on which stands a feline head and paws, modeled from stone and mud, and painted. By the paws was buried a woman, believed to have been sacrificed. At Moxeke and Pallca in the Casma Valley to the south, there are terraced, stone-faced pyramids with stone stairways. The first has niches containing clay-plastered reliefs of mud, stone, and conical adobes, showing felines, snakes, and human beings of Chavinoid character painted in polychrome. Also in Casma is a temple at Cerro Sechín, consisting of a series of superimposed platforms with a central stair, on either side of which, at the bottom level, stands a row of irregularly shaped flat stones with incised designs showing standing men carrying clubs, severed heads, and other designs. Lacking Chavín characteristics, these have been interpreted variously as ancestral Chavín or derived from it, the latter being the more plausible. There is a Chavín ceremonial centre at Garagay in the Chillón Valley but none to the south.

Chavín pottery

Chavín pottery is best known from the decorated types found in the galleries in the temple at Chavín and in graves on the northern coast, where it is called Cupisnique. Until the end of the period, the ware was monochrome—dull red, brown, or gray—and hard and stonelike. Vessels were massive and heavy, especially in the early part of the period. The main forms are open bowls with vertical or slightly expanding sides and flat or gently rounded bases, flasks, and stirrup-spouted bottles. The surface may be modeled in relief or decorated by incision, stamping, brushing, rouletting, or dentate rocker-stamping, all of which may be applied to particular zones in contrast with other smooth ones. Some bowls have deeply incised designs on both the inside and outside faces. Many of the forms and decorative features, apart from specifically Chavinoid designs (particularly feline fangs), were already present at Kotosh in the previous phase.

Considerable time changes are represented in Chavín pottery; for example, the earliest stirrup spouts were relatively small and very thick and heavy, and the spout had a thick flange. As time went on, the stirrups became lighter and the spouts longer; the flange was reduced and finally disappeared. The necks of the flasks underwent similar changes. The decoration on some of these is extremely striking; one has incised flower designs, and another has a roughened surface in which there are a number of concave circular depressions with a notably high polish. The Cupisnique stirrup-spouted vessels, some of which were modeled in the form of human beings, animals, or fruits, were the beginning of a north-coast tradition of naturalistic modeling, which persisted throughout its history. Toward the end of the period, a bichrome (dark red on cream) pottery came into use.

There is a considerable area on the south Peruvian coast with its focus in the Ica Valley, where strong influences from Chavín have been found in the Paracas pottery style, and two painted textiles in pure Chavín style have survived from the same valley. Paracas pottery was very different from that of Chavín, but various motifs have enabled the two to be correlated closely. Paracas began at practically the same time as Chavín, about 1000 BC, and lasted throughout its span and beyond it, perhaps to about 200 BC. The most characteristic form of Paracas pottery was a closed globular vessel with a somewhat flattened base, which had two narrow spouts connected by a flat bridge, or more frequently, with one spout replaced by a human or bird head. Simple round-based bowls were very common. The ware was most commonly black or very dark brownish, and much of the surface was covered with decoration outlined by incision and painted in polychrome with hard, shiny, resinous colours after firing. A panel bearing a feline face on one end of a spout-and-bridge vessel was one of the most frequent forms of decoration. Paracas is also distinguished for its gorgeous embroidered textiles, generally found in the mummy bundles of the important dead. Embroidery had a popularity at this time that it afterward lost, but a surprisingly wide range of weaving techniques were also used in various parts of the coast.

THE EARLY INTERMEDIATE PERIOD

The Early Horizon was succeeded by what has been termed the Early Intermediate Period. The onset of the Early Intermediate marked the decline of Chavín's cultural influence and the attainment of artistic and technological peaks in a number of centres, both on the coast and in the highlands.

The southern coast. The beginning of the period is best determined by the evolution of the Paracas pottery style into that of the Nazca (Nasca) area on the southern coast; this is traditionally estimated to have occurred about 200 BC, but Rowe's date of 400 BC is probably more reliable, since this is the area where his detailed succession was worked out. Nazca ware is marked by the introduction of slip painting applied before firing, which took the place of the resin painting applied afterward; but the style evolved continuously, and the polychrome tradition continued. The most common forms were bowls and beakers, all with rounded bases, but double-spout or head-and-spout jars were also characteristic. In contrast to the Moche area on the northern coast, figure modeling played a very minor role. Designs were painted in up to eight colours and fell into two main groups: one characterized by stylized but recognizable life forms, such as birds, fish, or fruits, with some humans; the other depicting mythical subjects such as complex demons. Between approximately middle and late Nazca, mythical figures became increasingly angular and elongated and developed a tangled mass of appendages. Trophy-head representations, which were modeled as complete vessels as well as painted in profile on simple vessels, increased greatly at the same time. Because Nazca art was less realistic than that of Moche, little can be learned of the appearance and life of the people.

Nazca pottery

In the time of the Nazca style what has been described as a small city was located in each of the south-coast valleys of Pisco, Ica, Nazca, and Acarí. At Cahuachi, in Nazca, this included a ceremonial centre consisting of six pyramids, which were terraced and adobe-faced natural hills associated with courts. Tambo Viejo in Acarí was fortified, which supports inferences drawn with some difficulty from late Nazca art that a concern with warfare developed at that time.

The northern coast. A cultural peak was reached in the valleys of Pacasmayo, Chicama, and Moche on the northern Peruvian coast. A large proportion of this area has been grouped by archaeologists into a Moche culture, although some of the territory encompassed by these valleys was not part of the polity called Moche. The Japanese archaeologist Izumi Shimada has referred to this kind of control as "horizontally discontinuous territoriality." The coast–highland "vertical" type of polity described above appears to have emerged later in coastal history. Thus, this "horizontal discontinuity" may have been related to coastal trade, as products were sought north of the desert coast, while at a later time it may have coexisted with "verticality."

The Moche culture

The Moche culture is distinguished by a ceremonial pottery style, commonly covered with a white or red-and-white slip, which may have had decoration painted on it, chiefly in red on the white parts. Some pots are molded in forms that include figures, animals, plants, and weapons; and some have molded designs in low relief. Molding and painting both convey highly realistic impressions of the people, things, and scenes they represent and are a vivid source of information about the life and activities of the people, though some important aspects, such as agricultural processes, are not represented. Moche pottery has been divided into five phases that were originally defined mainly by differences in the stirrup-spouted jars, but this has been extended to other forms—for example, bell-shaped bowls, double vessels, and jars with collars. The prevalence of stirrup spouts and the quality of the modeling connect Moche much more closely with Chavín–Cupisnique than with the intermediate styles, in which features such as the spout-and-bridge vessels suggest intrusive influences from the south. Among Moche buildings are adobe pyramids, like the enormous Huaca del Sol in the Moche Valley, palaces with large rooms (on terraces in the case of the Huaca de la Luna near the Sol), and fortified structures perched on the sides of valleys. These structures reinforce the evidence, provided by warriors and enthroned dignitaries depicted on pots, for the existence of an aggressive hierarchical state, and it may be inferred that this grew up as the result of dependence on highly developed irrigation systems in the restricted areas available in the valleys.

There were no towns in the northern valleys. Dispersed communities, built in places where they would not use the valuable irrigated agricultural lands seem to have been situated in ways suggesting dependence on one of the ceremonial centres.

Recuay pottery

The north highlands. In the north highlands, the remarkable pottery style of Recuay has been found in the Callejón de Huaylas region. This pottery is related to the negative-painted representative of Gallinazo in the Santa Valley and is painted with black negative designs over white and red, one of the most characteristic being a feline in profile with a comb on the head. There is a good deal of lively modeling, but it is much less naturalistic than that of Moche. A typical feature is a broad, nearly horizontal flange surrounding the mouth of a jar, and many jars also have a horizontal spout below the flange. Most of this pottery has come from stone-lined graves, and some stone buildings of two or three stories may have belonged to the people who made it.

The Cajamarca Basin is the site of a pottery style (called cursive) that was entirely independent of known outside influences and that spanned at least the Early Intermediate Period and the Middle Horizon. It has lightly painted running-scroll designs, which vaguely recall writing (whence the name cursive), as well as small animals and faces, in brownish black or red on a cream background, mostly on open bowls with ring bases. It was traded widely in the north, and south as far as Huari, during the Middle Horizon.

The south highlands. Large urban and ceremonial centres emerged at this time near the shores of Lake Titicaca.

Pucará

One site, Pucará, includes a well-built, horseshoe-shaped sanctuary of concentric walls of red sandstone enclosing a slightly sunken terrace lined with white-sandstone slabs. Within the terrace is a sunken court some 50 feet square and seven feet below the surface, also lined with white sandstone and reached by a stairway. This court contains two stone-lined grave chambers, and the outer horseshoe wall has small chambers, each containing one or two altarlike slabs in its thickness. There are also squat stone statues of men carrying trophy heads and stelae (upright sculptured slabs of stone) bearing recessed geometrical carvings and snakes. The pottery includes a reddish-buff micaceous ware painted in red, black, and yellow; cats, human or bird heads, and geometrical figures are all outlined by incision. The faces have the eyes divided vertically, one half of each eye black, the other half the natural colour of the ware. Pucará occurred early in the period, before the main development of Tiahuanaco, and it may have taken shape about 400 BC. It appears to have controlled an area between the site and Lake Titicaca or farther.

Tiahuanaco

Tiahuanaco is a well-known ceremonial centre whose stone remains are now a tourist attraction in the Andes second in popularity only to the ruins of Machu Picchu. The occupation of the ceremonial centre is believed to have begun very early in the period, since some of the earliest pottery shows similarity to that found at Pucará.

The ceremonial buildings—whose exact age is uncertain—include a large stepped pyramid or platform called Acapana (Akapana), with foundations of buildings on the top; a semi-subterranean temple with stone heads tenoned into the walls; and a low rectangular platform called Calasasaya (Kalasasaya), enclosed by a retaining wall of upright stones alternating with smaller rectangular blocks. In one corner of the platform stands a great monolithic doorway, not in its original position, cut from a large block of lava. At the top of the doorway is carved a central low-relief figure attended by three rows of smaller winged figures that appear to run inward toward him. The central figure, carrying staves that may represent a spear thrower and darts, has been likened to the Chavín Staff God and for convenience may be called the Doorway, or Gateway, God. Versions of the Doorway God and his attendants are found almost everywhere within the range of Tiahuanaco influence in the subsequent Middle Horizon. Another feature of the site is a number of large and finely finished stone blocks with niches, doorways, and recessed geometric decorations. Tiahuanaco masonry is sometimes held together by accurately cut notches, sometimes by copper clamps set in either straight or T-shaped grooves. Several massive monolithic statues have been found in and around Tiahuanaco, the largest being 24 feet high. They resemble pillars bearing relief designs and some carry beakers.

Decorated Tiahuanaco pottery is a finely polished polychrome, which commonly has a red slip with designs painted in various colours. Felines and hawks in profile, with eyes divided vertically into black and white halves, are common designs, as are geometric figures such as triangles and steps. Like all Tiahuanaco art, the designs are stiff and formal. The shapes include a tall, graceful, hollow-sided beaker, or *kero,* and various types of bottles and hollow-sided bowls with flat bases, including a form bearing a jaguar head and tail on the rim.

THE MIDDLE HORIZON

Both Pucará and Tiahuanaco were early forms of what became known as the Middle Horizon, an expansion of multiple-valley political rule that had two centres: one in the southern altiplano, the other centred on Huari (Wari), near the modern Peruvian city of Ayacucho. This development is usually dated around AD 600. Some Tiahuanaco effigy vessels have been discovered at Huari, but otherwise they seem to have been independent entities. Subsequent research has located parallel occupations near each other in the vicinity of the modern city of Moquegua.

Huari

The American archaeologist William H. Isbell has argued that Huari was a true state which displayed archaeological manifestations of administrative recording, had storage facilities on a scale suggesting major revenues, contained status tombs and palaces, and had other symbols and ornaments of a ruling class. Huari colonies and control also have been detected in the evidence. Attempts to explain what Huari and Tiahuanaco were doing outside the areas of their immediate control have pointed toward religious proselytism, as well as to trade. It has also been suggested that, although these polities employed an extensive form of control, they did not attempt to rule all of the intervening territories.

After a period of consolidation, the expansion was intensified, eventually reaching Cajamarca and the Chicama Valley in the north and the Ocoña Valley on the far southern coast, by about 800. The growth of the empire and its nature is shown by a number of features. One was the spread of Huari pottery styles and local copies of them, some bearing the Doorway God and other religious figures but many with neutral or secular motifs such as bands of chevrons. These generally were polychrome wares, and figures appearing on them—mythological, human, or animal—may have the eyes divided vertically into black and white halves, as at Tiahuanaco. A result of the increasing dominance of Huari styles was the obliteration of the old pottery styles over the whole coast from Nazca to Moche. The southern burial custom of huddled, cloth-wrapped mummies spread northward along the coast, displacing the older fashion of extended burial. The presence of large groups of storage buildings at Piquillacta in the Cuzco Valley and at Viracochapampa, near Huamachuco far to the north, suggests military activity like that of the later Incas. On the coast, some great cities in the north—of which Chan Chan, near modern Trujillo, is the best-known—originated at this time, apparently under southern influence, and the rectangular Great Enclosure Compounds in the Virú Valley may be an expression of the same phenomenon. All these changes, taken together, point strongly to military conquest.

Tiahuanaco designs, derived through Huari, are seen on coastal textiles as well as on pottery, and they are found particularly on tapestry. Besides recognizable figures like the Doorway God and his attendants, there are many examples, perhaps somewhat later in date, on which only the divided eyes—in black and white or other combinations of colours—can be inferred to belong to human or animal figures.

Pachacamac, on the central coast, which survived until Inca times as a great temple and oracle, was established as a ceremonial centre by the beginning of the Middle Horizon. At that time it also became a considerable town, with a degree of independence in the Huari empire, as is demonstrated by the presence of its own local variety of coastal Huari pottery—distinguished by the frequent depiction of a creature, sometimes called a griffin, with a winged feline body, human hands, and an eagle head, or sometimes the head alone—from Pachacamac north to Chicama, south to Nazca, and inland to Huancayo. Its influence may have been more religious than political, as in Inca times.

The Moche pottery style disappeared from the Chicama and neighbouring valleys under Huari pressure, but it is unlikely to have become entirely extinct because many features of it reappeared later on Chimú pottery. It probably survived, along with a remnant of the Moche state, in some valleys farther north (including perhaps Lambayeque), but the succession there has not been sufficiently worked out to demonstrate this.

The fall of Huari

When the Huari Empire reached its maximum extent, at about 800, it collapsed at the centre. Huari was abandoned, as it appears was Cajamarquilla, a large urban centre near Lima. Also at this time, it appears that construction peaked at Tiahuanaco—which is estimated to have had 5,000 to 10,000 inhabitants—although the city's influence on the region continued. Thereafter, few signs of urban life occur in the south, except at Pachacamac, until Inca times. Curiously, the decline of the cities in the south appears to have coincided with the beginnings of urban settlement on the northern coast at Chan Chan, Pacatnamú, and other places.

After Huari fell, signs of new influences from there dis-

appeared in the provinces, but various changes evolved in local pottery styles. Among these was the development of a new style on the north-central coast. One of the most distinctive products of this style was a face-collar jar, in many cases oval, decorated in pressed relief with cats and other Huari-derived designs and painted in washy black, white, and orange on a buff ground (Huari Norteño B). Other examples are the Epigonal styles of Nazca and Ica, characterized by bowls and flasks with occasional Huari motifs, such as animal heads, carried out in what has been described as "a slovenly, rounded and hasty" manner.

THE LATE INTERMEDIATE PERIOD

The Chimú state. The Late Intermediate Period began about 1000 (Rowe has said 900) with the dying out of the signs of unity imposed by Huari. The seeds of the Chimú state were probably sown at the same time, but they are not recognizable until considerably later. Elsewhere there were small independent states, which on the central and southern coasts were in most cases no bigger than a single valley, to judge from the distribution of the distinct pottery styles.

There were few new advances in techniques, and perhaps the most notable was the spread of bronze to the Peruvian coast from northwest Argentina and Bolivia, where tin ore was found and where the manufacture of bronze appears to have originated during the Middle Horizon. A hard alloy of copper and arsenic had been used previously in the centre and north. Pottery improved in quality in most areas, though its artistic character was not equal, for instance, to the earlier products of Moche and Nazca. There was a tendency toward standardization and toward reduction in the number of colours, which went with a degree of mass production. The modeling tradition of the north coast revived, but it was dull and lifeless by comparison with that of Moche times and was generally executed in black ware. In other parts, entirely new styles evolved. That of Chancay, on the central coast, was thin, dull red or cream in colour, with rather a dusty-looking cream slip and painted decoration in black. A common shape was an egg-shaped jar with a flaring collar and a pair of small loop handles, which were sometimes turned into a figure by modeling a face on the collar and adding ill-shaped limbs. Bowls and beakers with slightly bowed, almost vertical sides, were other common shapes. The porosity of the ware and the flaky nature of the slip made this pottery inferior in quality to that of other coastal areas. In the south, pottery of an attractive style was made in the Ica Valley. It was hard, well-burnished buff or red ware, covered with painted, textile-derived patterns in black, white, and red, although some vessels also depicted small birds and fish. Highly characteristic are bowls with a rounded base meeting the inward-sloping sides at a sharp angle and a thickened rim of triangular section.

Chimú textiles

It is difficult to determine whether any new textile techniques were adopted, but it is unlikely since the extreme versatility of the Peruvian weavers appears already to have covered most of the imaginable varieties at one time or another. On the other hand, fashions varied, and a relevant instance is the use of tapestry. Tapestry was known in the Early Horizon, suffered something of an eclipse in the Early Intermediate Period, and grew greatly in popularity in the Middle Horizon, when notable examples were produced. During the Late Intermediate Period its popularity waned again, although it was used for the finest garments on into the Inca Period; but in Chimú textiles it was generally confined to borders and other small areas. Textiles were similar over the whole coast, and to distinguish between those of different areas is a task for specialists. Some of the most characteristic types were garments adorned with regular rows, horizontal or diagonal, of stylized birds or fish, executed in brocade or double cloth.

In most of the northern valleys, irrigation systems reached their maximum extent; and the Virú Valley—which has been the most thoroughly studied—is deceptive in this respect, because much of the land went out of cultivation, possibly owing to removal of part of the population to other valleys by the Chimú rulers. In some cases, the systems of more than one river were connected, and water was taken, for instance, from the Chicama Valley to that of Moche, where Chan Chan was situated.

The legendary Chimú ruler Ñançen Pinco, who began to expand the state, is thought to have begun his reign about 1370, and the names of two predecessors are known; so it is a fair guess that the state was taking shape in the first half of the 14th century, when distinctively Chimú pottery forms appeared. Various rather exotic pottery styles dating before this time have been found in the northern area, but insufficient work has been done on their distribution in time and space. An early type consisted of bottles on a ring base with a loop handle and a narrow, tapering spout, decorated with geometric designs and cursive scrolls in black on a red-and-white ground. There were double whistling vessels, with modeled figures on one vessel connected to a tapering spout on the other by a flat, arched bridge; some early examples were of orange ware with a few dark-red stripes on the spout and bridge, but later ones were black. Carinated whistling vessels of black ware with hornlike projections above the ears on a ring base, with a tapering spout connected to a figure by a bridge, also have been found at an early stage. Similar vessels with two spouts connected by a bridge had a considerable range in time. Another form was a bottle, of black ware, with a tapering spout emerging from a modeled figure or head. These vessels had a strap handle and ring base of variable height. Many of the later blackware vessels had panels in pressed relief, a form of decoration that continued through Chimú times; these bore designs such as men holding staves, moons, or cats on a background of raised dots.

Chimú pottery

The overwhelming majority of Chimú vessels were of black ware. There was a revival of stirrup spouts, either on modeled vessels such as human figures or reed balsas, or on plain ones with or without pressed relief panels, and these normally had a monkey sitting on the stirrup at the base of the spout. Double vessels, often with a bird head to balance the spout and pressed relief panels on the bodies, continued. Many vessels were lentil-shaped. Jugs with strap handles and pressed relief panels sometimes took on a flattened section to become canteens. These are only a few of the forms that have been found, some resembling immediate predecessors, some new, and some, especially the stirrup spout, revivals of earlier types.

Ñançen Pinco is believed to have conquered the coast from the Saña River, just south of Lambayeque, south to Santa. After him came six rulers before Minchançaman, who conquered the remainder of the coast from at least as far north as Piura and possibly to Tumbes, south almost to Lima. His triumph was short-lived since he himself was conquered by the Inca in the early 1460s.

Chan Chan

The Chimú state originated in the Moche Valley, where its capital, Chan Chan, lay. There were other cities at Farfán and Pacatnamú in the Pacasmayo Valley and at Purgatorio and Apurlé in the Leche and Motupe valleys, respectively, which shared some features with Chan Chan. All included large walled compounds. Apart from the cities, there were defensive settlements, such as one in the narrow part of the Moche Valley, up which it straggled for five miles, occupying terraced hillsides and side valleys and commanding three of the main canals. A third type of settlement consisted of scattered compounds in the midst of large irrigated areas, one example of which was found in the Chicama Valley alongside an irrigation canal that took water to Chan Chan. Chan Chan covered an area of about 14 square miles (36 square kilometres), with a central area of about 2.5 square miles containing 10 or more large rectangular enclosures sometimes called *ciudadelas* ("citadels"). These were surrounded by tapering adobe walls, 10 feet thick at the base and about 30 feet high. They ranged in size from about 400 by 200 yards to 650 by 400 yards.

At least six of these citadels have similar plans, and one has been studied in detail. It has a narrow opening at the north end and is divided into three parts by high walls. The northern part contains a large entry court, flanked by a kitchen area and several smaller courts, leading to a densely built area of small courts, some of which have a U-shaped structure at one end, while others are filled

with small rooms. The U-shaped structures, which do not appear to have been roofed, may have been shrines, and the courts that contain them may have had walls covered with mud-plaster reliefs, such as bands of animals, birds, or fish, scrolls or step frets, arranged in a manner reminiscent of Chimú textiles. The central part has a somewhat smaller entrance court leading to several courts occupied by rooms, perhaps storerooms, although nothing was found in them. Another feature of this area is a great burial platform with rows of chambers arranged in three levels. All these features are connected by narrow and tortuous passages. The southern part is an open area, containing one or more *pukíos* (rectangular areas where the ground has been lowered to the water table, either to supply water or to grow plants). In the spaces between the enclosures, and elsewhere in the city, are large areas of dwellings, irrigated areas, and cemeteries.

It is now thought that the *ciudadelas* may have been the dwellings of the ruling classes and their immediate retainers, and it has even been suggested that they were the palaces of successive rulers, maintained by their descendants in the way that those of deceased Inca were maintained in Cuzco. The number of recognizable *ciudadelas* agrees with the number (10) of known Chimú rulers. This intriguing suggestion is further supported by the belief that the Inca learned a great deal from the Chimú after they conquered them, for, not content with carrying Minchançaman off to Cuzco, they established a colony of north-coast workmen there, and Topa Inca Yupanqui (Thupa 'Inka Yupanki) appears to have worked out the political organization of the empire at the same time, basing it largely on the Chimú system.

Roads between the valleys were always necessary to coastal states and were vital to the Chimú, and the Inca may have learned something in this connection also. There are almost continuous traces of a road from just north of Lambayeque to the Chao Valley just south of Virú, with remains even farther south in Santa, Nepeña, and Casma. The remains differ in elaboration and tend to be wider and more imposing near the cities; in the deserts between valleys they were tracks marked by posts or bordered by low walls, but in the valleys the simplest type is a leveled surface 15 to 25 feet wide, with walls of stone or adobe about three feet high and with the surface of the road sometimes being raised.

Although the Chimú had a powerful, aggressive, organized state, their dependence on elaborate irrigation systems for the maintenance of concentrated populations rendered them vulnerable to attack, which was one of the main factors that enabled the Inca to take them over comparatively easily. (G.H.S.B./J.V.M.)

The Chincha. The growth and expansion of Chimú were paralleled on the southern coast by Chincha, which was a similarly well-organized polity. Comparison between them has been difficult because of the very different evidence available. Whereas Chimú has become familiar through extensive archaeological research, data on the Chincha has come primarily from the study of historical sources.

In the first few years of Spanish rule, the Holy Roman emperor Charles V complained that he had not received any of the newly conquered lands as a personal fief. The conquistador Francisco Pizarro and his brother Gonzalo hurried to assign him three ethnic groups: (1) The Aymara kingdom of the Lupaqa, listed on the Inca quipu at 20,000 households; (2) the tropical island of Puná, in the Gulf of Guayaquil in modern Ecuador, with an unknown aboriginal population; and (3) the coastal Chincha polity, allegedly with 30,000 households. Unfortunately for the Chincha, their population vanished within the first three decades of the Spanish invasion; the royal affiliation and proximity to Lima did not help protect the Chincha.

Belonging to the crown, however, did promote account keeping and administrative reports to the Spanish court. The unusual feature about Chincha was its considerable orientation to the sea. Several thousand households were listed as high-seas fishers and sailors, and thousands more were engaged in long-distance trade with lands to the north. Because the waters off the Chilean and Peruvian coast were cold, there was a long-standing interest in the warm waters off the Ecuadorean coast, more than 1,000 miles away, where the Antarctic current was no longer present. The details of these exchanges are not known, but one feature was paramount in Andean eyes: throughout the central and southern Andes, wherever puna dwellers were the dominant population, there was a demand for the spiny oyster (*Spondylus*), the shells of which were believed to encourage rainmaking. The one Quechua literary text available lists the spiny oyster as the favourite food of the gods, although it was inedible for humans.

Chincha orientation to the sea

While there has been a long-standing archaeological interest in the shells of this mollusk, the extent and the organization of the shell traffic has not been verified archaeologically. One of the witnesses of the invasion, Pedro Pizarro (a cousin of Francisco), reported being told that the Chincha lord had 100,000 rafts on the "Southern Sea." The number need not be accurate: even 1,000 oceangoing rafts, with keels and sails, would imply a major economic operation.

Chimú and Chincha have received considerable attention from non-Peruvian scholars; understanding of the contemporaries of these peoples in the highlands, however, has remained sketchy. The oral tradition reported by the early European observers claimed that before the expansion of Tawantinsuyu, the Inca state, there were many polities large and small, all ruled by traditional lords and frequently at war with one another. To what extent the notions of ecological complementarity or the vertical archipelago were attempts to bridge these conflicts or their consequences cannot be stated with any certainty.

The 17th-century Andean writer Felipe Guamán Poma de Ayala (Waman Puma) reported the oral tradition that he had learned from his forebears, who were minor ethnic lords in the Huánuco region. In the century before the Inca conquest people had lived in the "epoch of the soldiers." During this period,

> they began to fight and there was much war and death . . . one lord against the other . . . bloodshed and taking of prisoners. And they also grabbed their wives and sons and took their fields and irrigation waters and pastures. And they were very cruel and stole each other's property, cloth, gold, copper, even their millstones . . . . And so they went and settled on the heights where they built walls and houses inside . . . and wells to draw water.

Poma de Ayala's description of Late Intermediate settlement patterns on mountaintops, at the very edge of and even beyond puna cultivation, has been confirmed by field research undertaken near Lake Titicaca by the American archaeologist John Hyslop. He found dozens of walled-in enclosures of 50 to 100 acres and larger. During the Late Horizon—which corresponds to the century of Inca rule—these populations were moved to the lakeshore, along the royal road.

The Inca

Forty years had elapsed since Columbus' landfall when in 1532 fewer than 200 Spaniards brought down the Inca (Inka) state. Ever since then, historians have been pondering the reasons for this sudden collapse. The evidence seems to favour internal subversion. Don Francisco Cusichaq, lord of the Huanca in central Peru, opened the country to alien rule; he wanted to destroy his hereditary enemies, the Inca. The Andean pattern of many dispersed regional polities that frequently were at war with one another—a situation that the Inca had manipulated but had not eliminated—and the diverse archipelago-like string of the communities may also have facilitated the relatively effortless Spanish victory.

Extent of the Inca state

By 1532 Tawantinsuyu, the Inca state, had incorporated dozens of coastal and highland ethnic groups stretching from what is now the northern border of Ecuador to Mendoza in west-central Argentina and the Maule River in central Chile—a distance roughly equal to that between New York City and the Panama Canal. By conservative estimates the Inca ruled more than 12,000,000 people, who spoke at least 20 different languages. A century earlier, during the wars of the Late Intermediate, they had controlled little beyond the villages of their own Cuzco

Table 2: Inca Rulers and Royal Corporations

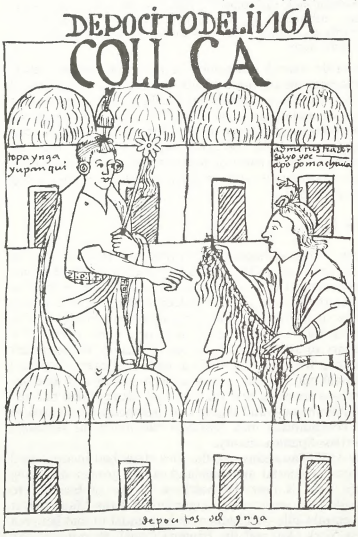
Inca name	Spanish spelling	*panaca* (royal corporation)
Manqo Qhapaq	Manco Capac	Chima
Zinchi Roq'a	Sinchi Roca	Rawra
Lloq'e Yupanki	Lloque Yupanqui	'Awaynu
Mayta Qhapaq	Mayta Capac	'Uska Mayta
Qhapaq Yupanki	Capac Yupanqui	'Apu Mayta
'Inka Roq'a 'Inka	Inca Roca	Wika-k'iraw
Yawar Waqaq	Yahuar Huacac	'Awqaylli
Wiraqocha 'Inka	Viracocha Inca	Zukzu
'Inka 'Urqon	Inca Urcon	none
Pachakuti 'Inka Yupanki (1438–71)	Pachacuti Inca Yupanqui	'Iñaqa
Thupa 'Inka Yupanki (1471–93)	Topa Inca Yupanqui	Qhapaq
Wayna Qhapaq (1493–1525)	Huayna Capac	Tumipampa
Washkar 'Inka (1525–32)	Huascar	Washkar
'Ataw Wallpa 'Inka (1532–33)	Atahuallpa	none
Thupa Wallpa (1533)	Topa Huallpa	none
Manqo 'Inka Yupanki (1533–45)	Manco Inca Yupanqui	none (?)

Valley. While forming their state they subordinated more than 100 independent ethnic groups; how much of this achievement corresponded to political experience gained during the Middle Horizon cannot be told. It is likely that the memory of that multiethnic expansion was alive in the ruling families of the major polities.

THE ORIGINS AND EXPANSION OF THE INCA STATE

Inca origins and early history are largely shrouded in legends that may be more mythical than factual. Their later history, particularly from the reign of Pachacuti Inca Yupanqui (Pachakuti 'Inka Yupanki) onward, is largely based on fact, even though it presents what the Inca wanted people to know. Whether these historical traditions are true, in the sense that they accurately related what happened, is not so important as the fact that the Inca used them to justify their various imperial conquests.

Courtesy, Library Services Department, American Museum of Natural History, New York City (Neg. No. 321546)



Bookkeeper (right) rendering accounts to the Inca ruler Topa Inca Yupanqui. The contents of the storehouses (foreground and background) are recorded on the bookkeeper's quipu of knotted strings. Drawing by Felipe Guamán Poma de Ayala from *El primer nueva coronica y buen gobierno.*

The nature of the sources. The Inca kept detailed accounts of their dynastic history, knotted onto the quipu records kept by professional accountants. The major local ethnic lords also kept records. As mentioned above, Don Francisco Cusichaq kept records of Spanish exactions, which were offered to and accepted in evidence by Spanish administrators. Through the study of Cusichaq's quipu, modern researchers have learned that there was both a quantitative and a historical dimension to Andean records. Cusichaq's quipu refers to more than 20 separate events—all recorded in perfect historical sequence—but the way in which these events were recorded has not been fathomed. The quantitative record, which was easier to decipher, lists counts of men and women on the first two cords, followed by the number of domestic animals (llamas being separated from alpacas). Cloth, the most valuable commodity according to Andean reckoning, comes first among the goods listed, followed by food and household items. The quipu could incorporate strings for new, Spanish items. Thus, in Cusichaq's records Spanish sandals are itemized separately from Andean footwear, and eggs and imported hens have their own strings.

The quipu records

The Cuzco bookkeeping records were used by the Spanish in the early days of their rule in order to divide the country and its population among the invaders. The accuracy of the information about distant places and peoples available to the Inca rulers astonished the Spanish observers. Some among them transcribed what they were told; these accounts became the source of the fragmentary information available to modern researchers. In 1549 and again in the 1570s systematic efforts were made by the Spanish to investigate the Andean past. Some of the interviewers were excellent ethnographers who noted discrepancies between separate oral traditions and contradictions from one set of claims to another. Just as in Mexico, where there were true ethnographers like Bernardino de Sahagún, so in the Andes a young soldier, Pedro de Cieza de León, was a remarkable interviewer, who constantly checked what he had been told by the members of one royal lineage against alternate versions.

Thus, the present knowledge of Inca society has been derived from a combination of archaeological studies and the written accounts sent to Spain by the early Spanish observers. Some of these accounts reached a wide public: within two years of the fall of the Inca, two quite different versions of what happened at Cajamarca (the place where Pizarro first met and kidnapped the Inca ruler Atahuallpa) were already in print in Europe. One of these was the official version of the Pizarro brothers, while the other criticized their actions. At a time when printing was still a rare skill and censorship was severe, such ample coverage of the invasion is notable.

The first serious study of the Andean peoples was written by Cieza de León, who had reached the Americas as a 14-year-old soldier and had settled in what today is Colombia. A decade or so later he drifted by horse to what is now Peru; he then rode for some 1,300 miles, traveling as far south as the mines at Potosí, in present-day Bolivia. Cieza de León was encouraged by the clergy, many of them partisans and correspondents of the Dominican missionary and historian Bartolomé de Las Casas, to interview both Spanish and Andean participants of the invasion and of the wars that some Andean factions had fought against one another.

Work of Cieza de León

The most widely read source during the colonial period was the work of Garcilaso de la Vega, also called El Inca—the son of an Inca royal woman and a Spanish nobleman (whose name the son adopted when he "returned" to his father's estate in Spain). He lived in Spain nearly 60 years, leading the life of a gentleman, reading, translating love poetry, editing the memoirs of one of the early invaders of Florida, and, finally, writing a vast account of his mother's ancestors, *The Royal Commentaries of the Inca.*

Guamán Poma de Ayala (Waman Puma) was one of the few Andean writers whose work is available. He wrote a 1,200-page "letter" to Philip III of Spain, consisting of two books combined into one. The first book was a "new chronicle," describing Andean achievements and history; the second, much larger part advised the king on how to

achieve a "good government." The second included 400 pages of pen-and-ink drawings, which have remained a major contribution to the modern understanding of Andean society. The manuscript somehow reached the Danish Royal Library in Copenhagen, where it was discovered in 1908 and where it still resides.

Settlement in the Cuzco Valley. Several of the modern Andean peoples trace their ancestries to mythical figures who emerged from holes in the ground. These places of origin, or *paqarina,* were regarded as shrines, where religious ceremonies had to be performed. The Inca *paqarina* was located at Paqari-tampu (Paccari Tampu), about 15 miles south of Cuzco. There are three caves at Paqaritampu, and the founders of the Inca dynasty—Manco Capac (Manqo Qhapaq), his three brothers, and his four sisters—supposedly emerged from the middle cave. They assumed leadership over 10 groups of people, or ayllus, that emerged from the caves on either side and led them on a journey lasting an unknown number of years.

Paqarina, the shrines of origin

During this period the Inca and their followers moved from village to village in search of enough fertile land to sustain themselves. Manco Capac succeeded in disposing of his three brothers. One of his sisters, Mama Ocllo, bore him a son named Sinchi Roca (Zinchi Roq'a). Eventually, the Inca arrived at the fertile area around Cuzco, where they attacked the local residents and drove them from the land. They then established themselves in Cuzco and gradually began to meddle in the affairs of their neighbours, forcing them to pay tribute in order to retain their freedom.

By this time Manco Capac was quite old and close to death. In order to ensure that all he had accomplished would be preserved for posterity, he named his eldest son, Sinchi Roca, to succeed him to the throne. He then directed his next eldest son to shelter and care for all of his other children and their descendants, who composed the Chima *panaca.* The traditions say little about Sinchi Roca, the second emperor, but apparently he was a peaceful man who made no military campaigns to add lands to the Inca domain. It is not clear whether or not Sinchi Roca married his sister, as his father had done. It is clear, however, that he did not follow his father's lead in naming his eldest son as his successor, for the third emperor, Lloque Yupanqui (Lloq'e Yupanki), had an older brother. Lloque Yupanqui, like his father, was not warlike and added no lands to the Inca domain.

The demand for additional lands and, more importantly, the resources they could provide first became apparent during the reign of the fourth emperor, Mayta Capac (Mayta Qhapaq). The reasons for the appearance of this need in the 14th century are undoubtedly complex, and any single-factor explanation is probably insufficient. But one possible explanation may lie in the fact that rainfall began to diminish very slightly about this time throughout the central Andes. In an area like the Cuzco Valley, this would imply that some of the marginal farmlands were either abandoned because they could not be watered adequately or were less productive than they had been earlier. Given this situation, if the Inca attempted to maintain their old standard of living, they might have placed some pressure on their food resources. One way of alleviating the problem would have been to acquire additional land and sources of water in an adjacent part of the valley. This is apparently what Mayta Capac did.

Mayta Capac

Mayta Capac is described in the chronicles as a large, aggressive youth who began fighting with boys from a neighbouring group when he was very young. Pedro de Cieza de León and Pedro Sarmiento de Gamboa (who also was one of the more reliable Spanish chroniclers) indicate that the quarrel began because the Inca were taking water from this group, although they differ on the details concerning who actually took the water. By the time Mayta Capac became emperor, this quarrel had grown into a full-scale war, which the Inca won. They looted the homes of their enemies, took some of their lands, and probably imposed some sort of tribute payment on them.

The beginnings of external expansion. The fifth emperor, Capac Yupanqui (Qhapaq Yupanki), was appointed ruler by his father before he died. He was apparently not the eldest son but was named emperor because his older brother was considered ugly. Capac Yupanqui was the first Inca ruler to conquer lands outside the Cuzco Valley, although these were only about a dozen miles away. Inca Roca ('Inka Roq'a 'Inka) succeeded his father and subjugated some groups that lived about 12 miles southeast of Cuzco. He is mostly remembered in the chronicles for the fact that he fathered a large number of sons, one of whom, Yahuar Huacac (Yawar Waqaq), was kidnapped by a neighbouring group when he was about eight years old. The boy's mother, Mama Mikay, was a Huayllaca (Wayllaqa) woman who had been promised to the leader of another group called the Ayarmaca ('Ayarmaka). When the promise was broken and Mama Mikay married Inca Roca, the Ayarmaca went to war with the Huayllaca and were defeating them. As a peace offering, the Huayllaca agreed to deliver Mama Mikay's son to the Ayarmaca. This tale says a great deal about the way war was waged around the Cuzco Valley at this time; the fact that the Ayarmaca held the boy for several years before returning him to his father suggests that the Inca were no more powerful than several other groups in the area.

Two years before his death, Inca Roca named Yahuar Huacac as the seventh emperor, ensuring a peaceful succession to the throne. Yahuar Huacac was never very healthy and apparently spent most of his time in Cuzco. His brothers Vicaquirao (Wika-k'iraw) and Apo Mayta ('Apu Mayta) were able military leaders and incorporated lands south and east of Cuzco into the Inca domain. Yahuar Huacac's principal wife was apparently an Ayarmaca, indicating that at that time sister marriage was not the rule (see below *Civil war on the eve of the Spanish conquest*). She bore him three sons, and he attempted to follow his father's example by naming her second son as the next emperor; the son was murdered through the intrigues of another of his wives, who wanted her own son named to the throne. The Emperor himself was apparently killed shortly thereafter, and the elders chose Viracocha Inca (Wiraqocha 'Inka) as his successor.

The Inca conquest began during the reign of Viracocha Inca in the early part of the 15th century. Up to this time, neighbouring ethnic groups were conquered and their lands taken, but no garrisons or Inca officials were placed among them. They were left undisturbed until the Inca felt it necessary to attack them again. This pattern of raiding and plundering changed during Viracocha Inca's reign. He planned to establish permanent rule over these groups and was ably assisted by his uncles, Vicaquirao and Apo Mayta, who developed military tactics that made permanent conquest possible. Their victory over the Ayarmaca kingdom in the southern Cuzco Valley provided a model for many subsequent campaigns. They first conquered lands in the upper part of the Urubamba Valley that lay behind the Ayarmaca territory. They then successfully attacked the Ayarmaca from two directions—one force coming from Cuzco and the other from the Urubamba Valley.

The beginning of permanent conquests

This was a relatively small-scale campaign, but it made the Inca a political power in the Urubamba Valley, an important passageway between Cuzco and the Lake Titicaca Basin. As a result of their conquest, the Inca were invited to interfere in a conflict between two Aymara-speaking kingdoms, the Colla and the Lupaca, in the northern part of the Titicaca Basin. The Inca allied themselves with the Lupaca, probably because the Colla were located between themselves and the Lupaca. But before the Inca could attack, the Colla attacked the Lupaca and were defeated. The battle was over by the time the Inca arrived; they joined in a victory celebration with the Lupaca but did not share in the booty.

During the early 15th century a group called the Chanca was emerging as a political power in the area west of the Inca territory. Presumably, they, too, may have been feeling the effects of diminishing food resources and were trying to maintain their standard of living by acquiring land outside their home territory. They moved from their place of origin in Huancavelica and conquered the Quechua (K'ichuwa), a large group whose lands lay immediately west of those controlled by the Inca. In about

Rise of the Chanca

1438 the Chanca attacked the Inca. One of the major effects of the Chanca invasion was to foment a civil war among the Inca.

Pachacuti Inca Yupanqui. For some time there had been palace intrigue in Cuzco over who would succeed Viracocha Inca to the throne. The Emperor chose Inca Urcon ('Inka 'Urqon) as his successor, but the two generals Vicaquirao and Apo Mayta preferred another son, Cusi Inca Yupanqui (Kusi 'Inka Yupanki). As the Chanca approached Cuzco, Viracocha Inca and Inca Urcon withdrew to a fort near Calca, while Cusi Inca Yupanqui, the two generals, and a few nobles remained to defend the city. They defended it successfully, and after their allies joined them they inflicted two heavy defeats on the Chanca. Cusi Inca Yupanqui then attempted to resolve the differences between his faction and that of his father; but the negotiations failed, and he set himself up as emperor, taking the title of Pachacuti (Pachakuti). At this point, there were two Inca states, one in Cuzco, headed by Pachacuti Inca Yupanqui, and the other in Calca, headed by Viracocha Inca. As the power and prestige of the Cuzco group increased, many people left the Calca faction to join Pachacuti Inca Yupanqui.

Victory over the Chanca

Pachacuti Inca Yupanqui had to deal simultaneously with two enemies—the Chanca and his father's forces. The Cuzco faction had made some gains during their two encounters with the Chanca; they took some Quechua lands from the Chanca and formed an alliance with the Quechua, who supported them against the Chanca. They then entered into an agreement with the Chanca that permitted either group to make independent military advances or gains as long as the other was not attacked. At this point, the Cuzco faction moved its army eastward to the edge of the tropical rain forest, thereby encircling the lands controlled by the Calca faction. By this maneuver, the Cuzco faction prevented the possibility of attack coming simultaneously from two directions. Viracocha Inca died about this time, leaving Inca Urcon as leader of the Calca faction. The latter was killed shortly thereafter in a skirmish with the Cuzco group. As a result, the differences between the two factions were resolved, and the Inca were reunited under a single leader.

The Inca forces crossed the Quechua territory and attacked the provinces of Vilcas and Soras, southwest of the area controlled by the Chanca. In about 1445, Pachacuti Inca Yupanqui sent his brother Capac Yupanqui (Qhapaq Yupanki) to explore the south coast, marking the first time the Inca reached the ocean. Returning to Cuzco, Capac Yupanqui passed through Chanca territory and captured a few of their villages. The Chanca retaliated by outflanking the Inca and conquering the Colla in the Lake Titicaca Basin.

The Chanca's action increased the tension between the Inca and the Chanca, but no fighting broke out. Instead, they decided to undertake a joint invasion of the area north of Vilcas. Pachacuti Inca Yupanqui appointed Capac Yupanqui to lead the Inca contingent, warning him of Chanca treachery and instructing him to go no farther than Yanamayo. As the expedition moved northward, the Chanca distinguished themselves in battle, to the embarrassment of the Inca. When Pachacuti Inca Yupanqui heard of this, he feared that the Chanca contingent might revolt and ordered his brother to kill the Chanca leaders. The Chanca, learning of this command, fled to the tropical rain forest near the headwaters of the Huallaga River before the order could be carried out.

Capac Yupanqui pursued the Chanca well beyond the Yanamayo, the limit set by his brother, before giving up the chase. Seeing that his forces were considerably overextended, he turned northward toward the rich province of Cajamarca, which was an ally of the powerful kingdom of Chimú on the north coast. Capac Yupanqui stormed and captured Cajamarca and left a small garrison there as he set out for Cuzco.

Pachacuti Inca Yupanqui was furious at this turn of events. His orders had been blatantly disobeyed, and he was apprehensive about his brother's intentions. Perhaps fearing that Capac Yupanqui would usurp the throne, Pachacuti Inca Yupanqui had him killed before he arrived in Cuzco. The Inca still had to contend with the Chanca and with the possibility of attacks from hostile groups in the north, including the kingdom of Chimú, which had set out on a campaign of conquest.

Murder of Capac Yupanqui

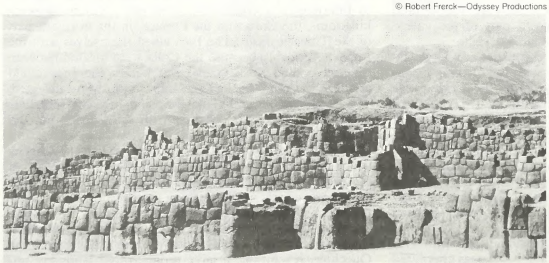
To alleviate this situation, Pachacuti Inca Yupanqui organized two expeditions: one to conquer the peoples of the Titicaca Basin and protect their exposed southern flank and the other to subdue the area to the north. According to Sarmiento de Gamboa, the Titicaca campaign was led by two of his older sons. They had subjugated the Colla earlier and now turned their attention to the Lupaca and their allies. When the campaign was over, the Inca controlled all of the territory between Cuzco and the southern end of the lake basin.

The northern expedition was led by another son, Topa Inca Yupanqui (Thupa 'Inka Yupanki), who subjected the territories of the Quechua and the Chanca. Topa Inca Yupanqui marched north through the highlands toward Cajamarca, subduing and pacifying the country as he went. After relieving the garrison at Cajamarca, which was being threatened by the kingdom of Chimú, he conquered as far north as Quito (Ecuador) in an attempt to outflank the Chimú armies. Frustrated during this drive by his ignorance of the geography of the region, he came out of the Ecuadorian mountains near Manta, north of the Gulf of Guayaquil; the local residents told him that he could not proceed southward along the coast because the mountains came down to the sea. So he returned to the highlands and sent a small force along the shores of the Gulf of Guayaquil toward the northern border of Chimú. As a result, the Inca were still able to attack the Chimú armies simultaneously from several different directions. After a brief but bitter battle, the Inca sacked the Chimú capital at Chan Chan and then advanced southward along the coast as far as Pachacamac, bringing the area under Inca control.

Defeat of the Chimú

Topa Inca Yupanqui returned to Cuzco, secure in the knowledge that Inca power could not be challenged. The rapid expansion of the empire, however, created a number of problems concerned with sustaining themselves and governing a large number of diverse ethnic groups.

© Robert Frerck—Odyssey Productions



Stone walls of the giant Inca fortress of Sacsahuamán on a hill above Cuzco, Peru.

Pachacuti Inca Yupanqui and Topa Inca Yupanqui were imaginative and made several important innovations in Inca institutions.

Pachacuti Inca Yupanqui began rebuilding Cuzco, the political and religious capital of the empire. Considerable effort was put into enlarging Sacsahuamán, the huge fortress built on a hill overlooking the city. At the same time he undertook a vast agricultural project over the entire upper end of the Cuzco Valley; rivers were channeled, the valley floor was leveled, and agricultural terraces were built on the surrounding hillsides. This reclamation project undoubtedly increased the agricultural productivity of the area and involved moving many of the original inhabitants of this part of the valley to other localities for several years while the work was being completed.

Pachacuti Inca Yupanqui also turned his attention to social problems. He decreed that no ruler could inherit property from his predecessor; instead, the property of a dead ruler was to pass to his other descendants, who could then support themselves from his lands and the labour taxes owed him. Consequently, each new emperor had to acquire land and labour to support his corporation and government. Pachacuti Inca Yupanqui thus ensured that the corporations of his eight predecessors had estates in the area around Cuzco so their members could support themselves adequately, attend certain ceremonies, and perform ceremonial obligations. Pachacuti Inca Yupanqui and his successors to the Inca throne formed corporations that had lands and estates scattered throughout the empire as well as in the Cuzco Valley itself.

Policy of *mitma*

He probably also began the policy of forced resettlement, or *mitma,* about this time, in order to ensure both loyalty to the state and better utilization of land resources, at least from the perspective of the Inca. This practice involved moving some members of an ethnic group from their home territory to distant lands. When a new area was conquered, loyal settlers were brought in from a province that had been under Inca rule long enough so that its residents knew how the Inca system of government worked. They were replaced in their home territories by recalcitrant groups from the newly conquered province. The policy had three important consequences: first, it broke up the size and power of an ethnic group by dispersing its members throughout the empire; second, it weakened the ability of an ethnic group to be self-sufficient; and, finally, it made it more difficult for the inhabitants of an area to revolt successfully.

Pachacuti Inca Yupanqui invented a state religion based on the worship of a creator-god called Viracocha, who had been worshiped since pre-Inca times. Priests were appointed, ceremonies were planned, prayers were prepared, and temples were built throughout the empire. He also expounded the view that the Inca had a divine mission to bring this true religion to other peoples, so that the Inca armies conquered in the name of the creator god. His doctrine was a relatively tolerant one. Conquered groups did not have to give up their own religious beliefs; they merely had to worship the Inca god and provide him and his servants with food, land, and labour.

Topa Inca Yupanqui. About 1471, Pachacuti Inca Yupanqui abdicated in favour of his son Topa Inca Yupanqui, thereby ensuring the peaceful succession to the throne. Topa Inca Yupanqui was a great conqueror who was to bring most of the Central Andes region under Inca rule. Yet his first military campaign as emperor, an invasion of the tropical rain forest near the Tono River, was not particularly successful. The Inca were always fascinated with the rain forest and its products but never got used to military operations in this type of environment. This campaign did, however, establish trade relations with the area and secured a contingent of archers in return for a few bronze tools. The Emperor soon abandoned the campaign because of a revolt that had broken out in the Titicaca Basin. The rebellion was led by the Colla and Lupaca and was fanned by the rumour that Topa Inca Yupanqui had been killed during his expedition into the jungle.

The Colla's mountaintop forts around Pucará fell one by one as the Inca attacked them. After subduing the Colla, the Inca moved against the Lupaca, who had retreated to the southwest corner of the Titicaca Basin, where they had allied themselves with another Aymara-speaking group, the Pacasa. The Inca armies were again victorious, and the revolt was ended. Topa Inca Yupanqui then turned southward, conquering all of highland Bolivia, northern Chile, and most of northwestern Argentina. He set the boundary markers of the Inca Empire at the Maule River in central Chile.

At this point, the southern coast of Peru still had not been incorporated into the Inca state. The area, however, was now surrounded by the Inca on three sides, and in about 1476 Topa Inca Yupanqui launched a campaign against this region. Each valley, beginning with those in the south, was attacked separately. Most valleys submitted peacefully or put up only minimal resistance; the inhabitants of the Cañete Valley, however, put up a stubborn fight; and it took the Inca nearly three years to subdue them.

During the remainder of his reign, Topa Inca Yupanqui concerned himself with the administration of the empire. He spent much of his time traveling throughout his territories, making assignments of land and establishing local administrations. He introduced a system of classifying the adult male population into units of 100, 500, 1,000, 5,000, and 10,000, which formed a basis for labour assignments and military conscription. He also instituted a system of tribute in which each province provided Chosen Women (Aqllakuna) to serve as temple maidens in state shrines or to become the brides of soldiers who had distinguished themselves in combat.

Huayna Capac. Topa Inca Yupanqui's unexpected death in about 1493 precipitated a struggle for the succession. It appears that Topa Inca Yupanqui had originally favoured the succession of Huayna Capac (Wayna Qhapaq), the youngest son of his principal wife and sister. Shortly before his death, he changed his mind and named as his successor Capac Huari (Qhapaq Wari), the son of another wife. Capac Huari, however, never became emperor. The claims of his mother and her relatives were suppressed by the supporters of Huayna Capac. This group was led by Huaman Achachi (Waman 'Achachi), the child's uncle and presumably the brother of the Emperor's principal wife. A regent named Hualpaya (Walpaya) was appointed from this group to tutor Huayna Capac in the ways of government until the child was old enough to rule in his own name. Hualpaya, however, tried to assert the claims of his own son to the throne and, as a result, was killed by Huaman Achachi. Huayna Capac's reign was mostly peaceful; he devoted much of his time to traveling, administering the empire, and suppressing small-scale revolts. He did extend the empire by conquering Chachapoyas, a mountainous country in northeastern Peru, and later northern Ecuador. After conquering Chachapoyas, he recruited part of his bodyguard from the warlike inhabitants of the area. The conquest of northern Ecuador occupied the last years of his life and took place shortly before the Spaniards arrived. During these campaigns, he pushed the frontiers of the Inca Empire to the Ancasmayo River, the present-day boundary between Ecuador and Colombia.

Conquest of northern Ecuador

While he was fighting in northern Ecuador, Huayna Capac received word that the Bolivian frontier had been invaded by the Chiriguano, a Guaraní-speaking group that periodically crossed the Gran Chaco from Argentina to raid Inca frontier settlements for bronze tools and ornaments made of precious metals. The Chiriguano were more of a nuisance than an actual threat to the empire, but Huayna Capac dispatched a general named Yasca (Yaska) to drive them from the area and to build forts along the frontier.

Meanwhile, he undertook another expedition in northern Ecuador to wipe out isolated pockets of resistance. During this campaign, he learned that an epidemic was sweeping Cuzco and the surrounding countryside. He left immediately for Quito, on the highroad to Cuzco, to deal with this crisis and arrived there about the same time the epidemic did. The pestilence had spread rapidly from Bolivia and, judging by its description, was either smallpox or measles, both of which were European diseases introduced into South America by the Spanish settlers at La Plata. The disease was probably communicated to the

Andean area by the Chiriguano, who had been in contact with the Spanish at La Plata. Whatever the ailment was, Huayna Capac contracted it and died about 1525, without naming a successor in the appropriate manner. This set off another struggle over the throne.

Civil war on the eve of the Spanish conquest. Huayna Capac's father had begun the custom of marrying a full sister in order to keep the royal bloodline pure and, more importantly, to prevent conflict over succession. According to this custom, one sister became the principal wife of the emperor, and one of their sons became the next ruler. As noted above, this system had failed at Huayna Capac's succession. Nor did it work at Huayna Capac's death because his principal wife had been childless. In this situation, the emperor could appoint any one of his sons as his successor, as long as one of them had "divine" approval registered on the lungs of a sacrificed llama. There were several candidates for the throne: Ninan Cuyuchi who was in Tumipampas with his father; Atahuallpa ('Ataw Wallpa), who was also in the north; Huascar (Washkar), who was apparently in Cuzco; Manco Inca (Manqo 'Inka), whose mother belonged to 'Iñaqa (the royal corporation of Pachacuti Inca Yupanqui); Topa Huallpa (Thupa Wallpa); and Paullu Topa (Pawllu Thupa).

Rule of succession

Huayna Capac, aware of imminent death, asked the priest to perform the divination ceremony to determine whether or not he should name Ninan Cuyuchi as his successor; if the signs were not favourable, then Huascar was to be the next candidate to be tested. The Emperor apparently died before the ceremony was performed. The priest then notified Ninan Cuyuchi that he was to be the next ruler, but the latter had contracted the same disease as his father and died shortly thereafter. The priest then named Huascar as the new emperor; this was highly irregular, because the priest apparently followed the old ruler's wishes without performing the required ceremony. The other candidates for the throne were not pleased with the situation.

The priest brought Huayna Capac's body back to Cuzco, while Atahuallpa remained in Quito. Huascar was so furious with the priest for leaving a rival for the throne in the north with a large army that he had him killed. This created animosity against Huascar among the members of the priest's royal corporation. Huascar then demanded that Atahuallpa return to Cuzco, but the latter ignored him and undertook a campaign to suppress a revolt around the Gulf of Guayaquil. While he was involved in this expedition, Huascar sent an officer to remove Atahuallpa's wives and insignias. Atahuallpa killed the officer and had a drum made out of him, which he sent to Huascar. This insult completed the breach between the two rivals, and a civil war resulted.

At this point, Huascar controlled the southern part of the empire, while Atahuallpa controlled Ecuador and parts of northern Peru. Atahuallpa won the first battle of the war, fought at Riobamba in Ecuador, and advanced to Tumipampas. There he lost to Huascar's army and was taken prisoner. He later escaped, rallied his forces, and drove his brother's army from the Cañari territory around Tumipampas. He then devastated the Cañari lands because he thought they had supported his brother's faction during his imprisonment. Apparently, the Cañari wanted little to do with either Inca faction and offered minimal support to whichever group controlled Tumipampas at the moment. After their lands were destroyed, they wanted nothing at all to do with the Inca, and later they became close allies of the Spaniards.

Victory of Atahuallpa

Atahuallpa's armies, led by the able generals Quisquis (Kizkiz) and Challcuchima (Challku-chima), marched south and won a series of decisive victories at Cajamarca, Bombon, and Ayacucho. As they moved southward, Huascar formed another army to defend Cuzco from the invaders. His forces were defeated, and he was captured a few miles from Cuzco in April 1532. The generals killed his entire family and fastened them to poles along a highway leading from the capital. They also killed a number of people in Topa Inca Yupanqui's corporation because they had supported Huascar during the civil war; and they burned the mummy of the deceased ruler, which was venerated by the members of this group. Atahuallpa was in the north, setting up his administration, when he learned of the victory. He ordered Challcuchima to bring Huascar to the north so he could insult him properly before being crowned.

The Spanish conquest. Meanwhile, the Spaniards had landed at Tumbes on the northern coast of Peru early in 1532 and were seeking an interview with Atahuallpa so that they could kidnap him. It is clear that they understood the nature of the Inca civil war and were dealing with emissaries from both factions. Their actions, however, must have seemed puzzling to Atahuallpa. On the one hand, Pizarro and his men were deposing and executing leaders who were loyal to him, and, on the other hand, they were sending messages that recognized him as the legitimate ruler of Tawantinsuyu. As the Spaniards moved toward Cajamarca, he sent them a message indicating that he was now the sole ruler of his father's domain. Furthermore, he reminded the Spaniards that they were far from their base of supply and in a land controlled by his armies. The Spaniards replied to this veiled threat by indicating that they would come to his aid against any group that opposed his rule. Atahuallpa clearly wanted the Spaniards as allies but continually misinterpreted their intentions and underestimated their abilities—even after he was kidnapped in Cajamarca on Nov. 16, 1532.

Imprisonment of Atahuallpa

Atahuallpa was allowed to meet with his advisers while the Spaniards held him prisoner, and he arranged to have the ransom they demanded paid. An enormous ransom was raised, but Pizarro did not free him because it would have been too dangerous for the Spaniards. While he was in prison, Atahuallpa decided that the Spaniards were indifferent to the idea of having his brother slain and ordered Huascar's death. The Spaniards, of course, wanted all pretenders to authority removed but later used this act to justify their execution of the Inca ruler. Realizing that Atahuallpa's death was a mistake because it weakened their position, they approved the coronation of Topa Huallpa, a candidate whom they thought would be acceptable to both Inca factions. But the Spaniards miscalculated. Topa Huallpa had not supported Atahuallpa and, in fact, had been in hiding as long as the latter was alive. He was supported by Huascar's group and was opposed by Atahuallpa's following, who believed that the legitimate heir was the deceased ruler's son in Lima. With this act, the Spaniards suddenly found themselves closely allied with Huascar's faction and were so viewed by both Inca groups.

Topa Huallpa died within a few months—poisoned, according to Huascar's supporters. At this point, the Spaniards reaffirmed their alliance with Huascar's following, placing Huascar's brother, Manco Inca, on the throne and assisting him in dispersing the remnants of Atahuallpa's army. The real Spanish conquest of Peru occurred during the next few years, when they prevented Manco Inca from reestablishing control over the coast and the north, much of which was still loyal to Atahuallpa or under no control at all. By 1535 the Inca ruler realized that the Spaniards were more dangerous than any threat posed by the remnants of Atahuallpa's followers. But it was too late. His attacks on the Spanish settlements were beaten back, and he was eventually driven into a remote mountainous area called Vitcos, where he established an independent Inca state that lasted until 1572.

INCA CULTURE AT THE TIME OF THE CONQUEST

The rapid incorporation of so many mountain and coastal desert polities before 1532 calls for explanation. It is tempting to view such expansion in the context of the instantaneous breakup in 1532, when some of the same forces were likely to have been at work: dispersed territories, interlocked with some belonging to other powers in the region, and multiethnic and polyglot agglomerations in neighbouring valleys. Each political unit—as eventually was the case with the Inca state itself—was likely to share pastures, cultivated terraces, and beach installations; hegemonies shifted according to local and regional circumstances. The Early, Middle, and Late Horizons were temporary concatenations, and none lasted for very long. The Spanish invasion interrupted these alternations: a

player had entered the field who ignored the local rules and who did not fathom the true sources of Andean wealth, which was not silver but an intimate familiarity with local conditions and possibilities and the ability to pool vastly different geographic and ecological tiers into single polities.

Social and political structure. According to the incomplete evidence provided by the Spanish eyewitnesses, the Inca themselves considered the term Inca applicable only to the descendants of the 12 individuals who traditionally are said to have ruled from Cuzco. Of the 12, only four or five can be documented to have been actual historical personages. The others may have been products of later efforts to legitimate and enhance the royal genealogy. There is also the possibility that some of the "earlier" names were actually a parallel line of personalities, possibly with different functions that may have been considered "heathen" by the Spanish. This hypothesis cannot be verified with the sources now available.

Terminology

In addition to the 12 lineages, the ranks of "Inca by decree" or "as a privilege" are also mentioned by some of the Spanish sources. Their origins and functions were just as nebulous as those of the royals: one of the few Andean sources, Poma da Ayala, claims that some of the inhabitants of the Cuzco basin who were conquered early during the expansion of the Late Horizon were "granted" or "promoted to" Inca status. They were "improved," according to Poma da Ayala, although his own case is weakened by his claim that his ancestors, who lived many hundreds of miles north of Cuzco, had benefited from such social mobility.

Administrative organization

The administrative organization of Tawantinsuyu is poorly understood, although its origins are known to lie in the earlier ethnic subdivisions. Claims have been made that authority was left in the hands of traditional lords who simply had to demonstrate their fealty. Other Spanish sources make reference to an administrative reorganization, in which all of the conquered groups were shoehorned into a decimal system. There is some evidence that decimal subdivisions were present in the Cajamarca region of northern Peru; and at the time of the conquest the decimal vocabulary apparently was in the process of being imposed on the rest of the country, presumably to rationalize the multiplicity of local and divided loyalties. The administrative papers available for a part of the Huánuco region allow the identification of a "hundred-households" unit with five actual hamlets, all of which were near each other. Since these records were kept house by house, it has been possible to test the significance of the decimal vocabulary at its lowest level. What is meant when the records speak of "lords of 10,000 households," however, cannot now be fathomed.

A clearer picture has emerged of the ethnic lords incorporated by the Inca into their realm. Some had ruled only small units—a few hundred households; others, like the Huanca or the Lupaca claimed to have had 20,000 domestic units. There is no record of the size of the coastal Chimú polity, which must have been quite large. The Chincha claimed 30,000 "fires," and the Chimú may well have been even larger before their defeat by Cuzco.

Usually, two lords ruled each ethnic group—which has been one of the arguments for considering as plausible a dual rule in Cuzco as well. The best evidence of the duties of the ethnic lords has come from the Aymara kingdom of the Lupaca: at one point in Inca history they rose in rebellion against Cuzco rule, and in the decades immediately prior to the arrival of the Europeans they were busy leading "6,000 soldiers" on faraway battlefields in what is now Ecuador. The testimony of the Lupaca, collected in 1567, claims that on such adventures they did not return to their lands for the harvest but devoted most of their energies to war, and in return they were exempted from farming, road building, and other state chores.

There was no tribute system in Inca statecraft, just as there had been no contributions in kind in earlier Andean polities. The peasantry owed only their energy, which was delivered through the well-understood *mit'a* system. Led by their traditional leaders, the people appeared for their obligations, lineage by lineage. The best quipu record of these obligations has come from a group who lived in the Huánuco area. Just as they had provided energy for their own lords, under Inca rule this group sent dozens of couples to labour on public works or to produce the grain that, as beer, was "fed" to the mummies of deceased Inca kings. Others became soldiers or helped fill the warehouses; some carried loads along the Inca highway system, while still others were soldiers under the command of their traditional lords. Using this quipu, it has been possible to test the claim that there was no tribute system: of its 26 cords only two deal with articles submitted in kind, wild honey and tropical feathers, both of which were lowland commodities that were gathered and not cultivated.

The absence of tribute was closely connected to the absence of markets. Just as all households owed some of their energies to their ethnic lords, to the shrines, and to Cuzco, so too their household needs were satisfied by the claims they could make to the reciprocal services of their kinfolk or their ethnic peers or to the administrative services of their ethnic authorities. It is probable that with the growth of the Inca state over time, this formula was breached, particularly in the case of prisoners of war and other populations moved from their traditional areas for state purposes.

The most elaborate example of the structural changes that emerged from the need to create new state revenues was the expansion and reorganization of corn production for military purposes in the Cochabamba Valley. This region was the largest single corn-producing area in the highlands. One of the later kings removed the native population and set up a large state enterprise (more than 2,000 warehouses), to which some 25 highland groups were sent on rotation, lineage by lineage. Each ethnic group was responsible for particular strips that were traced across the valley by Cuzco surveyors. In 1575 the Spanish viceroy Francisco de Toledo used this Inca precedent to establish the repartimiento system that provided labour for the silver mines at Potosí.

Inca technology and intellectual life. The intellectual tradition of the Inca emerged from their detailed and efficient knowledge and use of an extremely challenging environment. No system of writing, in the European sense, has been discovered, and the question remains as to how long-distance communication was achieved.

Textiles

Beyond oral transmission, the most promising domain for research is in textiles. In the highlands very few have been preserved because of the humidity, but on the coastal desert many burial cloths from widely different periods have been located and studied. Their artistic qualities have

Craig Morris



Large hall on the eastern side of the main plaza at Huánuco Pampa, Peru.

fomented grave robbing on a very large scale; museums throughout the world have dozens if not hundreds of such cloths, each of great beauty and enormous sophistication.

Fibre technology went beyond burial or sacrificial textiles: Viceroy Toledo wrote to Philip II that he was sending four gigantic cloths on which maps of his Andean realm had been painted. While the letter was carefully filed in the Archives of the Indies, at Seville, the maps have never been located. Other uses of textiles included the quipu used for bookkeeping and possibly also for historical recording; suspension bridges, some of which are still maintained on a regular basis by particular villages responsible for reweaving; and calendars and ceremonial accounting.

While in the field, Inca armies were rewarded with corn and cloth. One European observer was told that soldiers would rebel if they did not receive their issues of textiles and corn beer. A major manufacturing centre employing "a thousand" full-time weavers was established on the northeastern shore of Lake Titicaca. The craftspeople there were men, but every administrative centre along the Inca highway is said to have housed a group of secluded women weavers (Chosen Women); one such house, at Huánuco Pampa (administrative centre of the Huánuco region), has been located and excavated. The storehouses, full of thousands of textiles, were one of the wonders frequently mentioned by the early Spaniards in their letters.

As Tawantinsuyu grew and involved peoples of many different environments and cultures, techniques originating in any particular ethnic group were spread across the land. Prior to the Inca expansion, metals—gold, silver, copper, and their alloys—were used mainly for ornaments; and tools were made from wood and stone. Bronze tools—crowbars, chisels, axes, knives, and clubheads, to name only a few—became exceedingly common after the Inca conquest.

The Inca highway system

The remarkable Inca highway system was also noted by the earliest Spanish eyewitnesses, since these roads were in constant use, even by horses. Research since the 1950s has provided fresh insights into the engineering methods and geographic location of two parallel roads—one in the highlands, the other on the coast—the whole system adding up to at least 15,500 miles. While some of these roads may have been built first during the Middle Horizon and even earlier, it was during Inca times that the roads were maintained and unified into a single political and economic system. Travel units, adjusted to the pace of a loaded llama or human carrier, can still be detected along the Qhapaq Ñan, the main north–south royal road in the highlands. At the end of each day the caravan stopped at a tambo, a way station, which, although smaller than an administrative centre, was complete with warehouses and barracks. The maintenance of the road segment and the filling of storehouses was part of the *mit'a* responsibilities of neighbouring groups.

Measurement of both distance and surface area was done by units called *tupu*, since the Andean concern was with units of human energy expended. Somehow, two measurements that belonged to very different European systems of reckoning were part of a single Andean concern. Units of land measurement, called *papakancha*, also differed: where the land was in continuous cultivation, as in corn country, one unit was used; another unit was in use for highland-tuber cultivation, where fallowing and rotation was the dominant crop pattern. As one "measurer" explained to the viceroy's envoy, the *papakancha* was of one size when it was at a protected, lower altitude, but it could be up to seven times that size on the high, cold puna.

(T.C.P./J.V.M.)

Inca religion. Inca religion—an admixture of complex ceremonies, practices, animistic beliefs, varied forms of belief in objects having magical powers, and nature worship—culminated in the worship of the sun, which was presided over by the priests of the last native pre-Columbian conquerors of the Andean regions of South America. Though there was an Inca state religion of the sun, the substrata religious beliefs and practices of the pre-Inca peoples exerted an influence on the Andean region prior to and after the conquest of most of South America by the Spaniards in the 16th century.

Inca gods. The creator god of the Inca and of pre-Inca peoples was Viracocha, who was also a culture hero. Creator of earth, man, and animals, Viracocha had a long list of titles, including Lord Instructor of the World, the Ancient One, and the Old Man of the Sky. Some have said that he also was the creator of the Tiahuanaco civilizations, of which the Inca were the cultural heirs. Viracocha went through several transmogrifications (often with grotesque or humorous effects). He made peoples, destroyed them, and re-created them of stone; and when they were re-created, he dispersed mankind in four directions. As a culture hero he taught people various techniques and skills. He journeyed widely until he came to the shores of Manta (Ecuador), where he set off into the Pacific—some say in a boat made of his cloak, others that he walked on the water. This part of the myth has been seized upon by modern mythmakers, and, as Kon-Tiki, Viracocha was said to have brought Inca culture to Polynesia.

Viracocha was the divine protector of the Inca ruler Pachacuti Inca Yupanqui; he appeared to Pachacuti in a dream when the Inca forces were being besieged by the Chanca. Upon victory, Pachacuti raised a temple to Viracocha in Cuzco. He was represented by a gold figure "about the size of a 10-year-old child."

Inti, the sun god

Inti, the sun god, was the ranking deity in the Inca pantheon. His warmth embraced the Andean earth and matured crops; and as such he was beloved by farmers. Inti was represented with a human face on a ray-splayed disk. He was considered to be the divine ancestor of the Inca: "my father" was a title given to Inti by one Inca ruler.

Apu Illapu, the rain giver, was an agricultural deity to whom the common man addressed his prayers for rain. Temples to Illapu were usually on high structures; in times of drought, pilgrimages were made to them and prayers were accompanied by sacrifices—often human, if the crisis was sufficient. The people believed that Illapu's shadow was in the Milky Way, from whence he drew the water that he poured down as rain.

Mama-Kilya, wife of the sun god, was the Moon Mother, and the regulator of women's menstrual cycles. The waxing and waning of the moon was used to calculate monthly cycles, from which the time periods for Inca festivals were set. Silver was considered to be tears of the moon. The stars had minor functions. The constellation of Lyra, which was believed to have the appearance of a llama, was entreated for protection. The constellation Scorpio was believed to have the shape of a cat; the Pleiades were called "little mothers," and festivals were celebrated on their reappearance in the sky. Earth was called Paca-Mama, or "Earth Mother." The sea, which was relatively remote to the Inca until after 1450, was called Mama Qoca, the Sea Mother.

Temples and shrines. Temples and shrines housing fetishes of the cult were occupied by priests, their attendants, and the Chosen Women. In general, temples were not intended to shelter the celebrants, since most ceremonies were held outside the temple proper. The ruins of the Temple of Viracocha at San Pedro Cacha (Peru), however, had a ground plan that measured 330 by 87 feet, which indicates that it was designed for use other than the storage of priestly regalia.

The Sun Temple

The Sun Temple in Cuzco is the best-known of the Inca temples. Another, at Vilcashuman (which was regarded as the geographic centre of the empire), has a large temple still existing. Near Mount Aconcagua in Argentina, at the southern limit of the Inca Empire, "there was a temple . . . an ancient oracle held in high regard where they made their sacrifices," and on Titicaca Island, one of the largest of several islands in Lake Titicaca, there was a temple of the sun.

As the Inca conquered new territories, temples were erected in the new lands. In Caranqui, Ecuador, one such temple was described by a chronicler as being filled with great vessels of gold and silver. At Latacunga (Llacta cunga) in Ecuador there was a sun temple where sacrifices were made; part of the temple was still visible when the German explorer and geographer Alexander von Humboldt sketched the ruins in 1801.

The Sun Temple in Cuzco, built with stones "all matched

and joined," had a circumference of more than 1,200 feet. A fragment of the wall still extant is testimony to the accuracy of the chronicler's description. Within the temple was an image of the sun "of great size," and in another precinct, the Golden Enclosure (Corincancha), were gold models of cornstalks, llamas, and lumps of earth. Portions of the land, which supported the temples, the priests, and the Chosen Women, were allotted to the sun and administered for the priests.

Along with the shrines and temples, huacas (sacred sites) were widespread. A huaca could be a man-made temple, mountain, hill, or bridge, such as the great *huacachaca* across the Apurímac River. A huaca also might be a mummy bundle, especially if it was that of a lord-Inca. On high points of passage in the Andes, propitiatory cairns (*apacheta,* "piles of stones") were made, to which, in passing, each person would add a small stone and pray that his journey be lightened. The idea of huaca was intimately bound up with religion, combining the magical and the charm-bearing.

The priesthood. Priests resided at all important shrines and temples. A chronicler suggests that a priest's title was *umu,* but in usage his title was geared to his functions as diviner of lungs, sorcerer, confessor, and curer. The title of the chief priest in Cuzco, who was of noble lineage, was *villac umu.* He held his post for life, was married, and competed in authority with the Inca. He had power over all shrines and temples and could appoint and remove priests. Presumably, priests were chosen young, brought up by the more experienced, and acquired with practice the richly developed ceremonialism.

Functions of priests

Divination. Divination was the prerequisite to all action. Nothing of importance was undertaken without recourse to divination. It was used to diagnose illness, to predict the outcome of battles, and to ferret out crimes, thus giving it a judiciary function. Divination was also used to determine what sacrifice should be made to what god. Life was believed to be controlled by the all-pervading unseen powers, and to determine these portents the priests had recourse to the supernatural. Oracles were considered to be the most important and direct means of access to the wayward gods. One oracle of a huaca close to the Huaca–Chaca Bridge, across the Apurímac River near Cuzco, was described by a chronicler as a wooden beam as thick as a fat man, with a girdle of gold about it with two large golden breasts like a woman. These and other idols were bloodspattered from sacrifices—animal and human. "Through this large idol," a chronicler wrote, "the demon of the river used to speak to them." Another well-known oracle was housed in a temple in the large adobe complex of Pachacamac near Lima.

Divination also was accomplished by watching the meandering of spiders and the arrangement that coca leaves took in a shallow dish. Another method of divination was to drink *ayahuasca,* a narcotic that had profound effects on the central nervous system. This was believed to enable one to communicate with the supernatural powers.

Fire also was believed to provide spiritual contact. The flames were blown to red heat through metal tubes, after which a practitioner (*yacarca*) who had narcotized himself by chewing coca leaves summoned the spirits with fiery conjuration to speak—"which they did," wrote a chronicler, by "ventriloquism." Divination by studying the lungs of a sacrificed white llama was considered to be efficacious. The lungs were inflated by blowing into the dissected trachea (there is an Inca ceramic showing this), and the future was foretold by priests who minutely observed the conformance of the veins. On the reading of this augury, political or military action was taken.

The role of confession

Confession was part of the priestly ritual of divination. Should rain not fall or a water conduit break without cause, it was believed that such an occurrence could arise from someone's failure to observe the strictly observed ceremonies. This was called *hocha,* a ritual error. The *ayllu,* a basic social unit identified with communally held land, was wounded by individual misdeeds. Crimes had to be confessed and expiated by penitence so as not to call down the divine wrath.

Sacrifice. Sacrifice, human or animal, was offered on every important occasion; guinea pigs (more properly *cui*), llamas, certain foods, coca leaves, and *chicha* (an intoxicant corn beverage) were all used in sacrifices. Many sacrifices were daily occurrences for the ritual of the sun's appearance. A fire was kindled, and corn was thrown on the coals and toasted. "Eat this, Lord Sun," was the objuration of officiating priests, "so that you will know that we are your children." On the first day of every lunar month 100 pure-white llamas were driven into the Great Square, Huayaca Pata in Cuzco; they were moved about to the various images of the gods and then assigned to 30 priestly attendants, each representing a day of the month. The llamas were then sacrificed; chunks of flesh were thrown onto the fire, and the bones were powdered for ritual use. Ponchos of excellent weave or miniature vestments were burned in the offering. The Inca ruler wore his poncho only once; it was ceremoniously sacrificed in fire each day.

Humans also were sacrificed; when the need was extreme, 200 children might be immolated, such as when a new Inca ruler assumed the royal fringe. Defeats, famine, and pestilence all called for human blood. Even a Chosen Woman from the Sun Temple might be taken out for sacrifice. Children, before being sacrificed, were feasted "so that they would not enter the presence of the gods hungry and crying." It was important in human sacrifice that the sacrificed person be without blemish. Many were chosen from the conquered provinces as part of regular taxation; "blood money" was scarcely a metaphor.

Festivals. The 30-day calendar was religious, and each month had its own festival. The religious calendar is explained in considerable detail by Guamán Poma de Ayala (see Table 3). In his letter to Philip II he offered two different versions, one centring on state ceremonies and sacrifices performed at Cuzco and the other describing the agricultural practices at the local level in the highlands. Quite different calendars prevailed on the irrigated coast, but surviving sources do not record them in any detail.

(V.W.v.H./J.V.M.)

Table 3: Months and Celebrations of the Inca Calendar

Gregorian months	Andean months	approximate translation
December	Capac Raimi, Capac Quilla	the lord festival; the month of rest
January	Zarap Tuta Cavai Mitan	the time to watch the growing corn
February	Paucar Varai	the time to wear loincloths
March	Pacha Pucuy Quilla	the month of the land's maturation
April	Camai Quilla [Inti Raymi in state calendar]	the month of harvest and rest
May	Zara Muchuy Quilla Aymoray Quilla	dry corn to be stored
June	Papa Allai Mitan Pacha Haucai Cusqui	potato harvest rest from harvesting
July	Chacra Conaqui Quilla	the month of redistributing lands
August	Chacra Yapuy Quilla Hailly	the month to open lands coming into cultivation with songs of triumph
September	Zara Tarpuy Quilla Coia Raymi Quilla	the month for planting; also, the Festival of the Queen
October	Chacramanta Pisco Carcoy	the time to scare birds out of newly planted fields
November	Chacra Parcay	the time to irrigate fields

BIBLIOGRAPHY. GORDON R. WILLEY, *An Introduction to American Archaeology,* 2 vol. (1966–71); and GORDON R. WILLEY and JEREMY A. SABLOFF, *A History of American Archaeology,* 2nd ed. (1980), provide overviews of all of New World prehistory and place the Mesoamerican and Andean civilizations into the larger cultural-historical setting. See also LESLIE BETHELL (ed.), *The Cambridge History of Latin America,* vol. 1 (1984).

Meso-American civilization: (General works): The comprehensive, multivolume series "Handbook of Middle American Indians," especially vol. 2–3, GORDON R. WILLEY (ed.), *Archaeology of Southern Mesoamerica* (1965), and vol. 10–11, GORDON F. EKHOLM and IGNACIO BERNAL (eds.), *Archaeology of Northern Mesoamerica* (1971), is indispensable. Summaries of more recent developments are found in JEREMY A. SABLOFF (ed.), *Archaeology* (1981), vol. 1 in the series "Supplement to the Handbook of Middle American Indians." Historical sources include FRANCIS AUGUSTUS MACNUTT (trans. and ed.), *Letters of Cortes: The Five Letters of Relation from Fernando Cortes to the Emperor Charles V,* 2 vol. (1908, reprinted 1977); and BERNAL DÍAZ DEL CASTILLO, *The True History of the Conquest*

of New Spain, 5 vol., ed. and trans. by ALFRED PERCIVAL MAUDSLEY (1908–16, reprinted as *The Conquest of New Spain*, 4 vol., 1967). See also MURIEL PORTER WEAVER, *The Aztecs, Maya, and Their Predecessors: Archaeology of Mesoamerica*, 2nd ed. (1981); WALTER KRICKEBERG, *Altmexikanische Kulturen*, new ed. (1975); JACQUES SOUSTELLE, *The Four Suns: Recollections and Reflections of an Ethnologist in Mexico* (1971; originally published in French, 1967), and *Arts of Ancient Mexico* (1966; originally published in French, 1966); SUZANNE ABEL-VIDOR *et al.*, *Between Continents/Between Seas: Precolumbian Art of Costa Rica* (1981); ANNA BENSON GYLES and CHLOË SAYER, *Of Gods and Men: Mexico and the Mexican Indian* (1980); MICHAEL D. COE, *Mexico*, 3rd rev. ed. (1984); and JOYCE KELLY, *The Complete Visitor's Guide to Mesoamerican Ruins* (1982).

(*Preclassic and Classic periods*): For early prehistory and the beginnings of agriculture, see CHRISTINE NIEDERBERGER, *Zohapilco: Cinco milenios de ocupación humana en un sitio lacustre de la Cuenca de México* (1976); and BARBARA L. STARK and BARBARA VOORHIES (eds.), *Prehistoric Coastal Adaptations: The Economy and Ecology of Maritime Middle America* (1978). General introductions include WILLIAM T. SANDERS and BARBARA J. PRICE, *Mesoamerica: The Evolution of a Civilization* (1968); RICHARD E.W. ADAMS, *Prehistoric Mesoamerica* (1977); ERIC R. WOLF, *Sons of the Shaking Earth* (1959, reprinted 1974); and MARY W. HELMS, *Middle America: A Culture History of Heartland and Frontiers* (1975, reprinted 1982). For a coordination of regional chronologies within the area, see R.E. TAYLOR and CLEMENT W. MEIGHAN (eds.), *Chronologies in New World Archaeology* (1978). Books dealing with Olmec civilization include MICHAEL D. COE and RICHARD A. DIEHL, *In the Land of the Olmec*, 2 vol. (1980); MICHAEL D. COE, DAVID GROVE, and ELIZABETH P. BENSON (eds.), *The Olmec & Their Neighbors* (1981); and IGNACIO BERNAL, *The Olmec World* (1969; originally published in Spanish, 1968). For other cultures of the period, see the general works cited above.

(*Late Classic Period*): General books on the Maya include J. ERIC S. THOMPSON, *The Rise and Fall of Maya Civilization*, 2nd enl. ed. (1966, reprinted 1977); MICHAEL D. COE, *The Maya*, 4th rev. ed. (1987); NORMAN HAMMOND, *Ancient Maya Civilization* (1982); and SYLVANUS G. MORLEY and GEORGE W. BRAINERD, *The Ancient Maya*, 4th ed., rev. by ROBERT J. SHARER (1983). The rise and decline of Maya civilization are discussed in RICHARD E.W. ADAMS (ed.), *The Origins of Maya Civilization* (1977); and T. PATRICK CULBERT (ed.), *The Classic Maya Collapse* (1973, reprinted 1983). See also JOHN S. HENDERSON, *The World of the Ancient Maya* (1981), an ethnohistory. The transition from Maya Classic to Postclassic culture is examined in JEREMY A. SABLOFF and E. WYLLIS ANDREWS V (eds.), *Late Lowland Maya Civilization: Classic to Postclassic* (1986). A basic work on Maya hieroglyphic writing is J. ERIC S. THOMPSON, *Maya Hieroglyphic Writing: An Introduction*, 3rd ed. (1971). For later research in the field, see DAVID HUMISTON KELLEY, *Deciphering the Maya Script* (1976); ELIZABETH P. BENSON (ed.), *Mesoamerican Writing Systems* (1973); and LINDA SCHELE and MARY ELLEN MILLER, *The Blood of Kings: Dynasty and Ritual in Maya Art* (1986). For the complex subject of ancient Maya religion, J. ERIC S. THOMPSON, *Maya History and Religion* (1970), is an indispensable introductory guide. Native and early historical sources include DANIEL G. BRINTON (ed. and trans.), *The Maya Chronicles* (1882, reprinted 1969), and *The Annals of the Cakchiquels* (1885, reprinted 1969); *Popol Vuh: The Sacred Book of the Ancient Quiché Maya*, trans. from the Spanish work of ADRIÁN RECINOS by DELIA GOETZ and SYLVANUS G. MORLEY (1950, reprinted 1978); RALPH L. ROYS (trans.), *The Book of Chilam Balam of Chumayel* (1933, reprinted 1967); DIEGO DE LANDA, *Landa's Relación de las cosas de Yucatán*, trans. into English and ed. by ALFRED M. TOZZER (1941, reprinted 1978); and MUNRO S. EDMONSON (trans. and ed.), *The Ancient Future of the Itza: The Book of Chilam Balam of Tizimin* (1982). JACQUES SOUSTELLE, *Mexico*, trans. from French (1967), deals with the relation of Maya and other ancient Mesoamerican religions. Economic and demographic features are treated in PETER D. HARRISON and B.L. TURNER II, *Pre-Hispanic Maya Agriculture* (1978); and WENDY ASHMORE (ed.), *Lowland Maya Settlement Patterns* (1981). See also HENRI STIERLIN, *Art of the Maya: From the Olmecs to the Toltec-Maya* (1981; originally published in French, 1981).

(*Postclassic Period*): BERNARDINO DE SAHAGÚN, *General History of the Things of New Spain: Florentine Codex*, trans. from Náhuatl and ed. by ARTHUR J.O. ANDERSON and CHARLES E. DIBBLE, 13 vol. in 12 (1950–82), is the first full translation of this 16th-century Spanish writer of Aztec culture and is particularly informative on Aztec religion. J. ERIC S. THOMPSON, *Mexico Before Cortez: An Account of the Daily Life, Religion, and Ritual of the Aztecs and Kindred Peoples* (1933), is a shorter survey. See also JACQUES SOUSTELLE, *The Daily Life of the Aztecs: On the Eve of the Spanish Conquest* (1962, reissued 1970; originally published in French, 1955); RICHARD A. DIEHL, *Tula: The Toltec Capital of Ancient Mexico* (1983); NIGEL DAVIES, *The Toltecs, Until the Fall of Tula* (1977), and *The Toltec Heritage: From the Fall of Tula to the Rise of Tenochtitlán* (1980); and BURR CARTWRIGHT BRUNDAGE, *The Fifth Sun: Aztec Gods, Aztec World* (1979), and *The Phoenix of the Western World: Quetzalcoatl and the Sky Religion* (1982).

Andean civilization: (*General works*): A comprehensive and still useful survey is provided by the articles in *The Andean Civilizations*, vol. 2 of JULIAN H. STEWARD (ed.), *Handbook of South American Indians*, 7 vol. (1946–59, reprinted 1963). Introductions include JOHN HOWLAND ROWE and DOROTHY MENZEL (eds.), *Peruvian Archaeology: Selected Readings* (1967); LUIS G. LUMBRERAS, *The People and Cultures of Ancient Peru* (1974; originally published in Spanish, 1969); and RICHARD W. KEATINGE (ed.), *Peruvian Prehistory* (1988). Important research by Andean authors is presented in RAMIRO CONDARCO MORALES, *El escenario andino y el hombre* (1971); and SEGUNDO MORENO YAÑEZ and UDO OBERMEM, *Contribución a la etnohistoria ecuatoriana* (1981). See also JOHN A. MASON, *The Ancient Civilizations of Peru* (1968, reprinted 1979); JOHN V. MURRA, NATHAN WACHTEL, and JACQUES REVEL (eds.), *Anthropological History of Andean Polities* (1986; originally published in French in issues 5–6 of *Annales, Economies, Societés, Civilizations*, vol. 33, 1978); and SHOZO MASUDA, IZUMI SHIMADA, and CRAIG MORRIS (eds.), *Andean Ecology and Civilization: An Interdisciplinary Perspective on Andean Ecological Complementarity* (1985). For the material culture admired by the Spanish observers from the first days of the invasion, see CHRISTOPHER B. DONNAN (ed.), *Early Ceremonial Architecture in the Andes* (1985); HEATHER LECHTMAN and ANA MARÍA SOLDI (eds.), *La tecnología en el mundo andino: subsistencia y mensuración* (1981); and ANN POLLARD ROWE (ed.), *The Junius B. Bird Conference on Andean Textiles, April 7–8, 1984* (1986).

(*Pre-Inca periods*): Surveys of these civilizations include WENDELL C. BENNETT and JUNIUS B. BIRD, *Andean Culture History*, 2nd rev. ed. (1960, reissued 1964); GEOFFREY H.S. BUSHNELL, *Peru*, rev. ed. (1963); EDWARD P. LANNING, *Peru Before the Incas* (1967); DONALD W. LATHRAP, *The Upper Amazon* (1970); and JONATHAN HAAS, SHELIA POZORSKI, and THOMAS POZORSKI (eds.), *The Origins and Development of the Andean State* (1987).

(*The Inca*): The works of the chroniclers are available in modern English translations or in modern critical editions: PEDRO DE CIEZA DE LEÓN, *The Incas*, trans. by HARRIET DE ONIS, ed. by VICTOR WOLFGANG VON HAGEN (1959, reprinted 1967); GARCILASO DE LA VEGA, *The Incas: The Royal Commentaries of the Inca, Garcilaso de la Vega, 1539–1616*, trans. by MARIA JOLAS from the critical annotated French edition of ALAIN GHEERBRANT (1961, reissued 1971; originally published in French, 1959); and FELIPE GUAMÁN POMA DE AYALA, *El primer nueva corónica y buen gobierno*, trans. from Quechua by JORGE L. URIOSTE, ed. by JOHN V. MURRA and ROLENA ADORNO, 3 vol. (1980). Comprehensive histories include ALFRED MÉTRAUX, *The History of the Incas* (1969; originally published in French, 1963); GEORGE A. COLLIER, RENATO I. ROSALDO, and JOHN D. WIRTH (eds.), *The Inca and Aztec States, 1400–1800: Anthropology and History* (1982); and JOHN HEMMING, *The Conquest of the Incas*, rev. ed. (1983).

The following works study special aspects of the Inca civilization: R. TOM ZUIDEMA, *La civilisation inca au Cuzco* (1986), on social life and kinship systems; FRANKLIN PEASE (GARCÍA YRIGOYEN), *El dios creador andino* (1973), on religion; GARY URTON, *At the Crossroads of the Earth and the Sky: An Andean Cosmology* (1981); MARIA ASCHER and ROBERT ASCHER, *Code of the Quipu: A Study in Media, Mathematics, and Culture* (1981); FRANK SALOMON, *Native Lords of Quito in the Age of the Incas: The Political Economy of North-Andean Chiefdoms* (1986); CRAIG MORRIS and DONALD E. THOMPSON, *Huánuco Pampa: An Inca City and Its Hinterland* (1985); SALLY FALK MOORE, *Power and Property in Inca Peru* (1958, reprinted 1973); JOHN V. MURRA, *The Economic Organization of the Inka State* (1980); JOHN HYSLOP, *The Inca Road System* (1984); and GRAZIANO GASPARINI and LUISE MARGOLIES, *Inca Architecture* (1980; originally published in Spanish, 1977).

(G.R.W./W.T.Sa./J.V.M.)

Prehistoric Peoples and Cultures

This article provides a general account of man's cultural-historical development during the Pleistocene Ice Age and the early Holocene, or Recent, Epoch. During these approximately 500,000 years of prehistoric human development, the climates and environments of the world fluctuated considerably, and there were, of course, no national boundaries or ethnological regions that conformed in any meaningful way to those of the present.

It would be impossible to maintain, on the basis of the archaeological evidence, that the inhabitants of what is now France were already Frenchmen 25,000 years ago or that the inhabitants of what is now Zimbabwe were already Bantu-speaking peoples 250,000 years ago. Hence, the approach of this article is one of an overview of vast regions and of the great, general steps forward that human culture made from earliest times until the civilizations of the conventional ancient historians were initiated.

The article is divided into the following sections:

Cultural history of the Stone Ages 45
- Europe 46
  - Paleolithic 46
    - Lower Paleolithic
    - Middle Paleolithic
    - Upper Paleolithic
  - Mesolithic 48
    - The cultures
    - The general picture
  - Neolithic 49
    - The zones
    - Cultural elements
- Asia 52
  - Paleolithic 52
    - Middle East
    - Central Asia
    - South Asia
    - Far East
    - Siberia
  - Mesolithic–Neolithic: the rise of village-farming communities 53
    - Middle East
    - South and East Asia
    - Central Asia and Siberia
- Africa 56
  - Paleolithic 56
    - North Africa
    - Egypt
    - East Africa
    - Southern Africa
    - Central Africa
  - Mesolithic–Neolithic 57
- The Americas 58
  - Early cultures 58
    - Paleo-Indian tradition
    - Desert tradition
    - Archaic tradition
    - Western North America
    - South and Middle America
  - Rise of agriculture 59
  - Other developments 59
    - Village farming and towns 59
      - Hopewell culture
      - Mississippi culture
      - Pueblos
      - South America
      - The general picture
- Oceania 61

Civilizations 62
- The Urban Revolution 62
- The example of the Middle East 63
  - Evolution of Middle Eastern civilizations 63
    - Mesopotamia and Egypt to *c.* 1600 BC
    - New states and peoples
    - The Achaemenian Empire and its successors
    - Pre-Islāmic Arabia
  - Elements of civilized culture 65
    - Religion
    - Science and law
    - The alphabet

Prehistoric religion 66
- Practices and beliefs 66
  - Burial customs and cults of the dead 66
  - Cannibalism 66
  - Sacrifices 67
  - Hunting rites and animal cults 67
  - Female fertility deities 68
  - Shamanism, sorcery, and magic 68
- Evolutionary development 69
  - Stone Age cultures 69
    - Lower or Early and Middle Paleolithic
    - Upper Paleolithic and Mesolithic
    - Proto-Neolithic and Neolithic
  - Civilizations 69

Bibliography 69

CULTURAL HISTORY OF THE STONE AGES

Paleolithic archaeology is concerned with the origins and development of early human culture between the first appearance of man as a tool-using mammal, which is believed to have occurred about 600,000 or 700,000 years ago, and the beginning of the Recent geologic era, about 8000 BC. It is included in the time span of the Pleistocene, or Glacial, Epoch—an interval of about 1,000,000 years. Although it cannot be proved, modern evidence suggests that the earliest protohuman forms had diverged from the ancestral primate stock by the beginning of the Pleistocene. In any case, the oldest recognizable tools are found in horizons of Lower Pleistocene Age. During the Pleistocene a series of momentous climatic events occurred. The northern latitudes and mountainous areas were subjected on four successive occasions to the advances and retreats of ice sheets (known as Günz, Mindel, Riss, and Würm in the Alps), river valleys and terraces were formed, the present coastlines were established, and great changes were induced in the fauna and flora of the globe. In large measure, the development of culture during Paleolithic times seems to have been profoundly influenced by the environmental factors that characterize the successive stages of the Pleistocene Epoch.

Environmental factors

Throughout the Paleolithic, man was a food gatherer, depending for his subsistence on hunting wild animals and birds, fishing, and collecting wild fruits, nuts, and berries. The artifactual record of this exceedingly long interval is very incomplete; it can be studied from such imperishable objects of now-extinct cultures as were made of flint, stone, bone, and antler. These alone have withstood the ravages of time, and, together with the remains of contemporary animals hunted by our prehistoric forerunners, they are all that scholars have to guide them in attempting to reconstruct human activity throughout this vast interval—approximately 98 percent of the time span since the appearance of the first true hominid stock. In general, these materials develop gradually from single, all-purpose tools to an assemblage of varied and highly specialized types of artifacts, each designed to serve in connection with a specific function. Indeed, it is a process of increasingly

more complex technologies, each founded on a specific tradition, that characterizes the cultural development of Paleolithic times. In other words, the trend was from simple to complex, from a stage of nonspecialization to stages of relatively high degrees of specialization, just as has been the case during historic times.

In the manufacture of stone implements, four fundamental traditions were developed by the Paleolithic ancestors: (1) pebble-tool traditions; (2) bifacial-tool, or hand-ax, traditions; (3) flake-tool traditions; and (4) blade-tool traditions. Only rarely are any of these found in "pure" form, and this fact has led to mistaken notions in many instances concerning the significance of various assemblages. Indeed, though a certain tradition might be superseded in a given region by a more advanced method of producing tools, the older technique persisted as long as it was needed for a given purpose. In general, however, there is an overall trend in the order as given above, starting with simple pebble tools that have a single edge sharpened for cutting or chopping. But no true pebble-tool horizons had yet, by the late 20th century, been recognized in Europe. In southern and eastern Asia, on the other hand, pebble tools of primitive type continued in use throughout Paleolithic times.

French place-names have long been used to designate the various Paleolithic subdivisions, since many of the earliest discoveries were made in France. This terminology has been widely applied in other countries, notwithstanding the very great regional differences that do in fact exist. But the French sequence still serves as the foundation of Paleolithic studies in other parts of the Old World.

(H.L.Ms./Ed.)

The Recent (Holocene) Era

There is reasonable agreement that the Paleolithic ended with the beginning of the Recent (Holocene) geologic and climatic era about 8000 BC. It is also increasingly clear that a developmental bifurcation in man's culture history took place at about this time. In most of the world, especially in the temperate and tropical woodland environments or along the southern fringes of Arctic tundra, the older Upper Paleolithic traditions of life were simply readapted toward more or less increasingly intensified levels of food collection. These cultural readaptations of older food procedures to the variety and succession of post-Pleistocene environments are generally referred to as occurring in the Mesolithic Period. But also by 8000 BC (if not even somewhat earlier) in certain semi-arid environments of the world's middle latitudes, traces of a quite different course of development began to appear. These traces indicate a movement toward incipient agriculture and (in one or two instances) animal domestication. In the case of southwestern Asia, this movement had already culminated in a level of effective village-farming communities by 7000 BC. In Meso-America, a comparable development—somewhat different in its details and without animal domestication—was taking place almost as early. It may thus be maintained that in the environmentally favourable portions of southwestern Asia, Meso-America, the coastal slopes below the Andes, and perhaps in southeastern Asia (for which little evidence is available), little if any trace of the Mesolithic stage need be anticipated. The general level of culture probably shifted directly from that of the Upper Paleolithic to that of incipient cultivation and domestication.

The picture presented by the culture history of the earlier portion of the Recent period is thus one of two generalized developmental patterns: (1) the cultural readaptations to post-Pleistocene environments on a more or less intensified level of food collection; and (2) the appearance and development of an effective level of food production. It is generally agreed that this latter appearance and development was achieved quite independently in various localities in both the Old and New Worlds. As the procedures and the plant or animal domesticates of this new food-producing level gained effectiveness and flexibility to adapt to new environments, the new level expanded at the expense of the older, more conservative one. Finally, it is only within the matrix of a level of food production that any of the world's civilizations have been achieved.

(R.J.Br./Ed.)

Europe

PALEOLITHIC

Three major subdivisions—Lower, Middle, and Upper Paleolithic—are recognized in Europe. Although the dividing line between the Lower and Middle stages is not so clearly defined as that separating the Middle and Upper subdivisions, this system is still used by most workers.

Lower Paleolithic. On the basis of the very rich materials from the Somme Valley in the north of France and the Thames Valley in the south of England, two main Lower Paleolithic traditons have been recognized in western Europe. These are as follows: (1) bifacial-tool, or hand-ax, traditions (Abbevillian and Acheulean); and (2) flake-tool traditions (Clactonian and Levalloisian).

Abbevillian industry

The type tools of the Abbevillian (formerly Chellean), which takes its name from the town of Abbeville, France, on the 45-metre (150-foot) terrace of the Somme Valley, consist of pointed, bifacial implements, or hand axes. Their forms vary, and the flaking is generally irregular; it is probable that they were manufactured either with a stone hammer or on a stone anvil. Associated with these crude types of hand axes, simple flake tools are found, but they lack definite form. The Abbevillian has been reported from deposits of lower Pleistocene (First Interglacial) age.

The Acheulean, which begins in the Second Interglacial and persists to the close of the Third Interglacial, covers by far the longest time span of any of the Paleolithic traditions found in western Europe. The type site is on the 30-metre terrace of the Somme Valley at St. Acheul, near Amiens, in northern France. Acheulean hand axes, which display a marked technological refinement over their Abbevillian precursors, were apparently made by employing a wooden or bone billet rather than the more primitive stone-on-stone technique. But, except at the very end of the Acheulean cycle of development, there is very little typological difference in the types of hand axes found in the various layers.

The Micoquian, or Final (Upper) Acheulean, is characterized by elongated hand axes that exhibit very straight and finely chipped edges, in marked contrast with the Lower Acheulean, in which ovate forms predominate. Flake tools occur in all Acheulean levels, the side scrapers being the predominant type. Many of these tools were made from trimming flakes produced during the process of hand-ax manufacture. In general, flake tools, including points with a triangular cross section, are found in greater quantities in Micoquian deposits than in the older horizons.

The evidence from Clacton-on-Sea, Essex, and Swanscombe, Kent, in the Thames Valley of southeastern England clearly shows that the main development of the Clactonian occurred during early Second Interglacial times. The type artifacts are flakes, although core tools—single-edged choppers and chopping tools—do in fact occur. The flakes, which have large, high-angle (greater than 90°), plain striking platforms, as well as prominent bulbs of percussion, were detached from roughly prepared, discoidal cores by the stone-hammer or stone-anvil technique. Actual retouching or secondary working of the edge is found in some instances, but for the most part it is crude, and edge chipping resulting from use is far more characteristic.

Levalloisian tradition

Named after a locality at Levallois, a suburb of Paris, the Levalloisian is primarily a flake tradition, although hand axes are found in certain of the Middle and Upper Levalloisian stages. It first appears in deposits of the late Second Interglacial in association with hand axes of Middle Acheulean type and persists into Fourth Glacial (Würm) times. It is characterized by a new and improved method of producing flakes, which previously had been obtained in a more or less haphazard manner. This involves the careful shaping of the core by the removal of centrally directed flakes, and the preparation of an extremity for the detachment of a symmetrical oval flake. Since unstruck cores of this type exhibit a plano-convex section suggesting the form of a tortoise, they are known as tortoise cores. On the striking platforms of typical levallois flakes, small vertical flake scars, called facets, may be observed, and the scars of the converging core-preparation flakes are present

on the upper surface. The use of this technique resulted in the production not only of symmetrical flakes but also of larger ones in proportion to the size of the core. In the Middle and Upper Levalloisian a variation of this same basic technique was developed whereby it was possible to produce either triangular flakes (or points) or rectangular flakes (or flake blades) by modifying the method of core preparation.

Middle Paleolithic. The Middle Paleolithic comprises the Mousterian, a portion of the Levalloisian, and the Tayacian, all of which are complexes based on the production of flakes, although survivals of the old hand-ax tradition are manifest in many instances. These Middle Paleolithic assemblages first appear in deposits of the third interglacial and persist during the first major oscillation of the Fourth Glacial (Würm) stage. Associated with the Tayacian, in which the artifacts consist of very crude flakes, remains of modern man (*Homo sapiens*) have been found. Mousterian man, on the other hand, is of the Neanderthal race. By the 1960s no human remains had yet been found associated with the Levalloisian. It is in the Mousterian levels of the caves and rock shelters of central and southern France that the earliest evidence of the use of fire and the first definite burials have been discovered in western Europe. The cave of Le Moustier, near Les Eyzies in the classic Dordogne region of France, is the type site of the Mousterian. The typology of the artifacts is complex; it consists of three distinct increments: (1) the prepared striking-platform–tortoise-core (Levalloisian) tradition; (2) the plain striking-platform–discoidal-core technique of ultimate Clactonian tradition; and (3) a persistence of the bifacial core tool, or Acheulean tradition. The type artifacts from the Mousterian consist of points and side scrapers, in addition to a few hand axes (especially heart- or triangular-shaped forms), and the secondary working is coarse. A crude bone industry appears here for the first time. Judging by what is known concerning modern hunting groups, small bands or tribes of people already had developed simple social institutions, even at this early level of development.

Upper Paleolithic. The Upper Paleolithic, which occupies only approximately one-tenth of the time span of the period as a whole, first appears in horizons referable to the Würm I–II interstadial, and it persists to the very end of late Glacial times. Early man made his greatest cultural progress at this time. The hand axes and flake tools of the earlier assemblages were replaced by diversified and specialized tools made on blades struck from specially prepared cores. Many important inventions appeared, such as needles and thread, skin clothing, hafted stone and bone tools, the harpoon, the spear thrower, and special fishing equipment. Bone, ivory, and antler, in addition to flint, were extensively used. The earliest man-made dwellings are found, consisting of semisubterranean pit houses. Of prime importance and interest is the beginning of the basic techniques of drawing, modelling, sculpture, and painting, as well as the earliest manifestations of dancing, music, the use of masks, ceremonies, and the organization of society into patterns that were apparently fairly complex. Indeed, the location of certain settlements suggests a more complex social life, including perhaps collective hunting. There is evidence for fertility magic, private property, and possible social stratification. Furthermore, primitive types of early man disappeared, and the remains of men of modern type (*Homo sapiens*) alone are found in Upper Paleolithic sites.

Remains of *Homo sapiens*

Dmitri Kessel, *Life* (© 1955); Time Inc.



Views of an ivory female head (fragment of a "Venus" figurine), Aurignacian-Gravettian (*c.* 22,000 BC). From the Grotte du Pape, Brassempouy, Landes, France. In the Musée des Antiquités Nationales, Saint-Gérmain-en-Laye, France. Height 3.5 cm.

The chronology of this interval in western Europe shows a succession of cultures known as Lower Périgordian (or Châtelperronian; formerly Lower Aurignacian), Aurignacian, Upper Périgordian (or Gravettian; formerly Upper Aurignacian), Solutrean, and Magdalenian, each characterized by its distinctive types of artifacts. These latter occur, together with gravers (or burins), end scrapers, points, etc., which are common to all levels. The graver itself is a very important tool, for its invention made possible the extensive working of bone and facilitated the development of art. The climate of the Upper Paleolithic varied from cold steppe, or even Arctic tundra, to north temperate (taiga), similar to parts of Siberia and Canada of the present day.

Périgordian. In the Périgordian, named after a region in south central France, blades with steeply retouched backs are typical. The Lower Périgordian is characterized by large curved points with blunted backs that are known as Châtelperron points. These first appear, together with other types of blade tools, in horizons immediately overlying Upper Mousterian levels. It is believed that the straight points with blunted backs, called Gravette points and characteristic of the Upper Périgordian, were evolved from the Châtelperron type. In the final stage of the Upper Périgordian, tanged Font Robert points and diminutive multiangle gravers, known as the Noailles burin, are found. A number of small sculptured human torsos depicting the female form have been found at Upper Périgordian sites.

Aurignacian. The type site of the Aurignacian is near the village of Aurignac (Haute-Garonne) in southern France. At many sites it is found intervening between horizons referable to the Lower and the Upper Périgordian, a fact that is considered to indicate that more than one cultural element was present in western Europe at the beginning of Upper Paleolithic times. The tool types include various kinds of steep-ended scrapers, nose scrapers, blades with heavy marginal retouch, strangulated blades, busked gravers (or burins), and split-base bone points. Bone was extensively used, mainly for javelin points, chisels, perforators, and *bâtons de commandement,* or arrow straighteners. Articles of personal adornment, probably worn as necklaces, such as pierced teeth and shells, as well as decorated bits of bone and ivory, appear for the first time in the Aurignacian.

The oldest manifestations of art were produced during the Aurignacian, and the development continued during Upper Périgordian times. In general, Upper Paleolithic art falls into two closely related categories: mural art and portable art. The former includes finger tracings, paintings, engravings, bas-reliefs, and sculptures on the walls of caves and rock shelters; the latter is characterized by small engravings and sculptures on stone and bone found in the occupation layers. The whole development almost certainly owes its inspiration to the magico-religious idea, especially the custom of hunting magic as practiced today by living primitive peoples.

Solutrean. The Solutrean, which is named after the site of Solutré, near Mâcon (Saône-et-Loire), is noted for the beautifully made, symmetrical, bifacially flaked, laurel-leaf, and shouldered points, the finest examples of flint workmanship of the Paleolithic in western Europe. In addition, the usual types of gravers, end scrapers, points, perforators, etc., are present. Examples of Solutrean art are comparatively rare; they consist of sculpture in low relief and incised stone slabs. The fauna indicates that this culture flourished in a relatively cold climate.

Magdalenian. The rock shelter of La Madeleine, near Les Eyzies (Dordogne), is the type Magdalenian locality. This final culture of the Upper Paleolithic is noted for the



Stylized "Venus" figurines carved in ivory, Aurignacian-Gravettian (c. 24,800 BC). From Dolní Věstonice, Mikulov, Moravia, Czechoslovakia. In the Moravian Museum, Brno, Czechoslovakia. Height (left) 8.3 cm and (right) 8.6 cm.

By courtesy of the Czechoslovak News Agency, Prague

dominance of bone and antler tools over those of flint and stone and for the very remarkable works of art that were produced at this time. The wide variety of bone tools include javelin points, barbed bone points (or harpoons), eyed needles, *bâtons de commandement* (often elaborately decorated), perforators, spear throwers, chisels, etc. The flint and stone tools include a variety of special forms, among which small geometric forms, denticulated blades, scrapers with steeply retouched edges, and the parrot-beak graver are especially distinctive. The six phases of the Magdalenian have been established stratigraphically and are characterized mainly by the contained bone and antler implements. But the heights attained by the people responsible for this culture can best be evaluated on the basis of the art objects they produced. Magdalenian sites have yielded countless fine examples of both mural and portable art. Animals of the period, the usual subject matter, are portrayed in paintings (often polychrome), engravings, and sculptures. The fauna from the various Magdalenian horizons demonstrates that cold conditions prevailed in western Europe at the end of Paleolithic times. (H.L.Ms./Ed.)

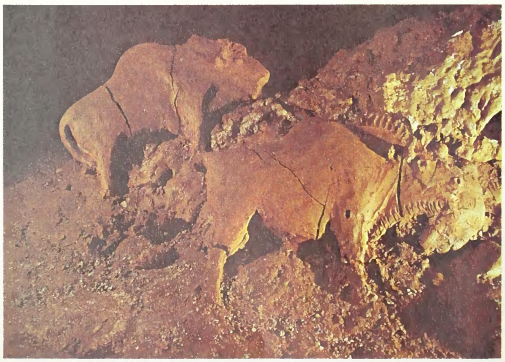
Mural and portable art

MESOLITHIC

In the Upper Paleolithic of Europe, certain evidence exists for what must have already been well-organized collective-hunting activities, such as the horse-stampede traces of Solutré, France, and the great concentrations of mammoth bones of the Gravettian hut settlements of Czechoslovakia and Russia. Cultural adaptations appear to have been made to restricted local areas or niches and to the fluctuations of climate and environment during the changing phases at the end of the Pleistocene range of time. In fact, it could be maintained generally that Upper Paleolithic traditions flowed rather smoothly into the Mesolithic, with no more significant indication of cultural development than further environmental readaptations. The people of the Mesolithic stage, or level of development, can be said to have "changed just enough so that they would not have to change."

The cultures. *The Maglemosian.* The level of intensified food-collecting cultures of the early Recent period in the Old World is best known from northwestern Europe, and it is with regard to this area that the term Mesolithic has greatest currency to denominate archaeological traces. A classic example of such traces comes from the Maglemose bog site of Denmark, although there are comparable materials ranging from England to the eastern Baltic lands. These bogs were probably more or less swampy lakes in Mesolithic times. At about 6000 BC, when the Maglemosian culture flourished, traces of primitive huts with bark-covered floors have been found. Flint axes for felling trees and adzes for working wood have appeared, as well as a variety of smaller flint tools, including a great number of microlithic scale. These were mounted as points or barbs in arrows and harpoons and were also used in other composite tools. There were adzes and chisels of antler or bone, besides needles and pins, fish-hooks, harpoons, and several-pronged fish spears. Some larger tools, of ground stone (*e.g.*, club heads) have appeared. Wooden implements also have survived because of the unusually favourable preservative qualities of the bogs; bows, arrow shafts, ax handles, paddles, and even a dugout canoe have been discovered. Fishnets were made of bark fibre. There is good evidence that the Maglemosian sites were only seasonally occupied. Deer were successfully hunted, and fish and waterfowl were taken, and it appears possible that several varieties of marsh plants were utilized. At Star Carr, in northern England, there are indications that four or five huts existed in the settlement, with a population of about 25 people.

This description of the Maglemosian must suffice to represent a considerable variety of European manifestations of the level of intensified post-Pleistocene food collecting. The catalogs of the Azilian and Tardenoisian industries of western Europe, of the Ahrensburgian of northern Ger-



J. Vertut—Ziolo

Two bison modelled in clay, Magdalenian (c. 15,000–c. 10,000 BC). In the Tuc d'Audoubert cave, Ariège, France. Length of bison at right 61 cm.



Head of an ibex, engraving on a bone staff, Magdalenian (*c.* 15,000–*c.* 10,000 BC). From Isturits, Basses-Pyrénées, France. In the Sainte-Périer Collection, Morigny, France. Height 13 cm.
J. Vertut—Ziolo

many, of the Asturian of Spain, etc., would each differ in detail, but all would point in the same general direction as regards cultural-historical interpretation.

The Nachikufan. As a further and far-distant example, the Nachikufan culture of southern Zimbabwe might be cited. Here again, microlithic flint bladelet tools, with certain types mounted as projectile points or in composite tools, existed. The Nachikufan cave walls show a few seminaturalistic drawings, and the caves also contain "pencils" of red and black pigment. Ground-stone axes and adzes, bored stones (digging-stick weights?), and normal-sized chopping and scraping tools of chipped stone also occurred. Grindstones of various types indicate a degree of dependence on collected vegetable foods, and the animal bones suggest specialization in the hunting of zebras, wildebeests, hartebeests, and wild pigs. These Nachikufan materials date back to at least 4500 BC. Again, an intensified level of food collecting is implied.

Cave drawings

The general picture. Though there are vast gaps in our knowledge of the Recent period in many parts of the Old World, enough is known to see the general cultural level of this range of time. Outside of the regions where food production was establishing itself, the period was one of a gradual settling-in and of an increasingly intensive utilization of all the resources of restricted regional niches. At first, the level seems nowhere to have achieved a climax of artistic expression, such as that for example, of Upper Périgordian–Magdalenian times. But, as time went on, certain climaxes within the matrix of an intensified level of food collection did occur. An often-cited example might be the complex art and social organization of the cultures of the northwest coast of British Columbia.

More often, however, as the culture history of the Recent period proceeded, cultures at the level of intensified food collecting were "captured" by being absorbed within an expanding matrix of the new elements, procedures, and traditions of food production or—subsequent to its appearance—by the expansion of civilized societies.

(R.J.Br./Ed.)

NEOLITHIC

The origins and history of European Neolithic culture are closely connected with the postglacial climate and forest development. The increasing temperature after the late Dryas period during the Pre-Boreal and the Boreal (*c.* 8000–5500 BC, determined by radiocarbon dating) caused a remarkable change in late glacial flora and fauna. Thus, the Mediterranean zone became the centre of the first cultural modifications leading from the last hunters and food gatherers to the earliest farmers. This was established by some important excavations in the mid-20th century in the Middle East, which unearthed the first stages of early agriculture and stock breeding (7th and 6th millennia BC) with wheat, barley, dogs, sheep, and goats. In a deep sounding in the Argissa Magula near Larissa (Thessaly, Greece), there have been early prepottery Neolithic finds (probably 6th millennium BC), while excavations in Lepenski Vir (Balkan Peninsula) have brought to light some sculptures of the same period. The independent origin of European Neolithic was established, and it was thought highly probable that the cradle of farming in the Middle East had not been the only one: there were others in Europe, too.

Ralph Morse, *Time* (© 1959) Time Inc



Cave painting of bulls, horses, and deer, Aurignacian (*c.* 28,000–*c.* 22,000 BC), Lascaux, Dordogne, France. Photograph covers a span of about 9 metres.



Woman gathering honey, water-colour copy by F. Benitez Mellado of the original painting in the Cueva de la Araña, near Bicorp, Valencia, Spain, Mesolithic (*c.* 10,000/8000–*c.* 3000 BC). In the Museo de Prehistoria de la Diputación Provincial, Valencia, Spain.
Instituto de Estudios Editoriales, Barcelona

The zones. Neolithic farming in Europe developed on its own lines in the four different ecological zones. These are: the Mediterranean zone of evergreen forest and winter rains; north of the Pyrenees, the Alps, and the Balkans, the temperate zone of deciduous forest and evenly distributed annual rainfall; still farther north the circumpolar taiga, or coniferous forest (the only zone to remain free of agriculture and stock breeding); and to the southeast the western end of the Eurasian Steppe. Each zone itself is subdivided into natural regions by physiographic boundaries and peculiarities of climate or soil. Only the three major divisions of the temperate zone are not obvious from every map. We may distinguish: western Europe, from the Atlantic to the Vosges and Alps and including the British Isles; the loesslands of central Europe, including the Ukraine and limited by the Balkans and the Harz; and the northern province, that portion of the Eurasiatic plain lying between the Rhine and the Vistula and including Denmark and southern Sweden. The substantial Neolithic communities that arose by 6000 BC must have been largely recruited from indigenous Mesolithic hunters and fishers, attested to so abundantly in western and northern Europe by various remains. (Some communities indeed seem to be composed entirely of such Mesolithic stocks, though they had adopted a Neolithic equipment from immigrant farmers; such are sometimes termed Secondary Neolithic. From these Mesolithic survivors, too, must be derived much of the science and equipment applied in Neolithic times to adapting societies to European environments. Upon the resultant distinctively European technology and economy was reared a no less original ideological superstructure expressed in distinctive sepulchral monuments, styles of ceramic decoration, and fashions in personal ornaments.

Cultural elements. *Rural economy.* In each of the above-mentioned provinces, the archaeological record begins with the early stages of farming, as in Thessaly. In the Mediterranean zone, this early farming is connected with the cardium pottery (decorated by shell impressions of *Cardium edule*), cultivation of the land having been proved by pollen-analytical methods in France, as elsewhere in temperate Europe, while northern Germany and southern Scandinavia revealed grain prints in potsherds (Ertebølle-Ellerbek). The process of cultural formation and modification during the Neolithic may be studied with the help of the different kinds of pottery and stone artifacts.

Mediterranean zone

Save in the taiga, where a Mesolithic economy persisted until the end of the Bronze Age, the basis of life everywhere was subsistence farming, supplemented by some measure of hunting and fishing—fish being a source of food curiously neglected in western and central Europe during the earlier phases of the Neolithic. Everywhere the same cereals were cultivated, together with beans, peas, and lentils. In the Mediterranean zone, orchard husbandry may already have begun, while around the Alps, apples were eventually cultivated and utilized for the preparation of a sort of cider. The balance between cultivation and stock breeding varied. Throughout the temperate zone, sheep, though bred even in Britain and Denmark, were at first rare. The damp temperate forests were uncongenial to these animals, and only toward the end of the Neolithic Period, when the greater dryness of the subboreal climatic phase and incipient clearing for plow cultivation were leaving their mark on the landscape, did flocks begin to multiply. On the loesslands, in early Neolithic times, animal husbandry may have played a subordinate role as compared with agriculture. But in the sequel, cattle raising combined with hunting proved to be the most productive pursuit among the deciduous forests with a Neolithic equipment; cultivation was relegated to an increasingly secondary place, until in the late Bronze Age more efficient tools for clearing land became generally available. The rural economy permitted the continuous occupation of permanent villages around the Aegean and in the Balkan Peninsula, perhaps also in southern Italy and the Iberian Peninsula. In the temperate zone, shifting cultivation may have been based on slash-and-burn clearance. Under this extravagant system, plots were presumably tilled with hoes, as in parts of Africa today. But by the beginning of the Bronze Age, the ox-drawn plow was beginning to replace the hoe.

Houses. Dwelling houses in Greece, Sicily, and the Iberian Peninsula were built, as in the Middle East, of pisé, or mud brick, on stone foundations. But in the Balkans and throughout the temperate zone, wood was used for the construction of gabled houses, stout posts serving to support the ridgepole and the walls of split saplings or wattle and daub. The earliest houses on the loessland of central Europe were very large, up to 42 metres (135 feet) in length and large enough to accommodate a whole lineage or small clan together with stalled cattle and grain stores. In the sequel these communal houses gave place to smaller two-roomed dwellings, 7.5 to 10 metres (24½ to 33 feet) long but still entered through one end. Finally in late Neolithic times clusters of one-roomed huts became the most widespread fashion. Around the Alps such two-roomed houses and, less often, one-roomed huts were raised on piles above the shores of lakes or on platforms laid on peat mosses. These are the world-famous Swiss "lake-dwellings" (*Uferrandsiedlungen*) that have yielded such precious collections of the organic substances from wood to bread that are otherwise missing from the archaeological record. In northern Europe, too, the earliest villages consisted of two parallel, long communal houses, but these were subdivided by cross walls into 20 or more apartments, each with a separate door. But here again the communal houses eventually broke up into free-standing one-roomed huts. Finally, Skara Brae on the treeless island of Orkney illustrates an ingenious adaptation of the one-roomed wooden hut to an inhospitable environment but shows how commodiously such huts must always have been furnished.

Stone tools. Carpenters used celts (ax or adz heads) edged by grinding and polishing of fine-grained rock or of flint where that material was available in large nodules. In Greece and the Balkans, all over central Europe and the Ukraine, and throughout the taiga, adzes were used exclusively, as in the earlier Baltic Mesolithic; in northern and western Europe axes were preferred. In the Iberian Peninsula axes and adzes occur in equal numbers in early Neolithic graves, but the proportion of axes in-

Use of axes and adzes





Megalithic gallery grave.
(Above) La Roche aux Fées, Essé, Ille-et-Vilaine, France, Neolithic (*c.* 3000–*c.* 1800 BC).
(Left) Interior of La Roche aux Fées.
D. Lesec—Ziolo

creased later. Often in western Europe, and occasionally in Greece and Cyprus, celts were mounted with the aid of antler sleeves inserted between the stone head and the wooden handle—a device that was already employed in the northern European Mesolithic. In Spain, the British Isles, and northern Europe axheads were simply stuck into or through straight wooden shafts, but adz heads must always have been mounted on a knee shaft (a crooked stick), a method regularly used for axheads, too, by the Bronze Age. Axheads like those in modern use, with a hole for the shaft, were rarely used for tools, but the Danubian peasants on the loesslands may sometimes have mounted adzes in this manner. They certainly knew how to perforate stone, using a tubular borer (a reed or bone with sand as an abrasive). From them the technique was adopted by various secondary Neolithic tribes in northern Europe for the manufacture of so-called battle-axes. The latter seem to derive their form from Mesolithic weapons of antler, but their splayed blades disclose the influence of metal forms.

Ax factories and flint mines. Celts, or axes, were manufactured in factories where specially suitable rock outcrops occurred, and they were traded over great distances. Products of the factories at Graig Lwyd, Penmaenmawr, North Wales, were transported to Wiltshire and Anglesey, those of Tievebulliagh on the Antrim coast to Limerick, Kent, Aberdeen, and the Hebrides. Similarly, large nodules of good flint were secured by mining in Poland, Denmark, The Netherlands, England, Belgium, France, Portugal, and Sicily.

The mine shafts, which were cut through solid chalk sometimes to a depth of six metres (20 feet) with the aid only of antler picks and bone shovels, may be simple pits, but often regular galleries branching from them follow the seams of big nodules. Although the ancient miners appreciated the necessity of leaving pillars to support the roof, skeletons of workers killed by falls have been discovered at Cissbury, Spiennes, and elsewhere. In the British Isles and Denmark, at least, there is evidence that the ax factories and flint mines were exploited and the products distributed by trade, for example, to the northern parts of Sweden. Still, the operators and distributors need nowhere be regarded as full-time specialists.

Art. Neolithic art, except among the hunter-fishers of the taiga, was geometric and not representational. It is best illustrated by the decoration of pottery. Pots, which were always handmade, were painted in southeastern Europe, southern Italy, and Sicily; elsewhere they were adorned with incised, impressed, or stamped patterns. Many designs are skeuomorphic—*i.e.,* they enhance the pot's similarity to vessels of basketry, skin, or other material. But on the loesslands of central Europe and the Ukraine and in the Balkans, spirals and meanders were favourite motifs.

Skeuomorphic designs

Trade. While Neolithic societies could be completely self-sufficient, growing their own food and making all essential equipment from local materials, luxury objects were transmitted quite long distances by some sort of trade. So ornaments made of the shells of the Mediterranean mussel, *Spondylus gaederopus,* are found all across

By courtesy of M.J. O'Kelly



Carved entrance stone to a Neolithic chamber tomb (*c.* 3200 BC) at Newgrange, County Meath, Ireland. Length 3.2 m.



Female figurine in the form of a jar, incised and with traces of paint, clay, Neolithic, *c.* 3000 BC. From Vidra, Bucharest, Romania. In the National Antiquities Museum, Bucharest, Romania. Height 42.5 cm.
Holle Bildarchiv, Baden-Baden

the Balkans, up the Danube Valley, and even on the Saale and the Main. Products of factories and flint mines were, as stated, traded widely throughout a single province, such as the British Isles, and some especially valued raw materials—the yellow flint of Grand-Pressigny (France), the obsidian of Melos and the Lipari Islands—became objects of "international trade" as much as shells. But the most prized object of such commerce was the amber of Jutland and Poland, whose electrical properties seemed evidence of potent mana. (Ri.P./Ed.)

Asia

PALEOLITHIC

During the Paleolithic, two major culture provinces can be recognized in Asia, each of which has yielded a distinctive sequence. The first of these includes the Middle East, Central Asia (formerly Russian Turkistan), central Siberia, and India; throughout this vast region a developmental sequence has been reported that, in all its essential respects, is related to that of Europe as well as to that of Africa in the early stages. The second of these provinces is in the south and east, and it embraces Pakistan, Myanmar (Burma), Java, Malaya, Thailand, and China. There the characteristic implement types consist of choppers and chopping tools that are often made on pebbles.

Hand-ax industries of Abbevilleo-Acheulean type are missing in southern and eastern Asia, together with the intimately associated prepared striking-platform–tortoise-core, or Levallois, technique. There the pebble-tool tradition persisted to the very end of Paleolithic times uninfluenced by contemporary innovations characteristic of the western portion of the continent.

Middle East. In this area, especially in Israel, Jordan, Lebanon, and Syria, a Lower Paleolithic development closely paralleling that of Europe is indicated by the widespread distribution of hand axes of Abbevillian and Acheulean type. Unfortunately, the majority of these finds are from open-air, unstratified sites that cannot be dated. A crude flake industry, reminiscent of the Tayacian of western Europe, has been reported from several cave sites. This is followed by a typical Upper Acheulean horizon in which there occur many developed hand axes of Micoquian type, a wide variety of flake implements, and the prepared striking-platform–tortoise-core technique. The Levalloiso-Mousterian found in the next-younger horizon is associated with a series of Neanderthaloid burials at one of the Mt. Carmel Caves of Israel and at Shanidar Cave in northern Iraq. Next in the sequence comes an early Upper Paleolithic development, which is characterized by various types of blade and flake-blade tools, including points that recall the Châtelperron type. This is overlain by the Antelian (formerly Middle Aurignacian), which in turn is followed by the Atlitian and the Kebarian. These assemblages, together with the recently discovered Baradostian of northern Iraq, constitute specialized late Upper Paleolithic industries that preceded various Mesolithic developments in the Middle East.

Neanderthaloid burials

Central Asia. In central Asia, few investigations of Paleolithic sites have been conducted. Surface finds of Acheulean-type hand axes have been reported from Turkmenistan, and several Mousterian localities have been excavated in southeastern Uzbekistan. At the most important of these sites—the cave of Teshik-Tash—the burial of a Neanderthal child who was surrounded by horns of a Siberian mountain goat has been discovered. No convincing evidence has been reported showing that this region was occupied during Upper Paleolithic times.

South Asia. Certain Paleolithic assemblages from India and Pakistan demonstrate that during Pleistocene times the region played an intermediate role between western Asia and the Far East. In the Punjab province of Pakistan, assemblages of implements that are characteristic of both the chopper–chopping-tool and the hand-ax–Levallois-flake complexes have been found. The former, which is called the Sohanian (or Sohan), has been reported from five successive horizons, each of which yields pebble tools that are associated with flake implements. Massive and crude in the earliest phases of the Sohanian, these implements reveal a progressive refinement in the younger horizons, where the evolved pebble tools are associated with flakes produced by the prepared striking-platform–tortoise-core technique.

In part contemporary with the Early Sohanian is a series of hand axes of Abbevilleo-Acheulean affinities, which occur in profusion at numerous sites in India from the Gujarāt region in the north to the Madras in the south. These sites yield hand axes, cleavers, and flake tools that are distinctly reminiscent of assemblages from southern and eastern Africa. As in the latter areas, the oldest materials are of Abbevillian type, and this is followed by the entire Acheulean cycle of development, just as in the case of the Stellenbosch of the Vaal Valley. Choppers and chopping tools made on pebbles and showing Sohanian affinities have been found throughout peninsular India in deposits of Middle Pleistocene Age. This suggests the probability that Lower Pleistocene horizons will ultimately be found in this area containing only pebble tools, as in the case of Africa.

No convincing evidence has been reported to indicate that a blade–burin complex was introduced into India before the close of Paleolithic times.

Far East. Pebble tools, including choppers and chopping tools, are found in the Pleistocene terrace deposits of the Irrawaddy Valley of upper Myanmar. This complex is known as the Anyathian. The Early Anyathian is characterized by single-edged core implements made on natural fragments of fossil wood and silicified tuff, and these are associated with crude flake implements. In the Late Anyathian, a direct development from the earlier stage, smaller and better made core and flake artifacts are found. No hand axes or flakes produced by the prepared striking-platform–tortoise-core technique have been found in Myanmar.

Elsewhere in the Far East, pebble tools have been reported from deposits apparently of Middle Pleistocene Age in western Thailand, for which the name Fingnoian has been proposed. In northern Malaya a large series of choppers and chopping tools made on quartzite pebbles and found in Middle Pleistocene tin-bearing gravels have been referred to collectively as the Tampanian, since they come from a place called Kota Tampan in Perak. Still another late Middle Pleistocene assemblage, called the Patjitanian, is known from a very prolific site in south central Java. In both the Tampanian and Patjitanian the main types of implements consist of single-edged choppers and chopping

The Patjitanian assemblage

tools that occur in association with primitive flakes with unprepared, high-angle striking platforms. Also in both assemblages is an interesting series of pointed, bifacial implements that have been described as crude hand axes. Since these tools are very rare in each instance and are absent in Myanmar, it is probable that they were developed in southeastern Asia independently of influences from the West. Several sites of Upper Pleistocene age in central Java have produced artifacts made on small to medium-sized flakes and flake blades. Antler and bone implements belong to this complex, known as the Ngandongian, which has also been reported from the Celebes and from the Philippines.

One of the oldest Lower Paleolithic occupation sites ever discovered is near the village of Chou-k'ou-tien, about 48 kilometres (30 miles) southwest of Peking in northern China. Associated with the remains of Peking man (*Homo erectus pekinensis*, formerly *Sinanthropus pekinensis*), pebble tools, together with quartz-flake implements, occur in quantity. This assemblage, which is known as the Choukoutienian, is of Middle Pleistocene age; it forms an integral part of the chopper-chopping tool tradition of the Far East.

The Choukoutienian assemblage

Also in northern China several Upper Paleolithic sites are known in the provinces of Shansi, Shensi, and northern Kansu, in the region encompassed by the great bend of the Yellow River (Huang Ho). Collectively known as the Ordosian, these materials are of Upper Pleistocene age. Typical of the Ordosian are blade implements of various types, points and scrapers of Mousterian-like appearance, and pebble tools of Choukoutienian tradition. Originally classified as Moustero-Aurignacian, it later became apparent that this development had much in common with that of the Yenisey–Baikal region to the north in central Siberia.

Siberia. The archaeological materials from the loess sites of Siberia between the Yenisey Valley and the Lake Baikal area is an interesting mixture of (1) blade tools, together with antler, bone, and ivory artifacts of classic Upper Paleolithic type; (2) points and scrapers made on flakes of Mousterian aspect; and (3) pebble tools representing a survival of the ancient chopper–chopping tool tradition of eastern Asia. Remains of semi-subterranean dwellings with centrally located hearths occur at certain of these stations, together with female statuettes in bone. One of the most striking features of this Siberian Upper Paleolithic is the fact of its comparatively late survival: in terms of the European sequence, it seems to have persisted as late as Early Mesolithic times. Indeed, in several instances it actually occurs in the uppermost layer of loess immediately below a horizon of humus containing Neolithic campsites. The problems of the Siberian Upper Paleolithic are of obvious importance to students of New World archaeology, since they have an intimate and direct bearing on the question of the peopling of the Americas. (H.L.Ms.)

MESOLITHIC–NEOLITHIC: THE RISE OF VILLAGE-FARMING COMMUNITIES

Middle East. There is little question that a level of an effective food-producing, village-farming-community way of life had been achieved in certain portions of southwestern Asia by at least 7000 BC. Furthermore, increasing evidence indicated that the effective village-farming level was preceded by one of cultivation and animal domestication and that this incipient level was at least under way by about 9000 BC.

Incipient cultivation and domestication. The level of incipient cultivation and domestication was essentially restricted to the piedmont and intermontane valley zone that flanks the Zagros–Taurus–Lebanon chain of highlands about the great basin of the upper Tigris–Euphrates and Karkheh–Karūn rivers and their tributaries. There are even hints that the zone extended to parts of the Iranian and Anatolian plateaus and that it may possibly have fingered northwest toward European Thrace. The significant point is that the zone appears to have formed a natural habitat for the cluster of plants and animals that were potentially domesticable. Most of these subsequent domesticates—wheat, barley, sheep, goats, cattle, and pigs, plus a possible wolf dog—still exist in their wild state in those parts of the zone that have been examined by prehistoric archaeologists and natural scientists.

The level of incipient cultivation and domestication is best manifested by the archaeological materials of the Natufian group in the Palestine-Syro-Lebanese littoral and parts of its hinterland, and by the Karim Shahir group in Iraqi and Iranian Kurdistan. The possibility of a continuation of the level into the northern Syrian and southern Turkish portions of the natural-habitat zone has been essentially untested by modern field research. Both of the available complexes of materials, the Natufian and the Karim Shahirian, appear to have been established by about 9000 BC.

The Natufian and Karim Shahirian. In both, there are clear indications of open settlements that were of modest size, and there are some traces of round huts, some of which were built on stone foundations, although caves are also known to have still been inhabited. Both groups yield traces of normal developments of flint industries that are based essentially upon local Upper Paleolithic antecedents, and both must have been influenced in their food getting by the already intensified food-collecting practices of their immediate predecessors. It is freely admitted that the postulation of this incipient level rests considerably on a judgment that is based on the materials of the succeeding level of effective village-farming communities. Nevertheless, it has been demonstrated that sheep were already being used at the incipient level, and there are such hints as flint sickles, ground-stone mullers, mortars and pestles, and probable hoe blades to suggest that food plants were also receiving marked attention. Claims for the domesticated dog in the Natufian are not universally accepted, however.

Evidence of settlements

It has been rightly stressed that the materials of this level will be exceedingly difficult to interpret, since the earliest plant and animal domesticates will show little morphological difference from their wild contemporaries and since the procedures and artifacts of the new food-getting and food-preparation techniques will have taken considerable time to develop.

The effective village-farming community. The next level, that of the effective village-farming community, yields, even in its earliest available phase (*e.g.*, at Jarmo, in Iraqi Kurdistan, *c.* 7000 BC), materials that leave little doubt about the presence of food production. In the Jarmo phase, wheat, barley, a pea, goats, sheep, and—before the phase is completed—pigs and probably dogs all appear. The Jarmo settlement suggests a permanent village of about 20 rectangular several-roomed huts, which probably had a population of at least 150 people. Several other variants of the Jarmo phase have been excavated or at least located in Kurdistan. One of these, Sarab, near Bakhtarān (formerly Kermanshah) in Iran, suggests a seasonal encampment of herdsmen. Sarab yields pottery throughout its shallow deposit; at Jarmo itself, similar pottery appeared only in the upper third of a much thicker deposit.

"Preceramic" village sites have been recovered in the Dead Sea Valley, along the Syro-Palestinian littoral, on Cyprus, in the southwestern Turkish highlands, and even in Thessalian Greece. Controversy exists regarding the very spectacular architectural remains of the Dead Sea Valley site of Tall as-Sultān (reputedly also the site of the later Jericho), with disagreements about its "town" or even "urban" nature in view of the normal small-object assemblage there, the radiocarbon determinations now available for it, and the relative lack of firm evidence for cultivation. These disagreements will certainly be resolved as more sites in the time range of about 9000 to 6000 BC are excavated in the Syro-Palestinian littoral and in parts of its hinterland.

Fully established village sequences in the Middle East. By 6000 BC or not long thereafter, a variety of more or less complete regional cultural sequences developed in the Middle East. In Iran, two sequences appeared. That beginning at the site of Sialk developed most characteristically in the northern and northeastern parts of the country and evidently extended into what is now Turkmenistan and northern Baluchistan and possibly beyond to

the Indus. A somewhat different tradition developed in southwestern and southern Iran, early traces of which may be seen at Jaʾfarabad in Susiana (Elam) and at Bakun B near Persepolis. This tradition exhibited a closer proximity to the earlier sites in Iraq; its eastern extension may also be traced as far as Baluchistan, if not beyond into the Indus Valley.

The Hassunan assemblage

The earliest full-bodied assemblage in northern Iraq, following that of Jarmo, is the Hassunan of the Mosul–Kirkuk piedmont. Next—either as elements in the developed Hassunan phases or alone at the mid-Euphrates site of Baghouz or at the mid-Tigris site of Samarra—comes the Samarran phase. Then, with further overlap, comes the Halafian phase of the upper (Syro-Turkish-Iraqi) piedmont. The overlapping of these three assemblages is indicated by the availability of a radiocarbon determination for an early Halafian level, which is as early as either of the two determinations of the Hassunan—about 5750 BC. The beginning of the food-producing sequence in classic southern Mesopotamia comes after this time and is, perhaps, partly an amalgam of (1) a southward extension of Hassuna–Samarra–Halaf traits; (2) the westward extension of early Susiana traits from southwestern Iran; and (3) the probable presence of indigenous riverine-oriented food collectors.

Another local tradition, at least contemporary with that of Hasuna (and perhaps earlier than that of Sialk), appears to have its focus in the Syro-Cilician corner of the eastern Mediterranean; its preceramic antecedents may be seen in the basal levels of coastal Ras Shamra. Later, this Syro-Cilician tradition appears to have been affected by the Halafian and later inland developments. To the north of Syro-Cilicia the early materials of Hacilar and of Çatal Hüyük must be given place, including the possibility of their implications for the early developments in the Aegean. To the south, the Syro-Cilician tradition merged gradually into a somewhat related coastal Palestinian tradition. But in the more arid reaches of inland Palestine, a somewhat different tradtion developed that appears to have culminated in the sites of semi-nomadic traders, such as that at Beersheba.

Food production appears to have reached Egypt (and northern Africa generally) relatively late, perhaps not much before 4500 BC. Such northern Egyptian occurrences as Merimde (on the western flank of the Nile Delta) and the Fayum (Fayyūm) A pit sites might argue for an expansion directly (by boats?) from the Asian coast. But some authorities favour the idea of a way into middle Egypt via the Red Sea and the Wādi Rawḍ ʿĀid to account for the available developments there.

General cultural level of the early villages. This very compressed sketch is meant only to suggest the variety of regional variations and adjustments within the general development of the effective village-farming level in the Middle East, from about 6000 to 4500 BC. Wheat and barley were the staple crops; cattle join sheep, goats, and pigs as major food animals, at least by the Halafian phase. Villages—except the Tall as-Sultān fortified establishment—were small; an informed guess would put their limit of population at about 500 people. Again, except for some dubious interpretations of certain rather modest buildings as "shrines," the architecture appears to be entirely domestic in nature. Aesthetic expression also took the form of an almost bewildering variety of regionalized and successive painted-pottery styles. The modelling of clay figurines—already well attested in the phase of Jarmo and its contemporaries—continues, with both animals and stylized human females being rendered. The latter, especially, may be suspected as having represented some magico-religious aspect of concern with fertility, upon which the livelihood of the communities depended. Flint tools were gradually replaced by copper and, eventually, by bronze implements, and the early trade routes in obsidian (a volcanic glass of restricted occurrence) were doubtless taken over by the metallurgists. Certain artifacts indicate the presence of weaving; in addition to their local utility, woven fabrics may also have served as media of exchange. It would be difficult to maintain that there was a strict subdivision of labour on a full-time scale (except perhaps on a basis of sex or age), but such a trend must have been setting in.

Evidence of woven fabrics

It should be emphasized that the complexity of this picture cannot readily be conceived apart from a system of effective food production. It may also be noted that an older trend was not being reversed. The intensified food collecting at the close of the Pleistocene was apparently accompanied by increasing regional specialization and a tendency toward full utilization of a rather restricted environmental niche. Now—with the establishment and spread of the effective village-farming community, its expansion beyond the confines of the natural-habitat zone, and the beginnings of trade—the horizon began to widen again. The *oikoumenē*, or known world of these first effective village farmers, became an ever-expanding one. Hence, just as it is probably not very fruitful to ask exactly where any particular element was "invented" or first discovered within the level of incipient cultivation and domestication in the natural-habitat zone, it is probably most useful to view the development of the way of life of the effective village-farming community as a general regional phenomenon of cultural interrelationships and stimulations. It might be further suggested that this general development took place over a broad area that had certain localized environmental variables and natural resources. These environmental conditions, however, had been there, just as the natural-habitat zone itself had been, long before incipient and effective food production came into being. The latter were human, cultural achievements; favourable environment, though it enabled them to come into being, did not cause them.

The threshold of town and city life in the Middle East. The end of prehistory and the threshold of urban civilization are first seen in classic southern Mesopotamia about 4500 BC. The materials of the Ubaidian assemblage make their appearance after a still rather poorly delineated phase in the basal levels of the mound of Eridu. Whatever elements combined in the earliest amalgam (northern Iraqian, Susianan, or indigenous), the resultant traits of the Ubaidian tradition are revealed in their greatest clarity, consistency, and variety in southern Mesopotamia by 4000 BC.

There are mound accumulations and at least one large cemetery, which suggest a scale of communities well beyond that of the simple village. Buildings sufficiently large, formal in design and size, and monumental in concept and decoration to be judged as temples were present. Great quantities of painted pottery of high quality appear in the excavations. This pottery, by its very uniformity and the somewhat cursive nature of its decoration, may already have been the product of specialized craftsmen. No unquestionable instances of metal tools were available by the early 1960s from Ubaidian contexts in southern Mesopotamia (although metal was available by that time in the north), but quantities of very highly fired clay tools (axes, adzes, sickles) had been found. These were useful for cutting the pithy woods, reeds, and grain of the southern alluvial environment or for dressing sun-baked bricks. The female clay figurines continued, but in a unique and highly characteristic stylization.

General cultural level of the Ubaidian Phase. A Ubaidian town supplied itself from fields of wheat and barley and its animal herds. The agricultural regime in the hot, dry alluvium of southern Mesopotamia depends, however, upon the utilization of the braided lower channels of the Tigris and especially of the Euphrates. Though elaborate irrigation works did not exist, the management of even quite informal ditches, with necessary shifts when the natural channels of the rivers shifted, added a new dimension to the sociopolitical necessities of Ubaidian culture. This system of irrigation may have been one of the factors that contributed to the expansion of society in late prehistoric Mesopotamia. Given the proper management and water, the yield of the rich alluvial soil was magnificent (until salinity became a problem several centuries later). There were also important dietary additions, such as dates from the groves of date palms and fish from the river channels and ditches.

With southern Mesopotamia as its focus, the Ubaidian tradition "exported" some of its elements at least as far as the Mediterranean coast and throughout the great up-

per drainage basin of the Tigris–Euphrates and Karkheh–Kārūn rivers. These exported traits doubtless reflect the growth of another *oikoumenē*, and one much more explicitly southern Mesopotamian in character. In southern Mesopotamia itself, the Ubaidian phase was followed (after a "Warkan" interval) by the proto-Literate period, in which the usual criteria of civilization are manifest.

The Indus Valley

South and East Asia. It is known that village-farming communities existed in the Indus Valley as early as 3000 BC, if not earlier. The original complexion of their assemblages resembled those of Iran (and perhaps those of the Ubaidian imprint on southwestern Iran), but this complexion gradually changed to something characteristic of the Indus Valley itself and evidently culminated in the Harappan urban civilization. Some degree of contact between the cities of the Indus and of Mesopotamia certainly continued to exist, however. It is becoming evident that the Harappan complex was not restricted to the Indus Valley alluvium but extended into the adjacent semitropical portions of India as well.

Knowledge of the developmental sequence in China is obviously incomplete. Except for a few snatches of typologically more simple materials, the first evidence of food production in China appears to pertain to a well-advanced phase of the effective village-farming-community level. This is the Yangshao complex, focussed in the basin about the confluence of the Yellow River (Huang Ho), the Fen Ho, and the Kuei Shui. Characterized by a handsome painted-pottery style, the Yangshao catalog also includes cultivated millet, rice, kaoling, and possibly soybeans, as well as domesticated pig, cattle, sheep, dog, chicken, and possibly the horse and silkworm. The village houses were built of tamped earth; there was a flourish of "ceremonial" pottery vessels and of elaborately worked objects in jade, as well as flint, bone, and ground-stone objects of daily use. The Yangshao phase is followed by that called Lungshan, after which comes the Yin, or Shang, early dynastic complex of about 1500 BC. The date for the beginning of the Yangshao is unknown; 3500 BC is probably much too early.

Even less is known of southern China and southeastern Asia; the former seems to have been affected by the expansion of the makers of the Lungshan black pottery and perhaps was also stimulated from the south. The rather amorphous Hoabinhian and Bacsonian sequence in Indochina, with ground-stone axes and adzes, appears to be quite late—perhaps of the 1st millennium BC. In Japan, on the other hand, the first appearance of pottery of early Jōmon type (evidently all of the Jōmon development lies before effective cultivation had begun) has several radiocarbon determinations at about 7000 BC, but some authorities suspect contamination of the samples. Positive cultivation (wet rice) appears in Japan about 300 BC, in the Yayoi phase. (R.J.Br./Ed.)

Central Asia and Siberia. The Mesolithic–Neolithic era and the settlement of northern Siberia started in the 7th to 6th millennia BC—the period of climatic optimum in Postglacial times, when forest conditions were introduced. Stratified sites in the Lake Baikal area show a long and gradual transition from the Paleolithic to the Neolithic stage. The Postglacial culture in Siberia was not a true "Neolithic" food-producing, but a "Mesolithic," or "sub-Neolithic," hunter-and-fisher, culture (except in southern Siberia around the Aral Sea) with a microlithic flint industry in western and southern Siberia and with polished-stone tools, pointed- or round-based pottery, and bow and arrow, starting in about the 4th millenium BC in almost all parts of Siberia.

Culturally and racially the territories of this vast area are divisible into two blocks: (1) the southwestern, covering the area from the Caspian Sea to the upper Yenisey, extending over the zones of semidesert, steppe, and forest steppe and (2) the eastern and northern, covering mountainous regions from Lake Baikal to the Pacific Ocean and the taiga (coniferous forest) and tundra belts of northern Siberia. The first is represented by Europoid people, the second by Mongoloid. These two blocks were in conflict until the Mongoloid overflowed the Europoid in several waves.

Europoid block. The earliest Neolithic culture in the steppes and in the oases may reach the 4th millennium or earlier, but its beginnings are not as yet satisfactorily investigated. The small flint industry continued from the earlier Mesolithic times. In the 3rd millennium BC, copper, painted ware, and other elements from the south entered the area. Sheep, cattle, and horses were the chief domesticated animals. Copper knives and stone sledges for mining appeared. Pottery was mostly round-bottomed, decorated with geometric stamped or scratched patterns in rows. Typical burial of the dead was in a contracted position under an earth mound. Excavations in Khwārezm (Khorezm, Khiva) revealed large communal houses of oval form. In the region of the Aral Sea (Khwārezm) this culture was given the name Kelteminar, in Altai and the region of Bisk, Krasnodar, and Minusinsk, Afanasievo, although related cultural features are found between southern Russia and the upper Yenisey, the area presumed to be Indo-European homeland. The Afanasievo was replaced by the Okun group of stockbreeders mixed with Mongoloid elements, famous for stone stelae incised with mythical figures, elsewhere in southern Siberia.

Burial practice and dwellings

Continuous cultural development is seen in the 2nd millennium BC. This culture, named Andronovo, is relatively uniform in this wide area, in spite of some local variations. Agriculture now played an important role. People lived in earth huts and reared cattle, sheep, and horses. Bowl- and flowerpot-shaped vessels were flatbottomed, well smoothed, decorated with geometric patterns, triangles, rhombs, and meanders, pointing to relationship with the painted pottery of the southern regions. Burial in contracted position persisted. The typical elements of a religion of food producers, the fire and sun cult, as well as bread offering, are evidenced. Wooden constructions in rich graves may have designated social differentiation. The Andronovo complex is intimately related to the Timber-Grave (Russian Srubna) group in southern Russia: both represent branches of the Indo-Iranian cultural block.

In the second half of the 2nd millennium BC in the region of Minusinsk, a Sinid group broke in that brought with it a bronze inventory of Ordos (northern China) type. Cemeteries of single graves covering the dead in extended position in stone cists, equipped with round-bottomed pots, appeared. New people mixed with the local Andronovo population. Through this immigration the so-called Karasuk culture originated and spread its influences farther to western Siberia and Russian Turkistan. Trade relations extended to central Russia. Exchange with the centres of the Far Eastern metallurgy introduced a new character of material culture (daggers and knives terminating in animal sculptures, series of ornaments) and stimulated the flourishing of metal industry in a wide area. The regions west of Minusinsk—Altai, Kazakhstan, and Kirghizia—show variations of Karasuk culture with strong local elements with which the persistence of the ancient racial type corresponds. Chronology of this period is based on comparisons with northern Chinese bronzes.

The Karasuk period persisted down to *c.* 700 BC. From *c.* 700 to *c.* 200 BC, culture developed along similar lines. Vital trade contact is traced from northern China and the Baikal region to the Black Sea and the Urals, influencing the uniformity of the culture. A mounted-warrior element occurred, although the agricultural and cattle-breeding elements persisted. In the high Altai, Tien Shan, and Pamirs appeared graves of nomadic warriors with co-burial of horses. Regarding the local facies, or separate political confederations, cultures of this period are called Tagar in the region of Minusinsk, Maicmiric in Altai, Sauromatian (Sarmatian) in western Kazakhstan, Sakian in Tien Shan and Pamirs, and Massagetian in Khwārezm.

The art of the steppe zone from southern and eastern Russia to China developed into specific animal style. The decorative talent is illustrated in the great ingenuity that the artist displayed in filling up with animal figures a shape determined by practical ends. The elk, ram, bird, and cat-animal portrayals of the middle of the 1st millennium BC exhibit a conjunction of the highest verisimilitude with rigorous stylization; later the organic form of the animal was ruled by extreme stylization. The elements of natural-

Animal-style decorative arts

ism link this style with the naturalistic animal style of the northern Eurasian forest belt. New motifs in the steppe and forest-steppe belt—portrayal of groups of animals, antithetic and intertwined groups of bodies, curled up animals, beasts, and birds of prey—originated in a borrowing of ideas from the Middle East and China.

Pre-Christian culture, although influenced by the Persian Empire, progressed gradually until the new flow from the east started. The territory between the lower Volga and Altai represents a unit with a common destiny. Chinese and Western sources report that the Sarmatian–Sakian time was followed by the supremacy of the Huns, who dominated the western steppes as far as the Urals and the Volga. Archaeological investigations show that the east–west movement started at a time when the Hun confederation had not yet been consolidated. In eastern Kazakhstan appeared an eastern group of Stone Tombs people not later than the 5th century BC. The main east–west stream ran presumably from Manchuria–upper Lena, along the northern border of the Gobi, into the Lake Balkhash territory, and from there on, avoiding powerful cities in Khwārezm, into the steppes north of the Caspian. For centuries up to the consolidation of the Turkish khanate in the 6th century AD, Mongoloid components were mingling with the local Europoid, which have never been wiped out. The known pre-Turkic tribes—Massagetians, Sakians, Usuns, Khakas—all show more or less Europoid traits.

The cultural pattern from Altai to Transbaikalia in the last centuries BC and first centuries AD is largely traced to China of the Han period. Social differentiation is evidenced by princely burials, extraordinarily well preserved in five large burial mounds of Pazyryk and Shibe in the high Altai. Complete burial places were frozen, and even perishable substances were preserved, including human bodies and horses with harness and saddles, textiles, felt and leather objects, clothing, fur coats, false beards, besides jewelry, mirrors, hair plaits, etc. All materials were finished with virtuosity. The art combined animal, plant, geometric, and human designs. Polychromy played an important part. Mummification, tattooing, scalping, and the use of amulets are evidenced.

Meanwhile, in the region of the Aral Sea, the apogee of the Khwārezm civilization was reached in the epoch of the empire of the Kushans. During the 1st and 2nd centuries AD the irrigation system attained its greatest development. Numerous cities were built along the banks of the canals.

Mongoloid block. The Arctic and sub-Arctic zones exhibit a continuous culture belt in a sub-Neolithic stage from Boreal times through several millennia. Making of pottery and polishing of stones, but neither farming nor domestication of animals, except the dog, were known. People lived in small, semi-nomadic communities, in semi-subterranean houses. The Arctic seashores demonstrate sea-hunter cultures. In the north this stage of life has lasted down to the present time. The region of the Amur River in eastern Siberia shows a long-lasting Neolithic, of which the oldest forms resemble certain finds of northern Japan (Proto-Ainu) and China. Cultural continuity is traced from the Neolithic through the stages in which copper smelting and iron were known. In the farthest northeast, archaeological and other data suggest that the Kamchadal, Koryak, and Chukchi entered the area from the west less than 2,000 years ago and found the coastal region occupied by a population related to the Eskimo.

The Ural region was linked with the northern Russian and western Siberian culture on one hand and with the Aral Sea region on the other. Throughout the Neolithic and Bronze Age times, two cultural branches were evident: the middle Ural (or Shigir) and that of the River Ob Basin. During the 3rd and 2nd millennia BC the culture of the middle Ural region is famous for its elk and waterbird sculptures portrayed in wood, found in the peat bogs of Gorbunovo and Shigir, and that of the upper Ob region for its cemeteries in the area of Tomsk, abundant art objects, including bear figurines, and rock carvings. Cultural relationships between the northern Baltic and northwestern Siberia, forming a continuum up to the early historic period, furnish this area with the characteristics of the homelands of the Finno-Ugric-speaking peoples.

The best explored regions are the shores of Lake Baikal, the Angara Valley, the upper Lena, and the lower Selenga. The earliest Neolithic culture shows Siberian Upper Paleolithic traits; the flint tradition of small implements persisted alongside a wood-working and quartzite industry, which developed as a result of adaptation to a taiga environment. Chronological phases are based chiefly on the Angara grave materials by means of stratigraphy and comparisons. The following successive cultures are discerned: (1) Isakovo, showing the earliest appearance of pottery, alongside flint and bone tools (arrowheads, knives, points, half-ground adzes). Pointed-based pots in Isakovo probably were copies of similarly shaped baskets. Art monuments are not numerous here. The period may reach back to *c.* 4000 BC. (2) Serovo, characterized by thinner pottery, decorated by dentate stamping, boss, pit, and net impressions and by stone inventory of more regular forms; reinforced bows with bone backing and fish effigies of stone appear. A marked increase of population is indicated by settlements covering hundreds of square metres, including storage pits for fish. In cemeteries, women's graves were richly equipped, which may indicate woman's equal rights in Serovo community. Serovo people migrated to the steppe and deserts of Central Asia and Inner Mongolia. The period belongs to the 3rd millennium BC. (3) Kitoi, placed before the middle of the 2nd millennium BC, shows a variety of more developed forms of equipment; the great number of fishhooks found in the graves indicates that subsistence was now based primarily on fishing instead of hunting; sculptures of human faces in stone, stone rings, and nephrite and copper objects appear; close parallels in stone and bone industry, as well as in art style, are found from northern Scandinavia and northern Russia to China. (4) Glazkovo, extending through the middle of the 2nd millennium BC to *c.* 1300 BC, continues a similar mode of life; novelties include the appearance of burial mounds and burials in stone cists, copper knives and arm rings, nephrite rings and disks.

Chronological phases

The first bronze inventory in the region of Lake Baikal is related to the bronzes of the Shang period in northern China and the earliest Ordos bronzes. Life was then of semi-settled character, and cattle breeding was known. Continuity of culture in the Bronze Age stage is traced up to *c.* 300 BC. The period between *c.* 700 and 300 BC in Transbaikalia, called Stone Tombs I, exhibits a transition to nomadism and mounted-warrior conditions. Cultural elements held in common with the Scythian steppe zone appear as far in the northeast as the Lena River. South–north and north–south movements are attested in the last centuries BC. The south–north movement is assumed as Yakut migration from the Baikal to the upper Lena region, the north–south movement from Cisbaikalia to Transbaikalia as migration of the taiga group, related to the Tungus of the present day. (M.G.)

Africa

PALEOLITHIC

The Paleolithic of Africa is characterized by a variety of stone-tool assemblages, some of which represent purely local developments while others are practically identical with materials from corresponding horizons in Europe. Geological investigations of the Late Cenozoic deposits of this continent indicate that, as the result of fluctuations in rainfall, the Pleistocene Epoch throughout most of Africa can be subdivided on the basis of a succession of pluvial and interpluvial stages. The pluvials, known as Kageran, Kamasian, Kanjeran, and Gamblian, are believed to represent the tropical and subtropical equivalents of the four major glacial stages of the Northern Hemisphere, but this has not yet been proved. The archaeological succession is well established in certain areas, although not in the continent as a whole.

Tool sequences

North Africa. In this area, very crudely worked pebble tools have been reported from one site in Algeria in direct association with a Lower Pleistocene (Villafranchian) mammalian assemblage. Throughout Tunisia, Algeria, Morocco, and the Sahara region, Lower Paleolithic hand axes of both Abbevillian and Acheulean type, to-

gether with flake tools, have been found in great numbers. The geological evidence shows that the Sahara region was far less arid during Pleistocene times than it is at present. The Middle Paleolithic of both Levalloisian and Mousterian facies is very widespread in North Africa, and it apparently persisted as late as the second maximum of the Würm glaciation in terms of the European sequence. A specialized Middle Paleolithic development, known as the Aterian, occurred there; it is characterized by tanged points made on flakes and flake blades. This was succeeded by two distinctive blade-tool complexes—the Capsian and Oranian—which are more or less contemporary. Their main development took place during the time span of the European Mesolithic. The Capsian sites are all inland, whereas the Oranian has a coastal distribution. Both are microlithic tool complexes that persisted after the introduction of Neolithic traits into the area.

Egypt. The Pleistocene terrace gravels of the Nile Valley in Egypt have produced a wealth of Paleolithic materials. The 30-metre terrace contains typical Abbevillian and early Acheulean hand axes, including a special form with a triangular section known as the Chalossian type. These are associated with primitive flake implements. In the 15-metre terrace, developed Acheulean has been recorded, while the nine-metre terrace yields large flakes and cores of Levalloisian type. In the low terrace, which occurs at a height of three metres above river level, developed Levalloisian (originally called Mousterian) has been reported. Overlying the low terrace, a local development known as the Sebilian is found. It contains very highly evolved flake implements of Levallois type and, in its later phases, a definite microlithic industry. Of approximately the same age as the Sebilian are several Epi-Levalloisian sites in the Lower Nile drainage, including the Fayyūm Depression and the al-Khārijah (Kharga) Oasis. In the latter area, where the specialized Levalloisian development is called the Khargan, an Egyptian version of the Aterian has been discovered.

East Africa. In Kenya, Tanzania, and Uganda, very simple types of pebble tools, roughly chipped to an edge on one side only, occur in deposits of Lower Pleistocene age. This development, known as the Kafuan, apparently evolved into an industry characterized by implements made on pebbles chipped to an edge on both sides, called the Oldowan. Overlying the latter are beds containing true Lower Paleolithic hand axes of Abbevillian and Acheulean type, together with flake tools. Associated with the Middle and Late Acheulean are cleavers made on flakes, as well as evidence of the use of the prepared striking-platform–tortoise-core (Levallois) technique in the production of flakes. In the next-younger horizon, two distinct toolmaking traditions are found: the Kenya Stillbay, a Levalloisian derivative characterized by small- to medium-sized, bifacially flaked points or minute hand axes; and the Kenya Fauresmith, basically of Acheulean inspiration and very similar to the true Fauresmith of southern Africa. Carefully shaped round stone balls, believed to have been used as bola weights in hunting, constitute part of the Fauresmith assemblage. In the post-Gamblian dry phase, microlithic tools appear for the first time in an assemblage known as the Magosian. This was followed by the introduction into the area of a true blade technique, called the Kenya Capsian, together with the art of pottery making. More or less contemporary with the localities where the earliest pottery is found in East Africa, a series of sites has been discovered yielding typical microlithic assemblages and referable to the Kenya Wilton, also found in South Africa, Zimbabwe, and Zambia.

Southern Africa. The sequence in southern Africa is well established on the basis of the terrace stratigraphy of the Vaal Valley. Just as in North and East Africa, the succession begins in the basal Pleistocene with the occurrence of simple pebble tools of Kafuan type. These develop into what is called the pre-Stellenbosch, which is found in the oldest gravels of the Vaal and which includes artifacts made on pebbles that recall both the Kafuan and the Oldowan. The true Stellenbosch complex occurs in the next-younger series of deposits; it is simply a southern African version of the Abbevillian and Acheulean of other parts of Africa and Europe. Typical are hand axes, cleavers, flakes struck from Victoria West cores, and (in its later phases) various sorts of flakes produced by the prepared striking-platform–tortoise-core technique. The Stellenbosch was followed by the Fauresmith, which is characterized by evolved hand axes and Levallois-type flakes. The Stellenbosch and Fauresmith together constitute what is called the South African Older Stone Age, a period roughly corresponding to the Lower and Middle Paleolithic stages of Europe. On the other hand, the South African Middle Stone Age belongs to the later part of the Upper Pleistocene. It is characterized by a series of more or less contemporary flake-tool assemblages, each of which displays local features. These are known as Mossel Bay, Pietersburg, Howieson's Poort, Bambata Cave, Stillbay, etc.; Stillbay, which occurs in Kenya and Uganda, is the only one of these found outside southern Africa. The characteristic tools are made on flakes produced by a developed Levalloisian technique, including slender unifacial and bifacial lances or spear points for stabbing or throwing. In the final stages of the Middle Stone Age, known as the South African Magosian, microlithic elements appear, just as in the case of East Africa. The Later Stone Age cultures of this region—the Smithfield and the Wilton—developed during post-Pleistocene times. These are closely related and, in their later stages, reveal varying degrees of influence as the result of contact with the culture introduced by the Bantu-speaking peoples. Both were extant at the time the first Europeans arrived in southern Africa, and there is little doubt that the Wilton, which is a typical microlithic assemblage, is to be associated with the modern San (Bushman). There are many paintings in the rock shelters and engravings on stones in the open-air sites of southern Africa, the oldest of which belong to the Later Stone Age. The naturalistic style of art revealed at these sites persisted until well into historic times.

Central Africa. The Lower Paleolithic sequence of Central, or Equatorial, Africa is essentially a repetition of what has already been outlined for East and southern Africa. At the beginning of Middle Stone Age times, however, a special development took place known as the Sangoan (formerly Tumbian). This is characterized by picks and adzes made on bifacially flaked cores, the tranchet type of ax, hand axes of developed Acheulean form, massive side scrapers, and many elongated, bifacially flaked points that probably served as lances or spearheads. The Sangoan seems to represent a response to the environmental conditions of this tropical rain-forest region. Its main development took place during Upper Pleistocene times, but it persisted after the introduction of Neolithic traits into the area. (H.L.Ms.)

Vaal Valley stratigraphy

MESOLITHIC–NEOLITHIC

The Paleolithic was everywhere followed by the Mesolithic, a period when man continued to use stone tools, mostly microlithic, and, while still in the hunting-and-gathering stage, depended less for his food supply on large mammals than on fish and mollusks. In Africa the evidence for the Mesolithic is still scanty. In the Lower Nile Valley, sites have been examined only at Ḥulwān (Helwan) and Kawm Umbū (Kom Ombo). At the latitude of Khartoum, for a considerable distance to each side of the Nile, have been found sites of a Mesolithic culture in which large, well-fired, unburnished pots decorated with designs impressed with a fish spine to make them resemble baskets were made and barbed bone harpoons were used for fishing. Arrows were mostly armed with stone lunates, and in general the microlithic industry shows relations with the Capsian (of northwestern Africa) and the Wilton (of east central Africa). The fauna indicates a climate much wetter than the present. The upper Kenya Capsian, with traces of similar pottery found at Gamble's Cave, probably represents the Mesolithic of Kenya. Its pottery also copies basketwork. And while it is impossible to say where pottery was invented, the discovery of a prepottery Neolithic in Asia, with the existence of modern mud-lined baskets among the Nilotes, the accidental burning of which could have led to the invention of pottery, suggests that pottery was possibly an African discovery.

Lower Nile Valley sites

The Neolithic inventions that led to the rise of man above the conditions of the Old Stone Age were made gradually in different places and probably over a long period. Some, such as the domestication of animals, took place more than once. In a famine, a wild animal will sell itself into slavery to man for the food that will preserve its life. Thus, cattle and goats, while certainly domesticated in Asia, may have been independently domesticated in Africa, too. African jackals may have provided one breed of domestic dog, while the donkey and the cat are African. The polishing of stone implements was probably a by-product of the grinding of red ochre, in wide demand for its magic properties since the Paleolithic and extensively used in Africa in the Mesolithic and later. One result of the grinding of ochre was to polish the grindstone, and another, when the upper grindstone was used at an angle, was to develop a sharp edge that, produced accidentally, may have led to the idea of grinding the cutting edge of celts or other tools. Repeated pecking of the flat surfaces of the grindstones that became too smooth to grind ochre efficiently led to perforation of the stone and thus to the development of the disk macehead of the Nile Valley. Archaeology must establish where and when celts were first ground; but the partly polished celts of the Fayum and Khartoum are probably the earliest forms of that tool known. The cultivation of wheat, barley, and flax probably were Asiatic developments that first entered Africa through the Nile Delta. The cultivation of one form of wheat may have originated in Ethiopia, however.

In Egypt, civilization first reached its full development *c.* 3000 BC, but though it passed through Copper and Bronze ages and introduced copper tools to The Sudan, there is no evidence of either of these ages in the rest of Africa, where a transition from the Stone Age, generally still Mesolithic in type, directly to the Iron Age took place gradually during the last two millennia and in a few places did not take place until the middle of the 20th century. In some localities, an intermediate state, when Neolithic forms were used, occurred (*e.g.*, Zaire and Ghana), but elsewhere (*e.g.*, Kenya) polished-stone celts, or axes, seem so rare that they may have been comparatively late imports from the north.

Transition from the Stone Age

(Ed.)

The Americas

The prehistoric sequence in the New World shares many essential developmental features with the Old World and provides a test for generalizations about cultural development based upon Old World materials. In the New World there is evidence for an early horizon of primitive food collectors, followed by an increasing specialization of food collecting based primarily upon differences in localized resources. These specialized collectors were followed by a tradition of food production independent of the Old World.

With food production came gradual increases in centres of population; villages were succeeded by towns and finally by centres of urban civilizations, which at the time of European contact were comparable to the ancient civilizations of the Middle East.

The absence of a suitable fossil record and of cultural remains from Early and Middle Pleistocene deposits in the New World have led prehistorians to look to the Old World as the ultimate source of the diverse populations of American Indians found in the Western Hemisphere by the early European explorers. Present knowledge of Pleistocene glaciations and of accompanying alterations in sea level indicates that the most probable route of entry for man from the Old World was via a land bridge between Alaska and Siberia, crossing what is now the Bering Strait. It appears that a dry-land crossing of this area was possible during periods of continental glaciation, until about 10,000 years ago. The subsequent flooding of this region has hidden whatever traces these early migrants may have left of their arrival on the threshold of the American continents, and it is necessary to look to the interior of North America for evidence of their presence. Although these early horizons of American prehistory are little known, a few sites in central Mexico have cultural remains or other possible evidences of man in a context suggesting occupation as early as 20,000 years ago. At no site in this early context are there any types of implements distinctive enough to be recognized in a context of crudely chipped stone tools from later horizons.

EARLY CULTURES

The earliest well-defined cultures in the New World have been placed by radiocarbon dating at about 9000 to 10,000 BC. At this period, two distinct traditions in North America are known: the Paleo-Indian big-game hunters of the Great Plains and eastern North America, and the Desert-culture peoples of the western basin–range region.

Paleo-Indian tradition. The oldest remains of the Paleo-Indian tradition are found on sites where large Pleistocene mammals were killed and butchered. The most distinctive artifact type of this horizon is the Clovis Fluted projectile point, a lanceolate point of chipped stone that has had one or more longitudinal flakes struck from the base of each flat face. These points are accompanied by side scrapers and, in one instance, by long cylindrical shafts of ivory. They are most frequently associated with mammoth, although associations with extinct species of bison, horse, and camel have also been reported.

The Clovis point

A second Paleo-Indian horizon, which seems in part to be contemporary with the Clovis material and partially to postdate it, is the Folsom phase of the central high plains. It is characterized by lanceolate points of more careful manufacture (including broader fluted surfaces) than Clovis, associated with the remains of extinct *Bison antiquus*. The Lindenmeier site, a Folsom campsite in northeastern Colorado, has yielded a wide variety of end and side scrapers, gravers, and miscellaneous bone artifacts. Clovis sites have been dated at about 9000 BC by radiocarbon, and Folsom sites at about 500 to 1,000 years later. Fluted points similar to western Clovis specimens have been found over most of the eastern United States south of the limits of the last major glacial advance. A single series of radiocarbon dates from the Debert site in Nova Scotia places the age of points of similar type at about 8500 to 9000 BC in that area. The distribution of this artifact type with respect to glacial events, however, suggests an appearance as early as 11,000 BC and a terminal date about 3,000 years later. In the east, several specialized varieties of fluted points may replace Clovis-type points toward the end of the Paleo-Indian occupation. While there is no instance of the discovery of eastern fluted points in association with an extinct fauna, the similarity of the accompanying assemblages of scrapers and gravers to those of the western industries suggests a similar carnivorous economic orientation in the east. Outside of the United States, fluted points have been reported at scattered sites from Alaska to Ecuador, but no certain temporal context has been established for any of these finds, and faunal associations are not clear.

Another variety of Paleo-Indian culture, which appears to be contemporary with the Clovis and Folsom phases, is characterized in its early horizons by rather crudely flaked lanceolate points that have been found associated with the bones of mammoth at two sites near Ixtapan in the Valley of Mexico and between the Clovis and Folsom horizons in a gravel pit near Portales, New Mexico. Similar points in circumstances suggesting comparable age have been found at San Jon, New Mexico, and Lime Creek, Nebraska. It appears that by about 7000 BC the fluted-point industries were replaced by a succession of lanceolate-point-using phases, which continued the Paleo-Indian hunting tradition, concentrating primarily on large, now-extinct species of bison until the onset of the Altithermal dry period about 5000 BC. The eastern limit of these cultures is in the vicinity of the western Great Lakes, while the most intensive occupation was on the western plains.

Desert tradition. The Desert-culture tradition, an adaptation of food-collecting peoples to the impoverished habitats of the basin–range area of western North America, seems to have been established by about 9000 BC. The most extensive knowledge of this way of life comes from cave or rock-shelter sites, such as Danger Cave in western

Utah, in which the desiccated remains of vegetal and animal materials have been discovered along with stone tools. The Desert peoples made intensive use of virtually all aspects of their habitat, specializing in the use of vegetable fibres for a wide variety of implements, including twine, nets, baskets, sandals, and snares. Projectile points appear to have been mostly leaf- or lozenge-shaped or lanceolate in earlier phases, with a greater use of notching for hafting in later phases. An essential feature of Desert assemblages is the milling stone, for use in grinding wild seeds. In earlier sites this is likely to be a small, thin, portable slab of stone used with a small pebble handstone, while later in the sequence, large, basin-shaped milling stones are more characteristic. Large choppers and scrapers are common in Desert sites and appear to have been used for the processing of plant materials.

RISE OF AGRICULTURE

Although the southern limits of the Desert culture are not yet clearly defined, it is known that it extended into Mexico, where, in the state of Tamaulipas, Desert materials have been found associated with the earliest known cultivated plants in the New World. Here, in the Infernillo phase, it appears that native American squash, peppers, and perhaps beans were being cultivated as early as 6500 BC. At this time, domesticates formed only a small portion of the total diet, the bulk of which was derived from wild animals and, to a lesser extent, wild plants. At about 2500 BC a primitive variety of corn (maize) first appeared in the Tamaulipas area in the La Perra phase. It appears, however, that corn was first domesticated elsewhere, possibly in the Puebla area of south central Mexico, where a date of 3600 BC is reported from materials associated with early corn in a cave near the town of Tehuacán. Even in the La Perra phase, cultivated species formed only a small part of the total diet, the majority of foodstuffs being wild plants. It appears that the development of efficient techniques of production of the three major New World domesticates—corn, beans, and squash—was necessary before real sedentary village and town life was possible in most of nuclear America. This level of efficiency seems to have been reached between 2000 and 1500 BC in Meso-America and Peru. Thus, there is evidence in the New World for plant domestication comparable in age to that of the Old World, but for many years this was unattended by the development of village life that closely followed domestication there.

Primitive variety of corn

OTHER DEVELOPMENTS

While the earliest cultivation was under way in Middle America, other areas of the New World also show evidence of interesting developments. At the site of Palli Aike, on the Strait of Magellan, the earliest cultural horizon has yielded a radiocarbon date of about 8000 BC, indicating that man reached the southern extremity of the New World well before 10,000 years ago. In the Northern Hemisphere, food-collecting cultures were well adapted to several specialized ways of life by about 4000 BC.

Archaic tradition. In the eastern United States, two basic traditions utilizing the woodland areas appear to have grown from an earlier culture that was present in that area by 6000 or 7000 BC. This early Archaic tradition is best known from the Modoc Rock Shelter in southern Illinois and from Graham Cave in Missouri and Russel Cave in Alabama. It differs from preceding Paleo-Indian horizons in its orientation toward a broad range of resources, including plant foods, as evidenced by the frequent use of milling stones. While some projectile points from these sites suggest Paleo-Indian varieties, the majority are stemmed or notched and differ in flaking technique from contemporary western Paleo-Indian specimens. By 2500 BC the Archaic cultures of eastern North America had separated into several distinct phases. There appears to have been a major division between peoples adapted to a riverine environment in the south and those adapted to the lacustrine resources of the north. Both depended, to a large extent, on the forest resources bordering these aquatic habitats. The Middle Atlantic coastal area appears to have supported another type of Archaic culture, and the boreal forests of the north yet another. In areas without concentrations of particularly favourable resources, a generalized Archaic culture similar to the earlier pattern seems to have persisted. Most Archaic cultures are characterized by a rather extensive use of ground-stone implements, both woodworking tools and other categories, such as bowls, knives, net sinkers, and elaborate weights for spear throwers. Projectile points vary widely but are usually rather large and crude and are stemmed or broadly notched for hafting. Perhaps the most interesting of the late Archaic manifestations is the Old Copper culture of the northern Great Lakes area. Here, exposures of native copper were quarried and cold-hammered into implements, such as projectile points, knives, awls, and axes; and highly valued copper from this region was traded over much of eastern North America.

Riverine and lacustrine environment

Western North America. In western North America, similar developments were under way during this same period. It appears that the more arid regions of the basin–range country were largely depopulated during the Altithermal dry period (from about 5600 to 2500 BC) and that in surrounding regions diversification and specialization took place. In the drainages of the major rivers of the northwest, such as the Columbia and the Fraser, the annual abundance of salmon was the basis of a cultural adjustment as early as 7000 BC. Implements of this horizon are similar to those found earlier in the Desert culture, with projectile points, the most diagnostic artifact types, tending to be long and leaf-shaped or slightly stemmed and with a few notched forms also present. Following the Altithermal drought, a broad horizon characterized by the use of indented-based points with serrate-edged blades (generally termed "Pinto-like," after the type locality in the Pinto Basin of California) is found over much of the southern portion of western North America. In at least one of the phases representing this horizon, the Chiricahua of southern Arizona and New Mexico, it appears that primitive corn cultivation was practiced. The site of Bat Cave in western New Mexico has produced specimens of a type of primitive corn that is also known from the Flacco phase in Tamaulipas at 2000 BC but that is here in association with a Chiricahua assemblage from which materials have been dated at about 1000 BC.

Salmon as basis of cultural adjustment

South and Middle America. By 2500 BC, techniques of cultivation had also reached the northern coast of Peru, where, at such sites as Huaca Prieta at the mouth of the Chicama Valley, there was a mixed dependence upon marine foods such as sea urchins, mollusks, and fish; upon wild plants, mostly tubers and roots; and upon cultivated plants, including beans, peppers, and a different genus of squash than that cultivated in the early horizons in Tamaulipas. Gourds and cotton were also grown, the gourds for use as containers and net floats, the cotton for twined fabric and cordage. The use of stone at Huaca Prieta is interesting in its simplicity. Crude flakes and shattered pebbles compose the entire chipped-stone industry, while pecked and ground-stone artifacts are chiefly perforated net sinkers. In the upper levels of the site are architectural remains consisting of one- or two-room, small cobble-walled subterranean houses. The absence of ceramics at the Huaca Prieta site poses a number of interesting problems. From the Valdivia site in Ecuador, several hundred miles to the north, radiocarbon samples indicate that ceramics may have been present there as early as 2500 BC, and another date from Panama indicates that the ceramics of the Monagrillo phase were manufactured by about 2000 BC. Present knowledge of the northern coast of Peru does not reveal ceramics before about 1200 BC, indicating an isolation of this area from cultural developments to the north. With ceramics, corn and other indications of Middle-American influence appear in Peru.

VILLAGE FARMING AND TOWNS

The appearance of village farming in the upper levels at Huaca Prieta and in the immediately succeeding Guañape phase in surrounding areas is roughly contemporaneous with the first appearance of this way of life in the Valley of Mexico at such sites as Zacatenco and El Arbolillo. Here a relatively sophisticated ceramic tradition (clearly derived

Ceramic tradition

from elsewhere) appears in the earliest levels. While evidence for architecture is not completely clear, it appears that by about 1500 BC there were small villages of wattle-and-daub huts scattered along the shores of the lakes of the Valley of Mexico, with inhabitants subsisting largely on corn–bean–squash cultivation, supplemented by the meat of game animals and by various aquatic resources.

Earliest evidences for the next cultural advances are apparent by about 800 BC in changes in architecture and settlement pattern in several areas of Middle America and Peru. At this time, fairly extensive public works are represented by temple structures and large sculptured monuments, which occupy a central position in towns and villages. Phases as widely separated as the Olmec of Veracruz and the Cupisnique of coastal Peru appear to be linked not only in time and patterns of basic subsistence but in specific ritual practices involving a jaguar or feline deity. Throughout Middle America and in the Andean area, this appears to have been a time of consolidation and establishment of the basic traditions that dominated the development of high cultures in the New World up to European contact.

Hopewell culture. The spread of cultivation into North America seems to have proceeded along two separate courses, one from northern Mexico into the southwest and the other from an unknown Middle American source into the Mississippi Valley. One of the earliest known phases in eastern North America in which corn cultivation appears to have had a role in subsistence is the Adena, which occupied the middle Ohio River Valley by about 800 BC. The stimulus of the Adena farmers was apparently instrumental in bringing about the spectacular Hopewell culture in the Illinois and Ohio valleys. The success of the Hopewell peoples (400 BC to AD 400) seems to have been due largely to their combining elements of the preceding Archaic cultures with elements of the Adena culture and perhaps with some features of a local cultivating tradition. It is evident that the Hopewell culture included a well-organized village-based society in which surplus resources were used in the construction of elaborate earthworks and were concentrated as wealth in a restricted group of individuals. The most outstanding feature of Hopewell culture is a burial complex that called for the deposition of concentrations of wealth in tombs of one or several deceased individuals. The interment procedure was elaborate and involved the construction of a large log tomb, later burned and covered by an earth mound. Artifacts found within these burial mounds indicate that the Hopewell were able to obtain goods from widespread localities in North America. Obsidian and grizzly-bear teeth were apparently derived from the Rocky Mountain region, copper from the northern Great Lakes, and conch shells and other exotic objects from the southeast and along the coast of the Gulf of Mexico. The ceramics of the Hopewell appear to be based in two major traditions, one derived from northern Asia, which reached eastern North America by about 1000 BC, and the other from Middle America, where the decorative technique of rocker-stamping, characteristic of finer Hopewell pottery, existed several hundred years prior to the earliest appearance of the Hopewell culture. In less favourable areas of eastern North America, a "generalized Woodland" culture paralleled the Hopewell in time, probably based more on collecting than on cultivation.

The Adena phase

Mississippi culture. The period of the Hopewell culture was followed by relative decline in social cohesion in the northern Mississippi and Ohio valleys, evidenced by the absence of unifying features comparable to the Hopewell in the succeeding generalized Woodland culture. At about AD 800 a new tradition, with much stronger and more specific Middle American elements, moved up the Mississippi Valley. This Mississippi culture was based on more intensive cultivating techniques than the Hopewell and resulted in impressive concentrations of population in large towns through the southern and central Mississippi Valley and in several areas of the southeastern United States. A central ceremonial plaza provided the nucleus of a Mississippi town, and each settlement had one or more pyramidal or oval earth mounds, surmounted by a temple or chief's residence, grouped around the plaza. This settlement pattern is typical of most of Middle America after about 850 BC but is not found in North America until the Mississippi culture appears. The scale of public works in the culture can be estimated from remains of the largest of the Mississippi earthworks, Monk's Mound near Cahokia, Illinois, which measures 1,000 feet in length, more than 700 feet in width, and is still 100 feet in height. The first European explorers in the southern Mississippi Valley in the early 16th century found the Mississippi culture still flourishing as warring alliances of towns, each ruled through a theocratic system based on kin ties.

Pueblos. In the southwest, the earliest villages of farmers appeared by about 200 BC, and this initial development in southern New Mexico and Arizona was succeeded by a gradual spread of this way of life as far north as southwestern Colorado, east to the Pecos River, and west into the lower valley of the Colorado River. The maximum expansion of the Puebloan culture of the eastern and northern portions of the southwest appears to have taken place by AD 1150 or 1200 and was followed by the gradual abandonment of much of the area by farming peoples. This decline seems to have been due to a combination of factors, including drought, deforestation, and lack of social cohesion within the villages. At the time of historic contact the Puebloan peoples were restricted to the Rio Grande Valley and adjacent localities and to scattered settlements in west central New Mexico and on the Hopi mesas of Arizona. The early explorers encountered other less well-organized farming groups, descended from the Hohokam and Patayan traditions of the southwest, in scattered localities along the Gila, Salt, and Colorado rivers.

South America. In South America, little is known of cultural development outside the Andean area, where, as in Middle America, urban civilization was well under way by the first few centuries AD. From a sequence near the mouth of the Orinoco, it appears that manioc cultivation, which formed the subsistence base for stable villages in the tropical forest, had been developed by about 1000 BC. Peripheral to the Andean area, numerous cultures are known, particularly in Colombia and northern Argentina and Chile, that show marked influence from Andean urban centres and yet preserve distinct local traditions throughout the late prehistoric period.

Manioc cultivation

The general picture. An overall view of the prehistory of the New World prior to the development of urban civilization reveals several general trends. The outline above follows the forefront of cultural development as it took place in several well-known areas. In localities less favourable to primary or intensive cultivation, the level of cultural development tended to stabilize at the point at which maximum food production was possible with the techniques at hand. Thus, in the Arctic and in the boreal forests of the north, as well as through most of southern South America and various other regions unfavourable for cultivation, cultural activity remained at an Archaic food-collecting level through the entire prehistoric period. In the tropical forests of South America and the woodlands of the northeastern United States, farming villages were the apex of cultural development under prehistoric conditions. In relatively favourable areas, such as the Mississippi Valley, the oasis regions of the southwestern United States, and several other regions peripheral to the South and Middle American high-culture centres, temple-centred towns were the climactic development. A general appraisal of cultural complexity reveals a trend from a single or few early cultural phases of uniform composition covering the entire New World, to the extremely diversified cultures of the last two millennia of the prehistoric period. Within the sequence of cultural development, it appears that the greatest diversity is present at the village-farming level, with hundreds of distinct phases indicating essentially locally oriented social groups that gradually united into larger units as communication and political pressures from more successful centres submerged the cultures of the weaker local phases.

When compared with the Old World sequence, a similar succession of cultural levels can be distinguished in the New World, but there are differences in such basic qualities as the lack of economically important domestic ani-

mals in the New World and the much greater diversity of habitats and forms in which the various cultivated plants originated. These factors seem basic in explaining the wide discrepancy in rapidity of cultural development between the Old and the New World once the idea of cultivation was present. It was not until several cultivated crops (corn, beans, and squash for most of the New World) were fully developed and assembled that higher cultural levels were possible. (R.M.As./Ed.)

Oceania

The archaeology of Oceania involves, for the most part, short time perspectives, because migrations within the open Pacific could have occurred only after the development of seagoing canoe navigation in Neolithic times. The exception is the New Guinea–Australia region, where the ancestors of the Australoid- and Negritoid-type peoples evidently arrived in Paleolithic times.

Theories of migration

The long-term history of the Oceanic peoples, especially the Polynesians, has been the subject of many theories. Scholars reject ideas involving a lost continent (*e.g.,* Lemuria, Mu) or direct relations with the Middle East (*e.g.,* the Ten Lost Tribes, migrations of Children of the Sun from Egypt), early India (*e.g.,* Indus Valley–Easter Island connections), or Japan (*e.g.,* supposed language relations). They also insist that, while eastern-voyaging Polynesians could well have reached the American continent and some may have found their way back into the islands, none of the various theories claiming that Oceanic peoples had their homelands in North or South America is scientifically credible (*e.g.,* E. Rout's imaginative *Maori Symbolism,* T. Heyerdahl's thesis for the "Kon Tiki" voyage). Similarly, they reject theories explaining the Pre-Columbian civilizations on the American continent in terms of influences by way of the tropical Pacific islands from Asia. Most archaeological as well as racial, linguistic, and ethnological evidence continues to support the long-standing hypothesis of the settlement of Oceania by a succession of migrants from the Southeast Asia region, with at most very minor contacts eastward to America.

The archaeological record begins when early *Homo sapiens* populations, comparable with the fossils of Wadjak in Java, Aitape in New Guinea, and Keilor and others in Australia, were moving eastward. This apparently occurred during the Fourth Glacial Epoch, when sea levels were lower, land pathways perhaps more uplifted, and inter-island channels narrower than now. Early man could migrate with lessened water obstructions from the Asiatic continental platform (Sunda Shelf) through the intermediate Celebes–Molucca–Lesser Sunda zones on to the Australian continental platform (Sahul, or Papuan, Shelf). Some scholars suggest such movements even during the Third Glacial, but this seems dubious. Core tools typologically Paleolithic and sometimes heavily patinated occur in both Southeast Asia and Australia, and crude flake tools often of microlithic size are found in the intermediate zones (*e.g.,* Timor) as well. Yet all these are surface finds or of dubious age when at subsurface levels. Doubtless most early groups moved, in glacial times, near shorelines now inundated, though valleys suitable for inland hunting or locally uplifted coasts might yield finds. The vital western New Guinea region, much of which is swampy and lacking stone, has had little systematic study.

The hypothetical picture has the isolated Australian and Tasmanian populations developing along regional and local lines characteristic of their later archaeological perspectives: generally Paleolithic, but with some Neolithic and also recent Malay trading contacts along the northern coasts. The New Guinea region was penetrated by canoe-migrating peoples carrying Neolithic elements that come to dominate the picture. The evidence suggests that the first comers were "dry" gardeners living in semi-sedentary hamlets, with crude stone tools including adzes and axes of oval cross section; such shifting cultivators are found still from Southeast Asia to the Solomon Islands and Vanuatu (New Hebrides). The Melanesian areas were also penetrated, apparently later and especially along the coasts, by village-living peoples with more sedentary cultivation; stone construction in various forms; finer tools, including quadrilateral cross-section types; stone pestles and mortars; pottery; and other later Neolithic elements. In New Guinea, as shown by A. Riesenfeld, these influences appeared to arrive by way of the northeastern coast from the outer fringe of islands, perhaps the Bismarcks, Admiralties, and others. Unhappily, archaeological work in all these Melanesian zones has been limited to sporadic surface collecting and recording and to occasional tools turned up by garden workers or miners; time-sequence definitions are out of the question except as they may be inferred from Neolithic chronology in Southeast Asia.

By contrast, archaeological work in Polynesia and some zones of Micronesia is considerably more advanced. For northwestern Micronesia a vital clue is a radiocarbon dating of approximately 1527 BC from a stratified deposit excavated in the Marianas. Collateral evidence suggests that occupation may go back to 2000 BC. The Mariana *latte* sites (rows of capped stone pillars, probably posts of important houses), together with stone mortars, pottery, and other artifacts, suggest migrations from the Philippines area in late Neolithic times.

The Mariana *latte* sites

Farther east, the low coral islands of the Carolines, Marshalls, and Gilberts yield limited artifacts of shell, bone, and coral rock capable of some comparative study. The few high islands in this part of Micronesia, however, have extensive stone construction and other more diversified elements. Yap, for example, has stone ceremonial platforms, stepped tombs, "stone money," pottery, etc. Ponape's most spectacular site, the "Venice" called Metalanim, has several acres of stone-faced islands and canals, the principal structure being a rectangular enclosure with double walls up to 40 feet high containing a central stepped tomb and also vault tombs. A much smaller Venice exists on Kosrae (formerly Kusaie), most easterly of the Carolines. Scholars generally attribute such elaborations of the basic stonework elements, in Polynesia as well as Micronesia, to local creativity rather than undemonstrated outside influences.

Micronesia and Polynesia may usefully be treated as a continuous zone. With them, too, may justifiably be placed the Fiji zone of eastern Melanesia, the most easterly limit of the potter's craft. In compiling known data on stonework in this whole area, one may distinguish two great types: one to the west (Micronesia, Fiji, Tonga, Samoa areas), characterized by ceremonial courts, perhaps with god houses, and platform tombs often of stepped types; the other to the east (Societies, Hawaii, New Zealand, and islands eastward to Easter Island), characterized by temple structures, usually with altars, standing stones probably as backrests for gods and priests in the rituals, and cist burials within the structures. From simpler shrines having a few standing stones (*e.g.,* inland Tahiti, New Zealand, small, outlying Hawaiian islands), the Polynesian temple became elaborated into various local forms including usually larger-size and "megalithic" stonework. In Hawaii wooden posts or images generally replaced standing stones. In the Tuamotus huge coralline slabs were often used, and in the Marquesas the stones were sometimes carved in human (or god) form. Some specialists consider that such comparative study solves the "mystery" of the Easter Island statues. Rather than being relics of some lost continent or pre-Polynesian migration, they follow this last pattern of carved figures, standing on altar platforms (called by the standard Polynesian name *ahu*) in which there are cist burials. Apparently a local inventive urge toward large size, combined with the presence of easily worked volcanic tuff, produced this one of the many variants of the Polynesian place of worship.

Types of stoneware

Such local constructions, however, together with other more spectacular elements, such as widely scattered petroglyphs and a dubiously old "script" on Easter Island, have less significance for historical reconstruction than detailed study of variability in minor artifacts.

Theories of contacts with the American continent must be treated with caution so far as they lean on gross parallels such as stone images or art resemblances; the most concrete evidence has been the presence of the sweet potato, apparently an American plant, in Oceania in prehistoric times. (F.M.Kg.)

CIVILIZATIONS

It is customary to regard the time span of modern civilization as immensely long, extending back through medieval Europe to the Greco-Roman world and to still earlier Oriental forerunners. But the length of time for this development is probably no greater than that which had elapsed between the prior introduction of agriculture and the rise of the first civilized states. And both of these intervals together probably constitute only about 1 percent or less of the time since two-legged, tool-using creatures made their first appearance. Furthermore, the extensive spread of civilization is much more recent still. Civilizations have prevailed over uncivilized food collectors, nomads, and primitive farmers across most of the world only within recent centuries. Thus, it must be stressed that civilization is a late condition in the human record.

Since at least the days of L.H. Morgan (1818–81), the achievement of civilization has been widely recognized as a major stage in man's social and cultural evolution. Coming ultimately as a consequence of the attainment of settled food production, its main attributes were substantial, qualitative increases in the scale and complexity of the limited number of societies that were its initial representatives. As there is civilization, so there are civilizations. And at least until well after the end of prehistory by any definition, the early civilizations must be regarded as very limited enclaves in a largely uncivilized world.

The earliest civilizations of the Old World occupied the great alluvial valleys of the Middle East by around the year 3000 BC. There is some evidence that Egyptian developments were stimulated by Mesopotamian influences, but the beginnings of civilization in the two areas were nonetheless essentially independent of one another and more or less contemporaneous. Farther to the east, civilization made its appearance in the valley of the Indus River a half millennium or so later. Here it was also essentially independent in at least its stylistic manifestations, but the time lag and the apparently abrupt inception of urban life—abrupt, that is, relative to the development of urban life in Mesopotamia—suggest that stimulus from the latter area probably played a considerable part in its origin. The civilization of northern China, still farther afield, did not cross the same threshold until perhaps the middle of the 2nd millennium BC, and controversies continue regarding the role of the already well-established Middle Eastern civilizations in its formation. From these four primary centres, lesser and later centres naturally developed across the Eurasiatic hinterland. This process of consolidation and expansion is, however, a part of the historic, and not the prehistoric, record.

A direct comparison of the Old World sequence with that of the New World is complicated by somewhat different cultural emphases. But while criteria for the identification of early civilization in the two hemispheres are not entirely the same, most authorities would concur in assigning the first appearance of a civilized way of life in the New World to the middle centuries of the 1st millennium BC. In view of the evident delay in this achievement, the possibility that it was prompted by influences from eastern Asiatic centres has long been debated. Apart from a few isolated and somewhat dubious cultural elements, however, the general course of development was sufficiently different in the New World to convince most authorities that it was essentially independent of external stimuli.

Two centres of focusses of early civilization in the New World came into existence roughly at the same time, in Middle America and in the central Andes. Stylistic affinities suggest the possibility of a brief period of interconnection between them at the outset, perhaps the result of the southward spread of an early Meso-American cult. As in the case of Egypt and Mesopotamia, however, the subsequent development of the two major focusses proceeded along substantially independent lines. As late as the time of the Spanish conquest in the 16th century, the powerful Mexican and Inca empires had little or no direct knowledge of one another.

The Urban Revolution

Civilization is of concern in the present context, of course, only as the terminus of the long prehistoric record. In most areas the rise of civilization coincided at least roughly with the development of writing, so that the later courses of civilization are illuminated by historical sources. In the absence of documents, the limitations of purely archaeological sources entail a stress on certain aspects of the rise of early civilizations—primarily those involving formal governmental or religious institutions and the material–technical facets of culture more generally. To be sure, a comparison with civilizations that have left written records or with non-Western civilizations of today suggests many other dimensions in which the early civilized societies must have diverged from their predecessors and noncivilized contemporaries. The U.S. anthropologist Robert Redfield, for example, identified the rise of civilization primarily with the formation of great traditions that systematized and redirected the cultural and moral order as it had been previously conceived in scattered folk communities. But except as indirectly inferred from objects associated with ritual or from tenuous parallels with modern non-Western societies, such a dimension of change as this is virtually irrecoverable by the prehistorian.

The possibility of attaining a civilized way of life rested ultimately on conditions that only intensive agriculture made possible. Among these were storable food supplies permitting permanent residence, agricultural surpluses supporting the proliferation of full-time administrators and specialists freed from primary subsistence activities, and a sufficient intensity of production within a given area to encourage the formation of urban centres. Recognizing these as prerequisites for civilization, V. Gordon Childe formulated the growth of civilization as the consequence of two successive revolutions. The first, or Neolithic, revolution has been dealt with in a preceding section; it encompassed the domestication of the plants and animals essential for an agricultural mode of subsistence, as well as the formation and spread of settled village communities. For the second, the Urban Revolution, a number of criteria have been proposed. Most importantly, these include: (1) cities, or large, dense settlements; (2) the differentiation of the population into specialized occupational groups; (3) social classes, including a ruling stratum exempt from primary subsistence tasks; (4) mechanisms for extracting a "social surplus," such as taxes or tribute; (5) monumental public buildings and other enterprises; and (6) writing.

Criteria for the Urban Revolution

It is important to note that these two proposed revolutions differ fundamentally in character. The first represented most importantly an advance in man's control of his environment and consisted of a series of discoveries permitting its vastly fuller exploitation. The Urban Revolution, on the other hand, affected man's relation to his environment only secondarily and much more slowly. As agricultural techniques became stabilized, the crucial changes increasingly were those in community size and composition, in the appearance of new institutions, and in the vastly greater complexity of patterns of social organization. These are changes less in man's interaction with his habitat than in that with his fellows. Hence, different interpretive skills are required to elucidate them than to explain the changes which accompanied the beginnings of agriculture, skills rooted more in the social sciences and humanities than in the natural sciences. Moreover, because of the increasing diversity of the social and cultural scene, our understanding of these changes is increasingly subject to distortions arising from the accidents of preservation or the unrepresentativeness of evidence so far obtained. The preoccupation of archaeologists with late prehistoric and protohistoric temple furnishings and architecture, for example, probably has exaggerated the importance of temples and certainly has left us with disappointingly little information on the nature of contemporary secular life in neighbouring precincts of the same towns. By contrast, it

can be (or, at least, has been) generally assumed that the relative absence of social and cultural differentiation in early villages permits them to be adequately sampled by random small-scale soundings.

If we accept the general descriptive validity of the criteria for civilization given above, the identification of the rise of civilization with the sequence of a Neolithic revolution and an Urban Revolution is still subject to three important qualifications. In the first place, to say that civilization depends on agricultural surpluses is not the same as saying that agricultural surpluses were all that was necessary to bring civilization into being in any given area. There is simply not enough known of the productivity of prehistoric agriculture to indicate whether increasing supluses might have triggered subsequent changes in other aspects of culture. In fact, on the basis of present evidence it seems at least as likely that it was primarily through changes in social organization that civilizations brought about that there were advances in agricultural productivity. Second, it must be stressed again that the criteria given above are only those that are in some way accessible to the archaeologist. The tendency to regard them as a self-contained list that somehow defines civilization, or even that somehow embraces its primary causes and consequences, must be avoided. Third, the term revolution implies sudden and violent change. Yet, if the rise of civilization was undoubtedly rapid as compared with the hundreds of thousands of years of preagricultural life, it is equally clear that it was spread over many generations and probably never was as apparent to its participants as the Industrial Revolution has been to recent generations. Authorities are not even agreed, moreover, that the Neolithic and Urban revolutions were events separable in time. It remains only a hypothesis that one was consummated before the other began, and an equally plausible hypothesis might see a single, rising tempo of change that began with incipient agriculture and culminated in the appearance of Oriental Bronze Age monarchies.

There remains a further difficulty with the criteria mentioned earlier: they apply most usefully to the earliest, or "type," sequence in Mesopotamia, and elsewhere diverge to a greater or lesser degree from the observed archaeological data. The Maya area of Meso-America, for example, lacked true cities yet exhibits the other elements in the list to such a degree that its status as a civilization seems indisputable. According to some scholars, cities were also absent in Egypt during the Old Kingdom, although the other appurtenances of a civilized state are unusually well represented there. Similarly, the highly organized Inca Empire got along entirely without writing, while in the Mayan and Mexican areas the much earlier invention of writing was only applied to economic and administrative tasks in the centuries immediately preceding the Spanish conquest.

To phrase the problem more broadly, the separate focusses of early civilization tended to crystallize into patterns that in some respects are quite different. Mesopotamia was characterized by extensive irrigation agriculture, by fairly rapid technological progress (though hardly rapid enough to "explain" the more profound social changes that took place concomitantly), by emphasis on urban settlement, and by an early separation between the fundamental economic and administrative units (which were often temples) and the political leaders. In Egypt, one finds irrigation agriculture again and an even more precocious elaboration of a state administrative apparatus, but there is an absence of the trends toward rapid technological advance, urbanization, and secularization that are so evident in Mesopotamia. A further variation on this pattern is found in the Maya area, where agriculture was largely of a nonintensive slash-and-burn variety, where the bureaucracy of a centralized state was entirely absent, and where primary cultural emphasis seems to have been placed on the elaboration of esoteric cults by hierarchies of priests for whom important nonreligious functions in the society are not apparent. Peruvian civilization, on the other hand, was somewhat closer to the Mesopotamian model, except perhaps in its greater emphasis on an ethnically distinguished ruling stratum. In both New World civilizations, but in neither of those of the Old World, it might be added, there are indications that monumental temple architecture long preceded the other criteria of civilization—and in fact may be as old as the earliest village settlements based on agriculture.

These divergences should not be taken to deny the regularities that are implied by Childe's criteria for the Urban Revolution. Given a sufficient level of abstraction, the early civilizations of both hemispheres shared most of their essential features. And certainly they constitute a distinctive sociocultural type when contrasted with all earlier folk societies. They form the culminating stage of the prehistoric record, and it is worth noting in conclusion that their essential attributes never subsequently disappear. In fact, in one sense the world's written history is the documenting of their spread: gradually and irregularly, based on a succession of different focusses (of which western Europe and the United States are only the most recent), but ultimately irresistible.

(R.M.As./Ed.)

The example of the Middle East

EVOLUTION OF MIDDLE EASTERN CIVILIZATIONS

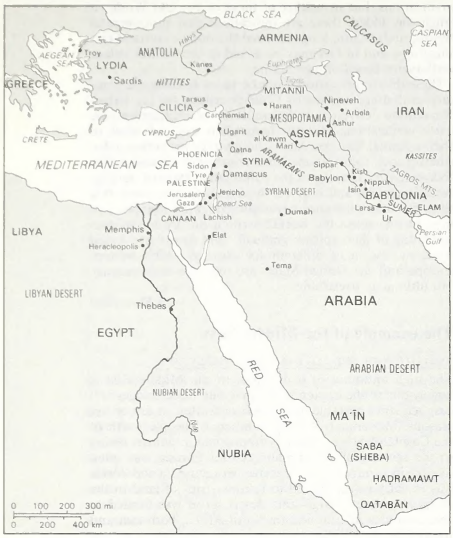
The high antiquity of civilization in the Middle East is largely due to the existence of convenient land bridges and easy sea lanes passable in summer or winter, in dry or wet seasons. Movement of large numbers of people north of the Caspian Sea was virtually impossible in winter, owing to the severity of the climate; central Eurasia was often too dry in summer. Land passage between Asia and Africa was in early times limited to narrow strips of land in the Isthmus of Suez. Large-scale desert travel was limited to special routes in Iran and in North Africa, both east and west of the Nile Valley.

Origins of agriculture

Another reason for the early significance of this area in world history is the fact that the water supply and the climate were ideal for the introduction of agriculture. Several species of grain grew wild, and there were marshes and tributary streams that could easily be drained or dammed in order to sow wild wheat and barley. The seed had only to be strewn over a sufficiently moist surface to ensure some kind of crop under normal conditions. It is therefore not surprising that there is evidence of simple agriculture as far back as the 8th or 9th millennium BC, especially in Palestine, where more excavating has been done in early sites than in any other country of the Middle East. Many bone sickle handles and flint sickle edges dating from between *c.* 9000 and 7000 BC have been found in Palestinian sites.

In Mesopotamia and Iran remains of this period appear in caves on the lower slopes of the Zagros Mountains between western Iran and Iraq. The date of the systematic introduction of irrigation on a large scale in Mesopotamia is somewhat doubtful because most of the early sites of irrigation culture were covered long ago by accumulation of alluvial soil brought down by the spring floods of the Tigris and Euphrates rivers. Archaeologists once thought that all irrigation originated in the foothills of the Zagros and that the earliest true farmers lived in the plains of Iran. But recent excavations and surface explorations have proved that irrigation around the upper Tigris and Euphrates, as well as their tributaries, dates from the early 6th millennium BC (*e.g.*, at al-Kawm on the Upper Euphrates). Small-scale irrigation was practiced in Palestine (*e.g.*, at Jericho) in the 7th millennium BC.

In northern and eastern Mesopotamia, main streams were soon partly diverted during moderate river floods into canals running more or less parallel to the rivers, which could thus be used to irrigate an extensive area. Such deflector dam irrigation avoided the self-destructive weaknesses of large storage dams, in particular the danger of depositing great masses of refractory mud in the storage basin behind the dam. In the north and east considerable urban installations developed at sites such as Nineveh no later than the 5th millennium BC, when southern Mesopotamia was still mostly swampland like the early Egyptian delta. The Euphrates had a much smaller flow of water than the nearby Tigris. The latter was much swifter,



The ancient Middle East.

Adapted from *Everyman's Classical Atlas* (1963), J.M. Dent & Sons Ltd.; map © John Bartholomew & Son Ltd.

however, so that it was potentially more important for irrigation, even though much harder to tame.

The Egyptian Nile had a much more predictable water flow than the Mesopotamian rivers because it flowed through hundreds of miles of swamp, where unusually high annual floods spread out, interfering with navigation but averting the danger of the occasional destructive inundations of Mesopotamia.

Mesopotamia and Egypt to *c.* 1600 BC. The oldest known urban and literate culture in the world was developed by the Sumerians in Mesopotamia beginning in the late 4th millennium BC. About 2300 BC a Semitic leader, Sargon I, conquered all of Babylonia and founded the first dynasty of Akkad (Akkadu), which held power for about a century and a half. Sargon and his successors were the first known rulers in southwestern Asia to gain control of the Fertile Crescent as well as of adjacent territories. They sent trading expeditions to central Anatolia and Iran and as far as India and Egypt. After the fall of the dynasty of Akkad there was a Sumerian revival under the 3rd dynasty of Ur (Ur III [21st–20th centuries]), followed by another influx of Semites. These people founded the first dynasty of Babylon (19th–16th centuries), whose most important king was Hammurabi. In the 17th century new ethnic groups appeared in both Babylonia and Syria-Palestine: Kassites from the Zagros Mountains, Hurrians from what is now Armenia, and Indo-Europeans from Central Asia. This period marked the end of the formative phase of Mesopotamian civilization.

The Sumerians

Shortly after 3000 BC the numerous small states that had arisen in the Nile Valley during the 4th millennium were united under the 1st dynasty of Egypt. At this time the Egyptians had already developed a system of writing. Between *c.* 2686 and *c.* 2160 BC their country was united under a powerful monarchy (the Old Kingdom) served by a complex bureaucracy.

Toward the end of the 3rd millennium there was a period of disunity, followed by reunification under the 12th dynasty (1991–1786).

During these two centuries Egyptian control was established over Nubia, Libya, Palestine, and southern Syria. Soon after 1800 BC the Egyptian Empire fell apart, and *c.* 1700 Egypt was overwhelmed by the Asian "Hyksos," who ruled the country for a century and a half.

New states and peoples. Before the close of the 16th century BC the native 18th dynasty rose in Egypt; it expelled the Hyksos and founded the New Kingdom. The New Kingdom rulers moved back into Syria-Palestine and came into conflict first with the Hurrian state of Mitanni and later with the Anatolian Hittites, who were expanding into Syria from the north in the 14th century BC. The Amarna Letters (diplomatic correspondence written in Babylonian script and language and discovered in Egypt by archaeologists) are an important source of information on this period. In Mesopotamia the dominant powers were Kassite Babylonia and Assyria (which emerged from subjection to Mitanni in the early 14th century BC). Relations between states were governed by elaborate treaties, which were constantly being broken. After the fall of Mitanni (*c.* 1350) the Hittites and Babylonians both directed their hostility against Assyria. Kassite Babylonia was subjugated by Assyria *c.* 1230. This, followed by the fall of the Hittite Empire (*c.* 1200), ended what has been called the first "International Age" in the civilized world.

The invasions of the 13th century BC

The latter part of the 13th century BC saw the irruption of new peoples into the Aegean, Anatolia, and the Fertile Crescent; their appearance coincided with the Trojan War, the collapse of the Hittite Empire, and the destruction of many coastal cities of Greece, Cyprus, and Syria-Palestine. Best known of the new settlers from the west are the Phrygians, who occupied most of the old Hittite heartland, and the Philistines, who moved into Palestine.

At the same time, in Transjordan and western Palestine, the Hebrews founded a tribal confederation that was changed into a monarchy by Saul and David (*c.* 1020–960 BC).

In the east the Iranian tribes, led by the Medes, were pouring into Iran from Turkistan. From the south and west came the Semitic Aramaeans. The Aramaeans and Medes were to transform the ancient Middle East.

The Assyrian state suffered an eclipse in the 11th century BC, when the Aramaeans and related tribes occupied most of its territory. It was not until the late 10th century that the Assyrians began to recover, but by 850 they had conquered much of western Media and southern Armenia as well as Babylonia and Syria. In the following centuries, until just before 630, the empire was greatly expanded. It was also highly organized administratively; its language became Aramaic.

The Canaanite Phoenicians on the Syrian coast re-established their trading communities after the Philistine and Aramaean invasions; in the 10th and 9th centuries they moved out into the Mediterranean, establishing colonies in North Africa and as far west as Spain. Their influence in the western Mediterranean declined after the 6th century. Their Carthaginian colony then took over Phoenician trade in the western and central Mediterranean.

Farther east the Medes and Chaldeans destroyed the Assyrian Empire at the end of the 7th century. The Chaldean dynasty in Babylonia carried on Assyrian traditions of administration and encouraged commerce; under Nebuchadrezzar II (*c.* 605–*c.* 561 BC) their Neo-Babylonian Empire became the most powerful political entity of its time. Its rule extended from the Taurus Mountains in Anatolia to eastern Arabia and deep into southern Iran. This short-lived state made a tremendous impression on contemporaries, especially on the Jews, whose state was destroyed and who were carried into the Babylonian Captivity, and on the Greeks, to whom the glory of Babylon became legendary.

The Achaemenian Empire and its successors. In the 6th century the Iranian Persians under Cyrus the Great conquered their Median cousins and established the Achaemenian state (549). This was followed by the conquest of Lydia (546) and the Babylonian Empire (539). Aramaic became the official language of the Persian Empire, and its official religion was Zoroastrianism. Cyrus' enlightened policy put an end to the Assyro-Babylonian practice of deporting conquered peoples and trying to destroy all local nationalisms.

At its height the Achaemenian Empire ruled the whole of the Middle East; Greek resistance prevented it from expanding successfully into Europe.

The Macedonian conquest

In 334 BC Alexander of Macedon invaded Anatolia and nine years later completed the conquest of the Persian realm. His vast empire was broken up into Macedonian "successor states" after his death. The Seleucid kings of Syria controlled most of Anatolia, Mesopotamia, and Iran. About 250 BC the still seminomadic Parthians emerged from a small area southeast of the Caspian Sea. Establishing control over Iran, they declared their independence of the Seleucid Empire and in the 2nd century BC expanded westward into Mesopotamia. In the 3rd century AD the semi-Hellenized Parthians were replaced by the Persian Sāsānians. The Sāsānians ruled Iran from AD 224 to 642; they extended its boundaries, reinvigorated its administration and cultural life, and challenged Roman power in the Middle East. In 636 the Sāsānian Empire was conquered by the Muslim Arabs, bringing to an end the last phase of ancient Middle Eastern civilization.

Pre-Islāmic Arabia. Arabia was drawn into the orbit of western Asiatic civilization toward the end of the 3rd millennium BC; caravan trade between south Arabia and the Fertile Crescent began about the middle of the 2nd millennium BC. The domestication of the camel around the 12th century BC made desert travel easier and gave rise to a flourishing society in South Arabia, centred around the state of Saba (Sheba). In eastern Arabia the island of Dilmun (Bahrain) had become a thriving entrepôt between Mesopotamia, South Arabia, and India as early as the 24th century BC.

The discovery by the Mediterranean peoples of the monsoon winds in the Indian Ocean made possible flourishing Roman and Byzantine seaborne trade between the northern Red Sea ports and South Arabia, extending to India and beyond.

In the 5th and 6th centuries AD, successive invasions of the Christian Ethiopians and the counterintervention of the Sāsānian kings disrupted the states of South Arabia. The resulting economic decline made the rapid Muslim conquest of the area an easy task in the 7th century.

ELEMENTS OF CIVILIZED CULTURE

Religion. Middle Eastern religious thought had a strong influence on the ancient Greeks. The cosmogonies of Egypt, Babylonia, Phoenicia, and Anatolia were transmitted in part to the West and formed the basis of much of the cosmogonies of Hesiod and the Orphics before 600 BC, as well as the background for the cosmogonies of Thales and Anaximander in the 6th century BC. There is some influence from the Middle East in Pythagorean and Platonic thinking, but it is often hard to define and still harder to prove in detail. From the early 3rd century BC on, the Middle East began to influence Greek thought increasingly. Babylonian astrology influenced Stoic philosophy, and some Jewish influence on Stoic ethics is likely as well.

Late Egyptian religious speculations were transmitted to the Greeks, especially through the "Physiologists" and Hermetic writers who flourished in Hellenistic Egypt beginning in the 2nd century BC. Astrology and alchemy were transmitted to our time in substantially the forms they received in Hellenistic and Roman Egypt.

Jewish influence

With the partial Hellenization of Judaism and its Christian offshoot in the 1st century AD, Jewish influence on the West rapidly became dominant. Most of it came through the Pharisees, but such Jewish sects as the Essenes and Baptists were directly involved.

Even such a heterodox Jewish sect as the Samaritans exerted disproportionately great influence. In the first years of the nascent Christian Church a Samaritan diviner named Simon, later called "Magus," founded a new faith known as Gnosticism. The Gnostics spread rapidly over both the Roman and Iranian worlds, and by the end of the 3rd century AD they had subdivided into a multitude of different sects that covered a wide spectrum of possible combinations between Judaism, Christianity, Zoroastrianism, and Greco-Roman paganism.

In four centuries Christianity conquered the entire Roman Empire and many outlying regions, thanks to the intensity of its faith and the tenacity with which Christians held to their views, following Jewish models, through the bitterest persecutions.

In the East, Zoroastrianism maintained its hold over the Iranians and neighbouring peoples until the Islāmic conquest, when it was replaced by Islām, itself an offshoot of the monotheistic Judeo-Christian stem.

With the Hebrew Bible (which became the Christian Old Testament) almost the whole of traditional Israelite and early Jewish religion passed into Christianity, which was to a great extent an extension of Judaism. What was lost by the surrender of a large part of Jewish ritual law, with its stress on purity, was compensated by the triumph of monotheism over Greco-Roman polytheism and the transformation of ethical and spiritual ideas in the Greco-Roman world. Judaism itself survived to coexist with Christianity and Islām through all the intervening centuries to our own day.

Science and law. During the 3,000 years of urbanized life in Mesopotamia and Egypt tremendous strides were made in various branches of science and technology. The greatest advances were made in Mesopotamia—very possibly because of its constant shift of population and openness to foreign influence, in contrast to the relative isolation of Egypt and the consequent stability of its population. The Egyptians excelled in such applied sciences as medicine, engineering, and surveying; in Mesopotamia greater progress was made in astronomy and mathematics. The development of astronomy seems to have been greatly accelerated by that of astrology, which took the lead among the quasi-sciences involved in divination. The Egyptians remained far behind the Babylonians in developing astronomy, while Babylonian medicine, because of its chiefly magical character, was less advanced than that of Egypt. In engineering and architecture Egyptians took an early lead, owing largely to the stress they laid on the construction of such elaborate monuments as vast pyramids and temples of granite and sandstone. On the other hand, the Babylonians led in the development of such practical arts as irrigation.

Mesopotamian astronomy and mathematics

Both sciences and pseudosciences spread from Egypt and Mesopotamia to Phoenicia and Anatolia. The Phoenicians in particular transmitted much of this knowledge to the various lands of the Mediterranean, especially to the Greeks. The direction taken by these influences can be followed from Egypt to Syria, Phoenicia, and Cyprus, thanks to a combination of excavated art forms that prove the direction of movement, as well as to Greek tradition, which lays great stress on what the early Greek philosophers learned from Egypt. Mesopotamian influence can be traced especially through the partial borrowing of Babylonian science and divination by the Hittites and later by the transmission of information through Phoenicia. The Egyptians and Mesopotamians wrote no theoretical treatises; information had to be transmitted piecemeal through personal contacts.

The westward transmission of Babylonian mathematics was associated with that of law. All early Babylonian mathematics was transmitted in case form, introduced by a condition followed by its solution. This model appears first in late Sumerian law: from law it was extended to scientific problems, and the form remained the same until well into the 1st millennium BC. This was true also of such derived systems as Hittite law and the so-called covenant law of Israel, as well as the earliest Greek codes (Draco and Gortyn), all of which are formulated similarly: condition (protasis), secondary condition or conditions, and conclusion (apodosis). The Babylonian Code of Hammurabi provides an example: "If a man accuses another man and brings a charge of murder against him but does not prove it, the accuser shall be executed."

Logical thought

With the spread of case law in such conditional formulation, it was eventually discovered that formulation as generalized propositions or prohibitions was a simpler and more logical means of setting up a coherent code of laws than the case-law formulation (*e.g.*, in the Hebrew Decalogue). The next step was the formulation of generalized geometrical propositions. This was first done by the Ionian Greek Thales (about 600 BC), whose listing of

mathematical propositions in this generalized form instead of as conditional sentences was quite naturally described later as the "discovery" of mathematical theorems. With Thales logical reasoning made a giant step forward from the age of empirical modes of thought.

The Judeo-Christian concept of ethics and morals in law often prevailed in the Roman law of Christian times. Roman forms of law were ultimately adopted in almost the entire Western world, and through their universal sway many biblical approaches to legal problems became dominant.

The alphabet. Of all the accomplishments of the ancient Middle East, the invention of the alphabet is probably the greatest. While pre-alphabetic systems of writing in the Old World became steadily more phonetic, they were still exceedingly cumbersome, and the syllabic systems that gradually replaced them remained complex and difficult. In the early Hyksos period (17th century BC) the Northwestern Semites living in Egypt adapted hieroglyphic characters—in at least two slightly differing forms of letters—to their own purposes. Thus was developed the earliest known purely consonantal alphabet, imitated in northern Syria, with the addition of two letters to designate vowels used with the glottal catch.

This alphabet spread rapidly and was in quite common use among the Northwestern Semites (Canaanites, Hebrews, Aramaeans, and especially the Phoenicians) soon after its invention. By the 9th century BC the Phoenicians were using it in the western Mediterranean, and the Greeks and Phrygians adopted it in the 8th. The alphabet contributed vastly to the Greek cultural and literary revolution in the immediately following period. From the Greeks it was transmitted to other Western peoples. Since language must always remain the chief mode of communication for *Homo sapiens*, its union with hearing and vision in a uniquely simple phonetic structure has probably revolutionized civilization more than any other invention in history. (W.F.A.)

PREHISTORIC RELIGION

Practices and beliefs

BURIAL CUSTOMS AND CULTS OF THE DEAD

The oldest known burials can be attributed to the Middle Paleolithic Period. The corpses, accompanied by stone tools and parts of animals, were laid in holes in the ground and sometimes the corpses were especially protected. In some cases, the findings give the impression that the dead were to be "held onto." Whether or not that meant that the dead were to be cared for lovingly or that their return was to be feared, it implies, in any case, a belief in life after death in some form. But it is not necessary to infer a belief in separate souls; rather, it could also indicate the concept of a "living corpse."

Implication of life after death

From the Upper Paleolithic Period on, the burials manifest richer grave goods; however, it is not possible to conclude from this that religious concepts had changed. The same holds for the adoption of other burial practices, as, for example, secondary burials, in which the bodies were first allowed to decompose fully and then the bones were buried, or in the burning of bodies (evident from the Neolithic Period). From these facts it is not possible to infer the existence of a definite belief in souls; it is also not possible to determine the advent of such concepts from archaeological evidence. Even the increase in the discoveries of grave goods, occasionally also including other human remains, is evidence not for a change of religious concepts but for increased needs of the dead in the beyond—*i.e.*, needs after death that are dependent on economic and social status in life. Analogies to recent (primitive) phenomena demonstrate that it is not possible to connect particular burial customs with particular notions of the beyond, or to any other religious conceptions. Other than the burial of the whole body, the disposition of the individual parts of the body, and especially the skull, is important. Ritual deposition of skulls is confirmed for the Middle Paleolithic Period. From even earlier periods, however, individual or multiple human skulls and long bones have been found within a single site (for example, associated with Peking man). It is not necessary to interpret these findings as remains of headhunting or developed skull cults; for even today some simple hunting and gathering societies have the custom of preserving such parts of corpses for long periods of time and even of carrying them around on their bodies. The same practice is observed also to have occurred in the Upper Paleolithic and even later periods; but it is not possible to infer an elaborated ancestor cult directly from such prolonged connections of the living with the dead.

Burial of skulls and skull cults

The situation is different with findings from permanent settlements of agrarian people, in contrast to constantly shifting hunter-collectors. Evidences for ancestor cult practices dating to the 7th century BC were first discovered at Jericho in Palestine, where several skulls were found to have been deposited in a separate room, some of them covered with a plastic modelling of faces similar to that found on the ancestral skulls preserved by present-day agrarian peoples of South Asia and Oceania. An elaborated skull cult is usually connected with the veneration of ancestors. An important theme of ancestral cults is the belief in a connection between the dead and the fertility of the land of their descendants.

An especially noteworthy kind of burial is that of the megalithic (huge stone) graves that appear in various areas from the Neolithic Period on. It is probable that in this practice there was also a vital believed link between the living and the dead, and that occasionally sacred areas and gathering places were connected with such graves. The practices of the megalith builders were probably rooted, to a considerable extent, in ideas about the dead and in ancestor cults to which their stones gave a particular durability and a monumental form. It is more difficult to explain the individual erect stones (menhirs), which, of course, could be the symbol or seat of ancestors, especially where they show indications of being sculpted in human form. It surely would be a mistake, however, to look for a uniform interpretation of all megalithic monuments or even to speak of a distinct megalithic religion. The megalithic monuments are rather to be understood as a complex of grandiose manifestations of ideas that could well have been diverse, but among which the cult of the dead, nevertheless, played an important role.

Megalithic graves

CANNIBALISM

In finds belonging to the Paleolithic Period, pieces of human bodies as well as the bones of other animals are found scattered throughout the archaeological layers and are sometimes broken or charred. This is often taken as evidence for cannibalism, but other interpretations are just as likely (*e.g.*, the action of carrion-eating animals [such as hyenas] turning up the bones to the surface and thus causing their burning by later fires at the same place). To be sure, the finds allow the interpretation of cannibalism; however, they do not necessarily or intrinsically require it but rather permit that explanation if one proceeds from the prior conviction that cannibalism already existed at that time. This obsolete conception, still held by some scholars today—*i.e.*, that cannibalism is an especially "primitive" phenomenon and therefore very ancient—must be abandoned. Ethnological studies show clearly that cannibalism appears almost exclusively in the practices of agrarian peoples, that is, in a later cultural stage, and evidently is essentially bound up with religious or magical conceptions in which cultivated plants play a large role. Even if a Paleolithic cannibalism existed on a large scale, it could not be explained by means of concepts that originated in a cultural stage so differently structured.

The situation in later periods, especially in the Neolithic, is different. Here, rather than isolated parts of human skeletons scattered about a settlement, human remains

occasionally are found in association with remains of foodstuffs in waste pits or in holes and tunnels that served as sacrificial sites. Especially where human skulls have been broken open and the hollow bones split, the interpretation of cannibalism is unavoidable. Since this inferred practice occurred in the realm of agrarian cultures, it is more feasible to make comparisons with present-day cannibalism, where the meaning is generally the acquisition of the powers and other qualities of the victim.

SACRIFICES

Sacrifices (*i.e.*, the presentation of offerings to higher beings or to the dead) appear as early as the Middle Paleolithic Period. Pits with some animal bones have been found in the vicinity of burial sites; thus, it is a likely possibility that they represent offerings to the dead. There is a dispute over the interpretation of the arrangement of the skulls and long bones of bears, since they are deposited in such a manner that it is hardly possible to discern a profane explanation. It is assumed that they had a cultic or magical significance. Most likely, certain parts of the prey, such as the head and the meaty shanks, or at least the bones with brain and marrow, were sacrificed. Even if it cannot be definitely stated who the recipient of these sacrifices was, analogies with present-day "primitive" phenomena make it likely that a part of the prey was offered to a higher being who was believed to dispense nourishment. It could also, however, have been a matter of preserving parts of animals in order to resurrect the entire animal and preserve the species. Furthermore, finds of bones and drawings show that the preservation of skulls with still attached vertebrae, ribs, and front legs of oxen and reindeer played a certain religious or magical role. The sinking of whole reindeer into lakes is hard to explain other than as a sacrifice. This might be traced to the idea that what occupies the centre of attention is not the individual hunted animal but the whole herd; no longer only a part of an animal but a whole animal as part of a herd is sacrificed. The custom also existed in recent times among hunters and herders of central and north Asia. As such finds become more numerous, it seems evident that certain specific animals and parts of their bodies are selected for sacrifice.

Cultic and magical significance of sacrifices

It is difficult to differentiate between animal sacrifices and the immediate cultic veneration of an animal at the burial sites of animals. In the Neolithic Period, the sites become especially profuse and are usually found in connection with human burials; nevertheless, there are such burial sites of animals that are not related in this manner and that occur with pronounced frequency, characteristically in particular groups of cultures. In these cases, domestic animals almost exclusively are involved, and among them the dog and the ox predominate.

The question of human sacrifice is of special significance here. Human sacrifices often were related to cannibalism and to the sacrifice of animals. With conspicuous frequency victims discerned in ceremonial remains are females and children, sometimes along with young pigs. This practice is similar to fertility and agricultural rites that are known to have been practiced in the early Mediterranean civilizations. It is also similar to beliefs and practices observed among present-day "primitive" agrarian peoples (in which pigs are often substituted for humans), such as in ceremonies of secret societies, initiation rites, sacrifices, celebrations of feasts of the dead, and notions about fertility, especially in connection with the growing and ripening of cultivated plants.

In comparison, the inclusion of servants or women in the burial sites of highly placed persons can hardly be called sacrifice in a strict sense—that is, an offering to a higher power or deity. Such inclusions most likely reflect the social status of the deceased leader and his need for servants in the afterlife, rather than an offering. It is a sacrifice in the wider sense of respect and awe for the person and status—and all that this conveyed—of the deceased leader. This practice becomes more important only where correspondingly differentiated social conditions are found (such as in the royal graves at Ur in Mesopotamia and in those of the Shang dynasty in China). Sometimes it took on almost unbelievable forms, especially in terms of the numbers of persons and animals interred with the deceased leader.

The ritual preservation of objects also must be included in the realm of sacrifice (in a wider sense). This can be demonstrated for the first time in the Neolithic Period (for instance, the ritual depositing of axes); in later periods, it plays a large role. In finds from the Bronze Age on, weapons and jewelry frequently are found in wells and springs. In Iron Age finds, such objects are found in almost unbelievable quantities in a number of swamps and other bodies of water. It seems probable that they represent the sacrifice of war booty.

Ritual preservation of objects

HUNTING RITES AND ANIMAL CULTS

In the oldest known examples of graphic art, the representations of animals play a large part; humans appear rarely and then frequently with animal attributes or as mixed human–animal figures. In the context of the whole situation, the view that these representations were merely ornamentations or served a purely artistic need may be dismissed; they are found without boundaries and background on rock walls and are not part of an interrelated scene. It is evident that animals played a predominant role in the mental world of the Upper Paleolithic Period insofar as this role is reflected in the art of the period. What is represented is, first of all, that which is essential to the animal, partly in its relation to the hunt, but also in relation to anthropomorphic figures showing the intermixing of human and animal forms. This indicates a special and intimate relationship between humans and animals that transcends and overcomes the boundaries between different realms of being that modern concepts and understanding require.

This phenomenon is similar to what is still known today as animalism (or nagualism or theriocentrism). It is characterized by close magical and religious ties of humans with animals, especially with wild animals. It is also characterized in terms of otherworldly and superworldly realms and practices, such as placating and begging for forgiveness of the game killed, performing oracles with animal bones, and performing mimic animal dances and fertility rites for animals. Animals were thought to be manlike, to have souls, or to be equipped with magical powers. Animalism thus expresses itself in various conceptions of how animals are regarded as guardian spirits and "alter egos," of the facile and frequent interchangeability between human and animal forms, and also of a theriomorphically (animal-formed) envisioned higher being—one who changes between human and animal forms and unifies them. Higher, often theriomorphic, beings are gods who rule over the animals, the hunters, and the hunting territory, or spirits in the bushland and with the animals. It is obviously not possible to identify special occurrences or forms of such higher beings during the Paleolithic Period, but their general features may be safely assumed.

Animalism is, to a large extent, a basis for totemism, which involves various permanent relationships of individuals or groups to certain animals or other natural objects; hence animalism is occasionally called "protototemism." Individual and cultic totemism, as opposed to group totemism of an almost solely social function, are particularly close to animalism, whereas religious and cultic meanings in group or clan totemism are usually poorly developed. It is not possible to determine to what extent animalism had already assumed the character of true totemism in the Paleolithic Period; the early existence of clan totemism is improbable because it occurs primarily among peoples who are to some extent agrarian, and possibly a certain kind of sedentary life was prerequisite to its development.

Totemism

Also, special sacrificial traditions were closely connected to game, particularly the custom of preserving the animal skeleton or a part of a skeleton in order to placate the ruler of the animals (see above) and to provide for continuation of the species.

A certain kind of bear ceremonialism is rooted in this conception and is to be recognized in several finds and pictures from the Upper Paleolithic Period on. A skin with attached head was evidently draped over the body of



Cave bear skull, found near Petershöhle, Germany, probably 50,000–30,000 BC, suggesting an offering of this part of the prey.

By courtesy of the Naturhistorische Gesellschaft, Nurnberg, Germany; photograph, Georg Pandura

a bear made out of clay; the skull and long bones of the bear were buried separately (a practice begun in the Middle Paleolithic Period); the bear was shot with arrows and killed by a shot or a thrust into the lungs; the animal or a bearlike figure was surrounded by dancers. Similar phenomena are documented for more recent periods, above all for the hunting cultures of Neolithic Siberia. These observations can be effortlessly fitted into the practice of bear ceremonialism that is still widely distributed in northern Eurasia and North America.

The question of whether animals were the immediate objects of a cult is extremely difficult to judge in each particular case. Nevertheless, with the beginning of the Neolithic Period, animal phenomena appear that probably go beyond functioning merely as a sacrifice and symbol. This applies especially to representations of oxen and bulls and to the symbolism of bull heads and bull horns.

FEMALE FERTILITY DEITIES

Fertility statuettes

Small female figures, the so-called Venus statuettes, appear for the first time in the Upper Paleolithic Period. In some cases they are very schematically formed, and it is often difficult or impossible to recognize female attributes. In other cases, however, they are naturalistic representations of corpulent women whose secondary sexual characteristics (their breasts and buttocks) were given special prominence, though their faces, feet, and arms were almost completely neglected. Such strong emphasis on the anatomical zones that are related to the bearing of children and nourishing them easily conveys to one the idea of female fertility. Nevertheless, it is not necessarily true of all these small figures.

Dortes—B. Arthaud



Female holding a bison horn, stone bas-relief from Laussel, Dordogne, France. Gravettian period (25,000/20,000–18,000/15,000 BC). In the Musée d'Aquitaine, Bordeaux, France. Height of figure 37.72 cm.

Ethnological analogies with present-day primitive phenomena offer the equally plausible view that such figures were regarded as the representations of the abodes of spirits whose function was to help and protect, and especially during hunting. They also may have been conceived, among other things, as mothers or rulers of the animals, goddesses of the underworld, helpers during hunting and donators of game, and as sovereigns of the land and other regions and of natural forces, including that of fertility.

No known direct continuum connects these earlier Paleolithic figures to similar ones of the early Neolithic and later periods. In settlements and shrines of these later periods are found large numbers of female figurines of widely differing types. They may have been representations of deities and symbols or, perhaps, votive offerings, somehow connected with female fertility. This can be safely assumed for figurines that show an obvious indication of fertility or are connected with children, and even more for shrines containing figures with sculptured pairs of breasts, and figures on the walls of women in childbirth. Not all female figures can, however, be understood merely as fertility symbols; rather, in many cases they are assumed to be house gods or representations of ancestors, and, especially when appearing in graves, as substitutes for the bodies of maids, wives, and concubines. An appearance of a large number of smaller figures suggests a votive or magical usage.

SHAMANISM, SORCERY, AND MAGIC

Belief in guardian spirits

Shamanism is a rather variable and highly stratified complex of practices and conceptions; characteristic among these are the use of ecstasy, the belief in guardian spirits (who are often in animal form, with the function of helping and guiding the dead on their voyage to the beyond), and beliefs concerning metamorphosis (change of form) and travelling to the beyond. Pictures from the Upper Paleolithic Period indicate the existence at that time of ecstatic practices and of beliefs in protective and helping spirits, which assume the form of birds and other animals. On the other hand, it is doubtful whether shamanism existed in fully developed form at that time. Also, in the course of prehistory, objects appear that may well have belonged to the paraphernalia of shamanism. Noisemaking objects (to drive away evil spirits) are often found in the material remains of the Iron Age and probably are connected with shamanism.

Recent studies stress the religious character of shamanism, though in practice it is related to sorcery and magic. Shamanism is not to be identified with sorcery and magic if they are understood as attempts to manipulate the supernatural through certain human techniques, in contrast with religion, in which man approaches higher beings (gods) in an attitude of supplication. Magic or sorcery thus appears as the opponent of true religion and gains importance when religion declines or is overwhelmed. In fact, magic and sorcery may take over cultic forms and rob them of their religious meaning when this occurs. For these reasons, it is often difficult to decide whether prehistoric phenomena were of religious or magical character.

Magic also can be practiced to a large degree without the use of material objects, and it is, therefore, as hard to grasp archaeologically as true religion. In the interpretation of the art of the Upper Paleolithic, scholars have given great importance to magic because, for example, missiles (spears and arrows) were drawn on the pictures of animals. This has been interpreted to mean that an effort was made to ensure and compel the success of hunters through magical action. But this interpretation is highly speculative, and it remains uncertain what these drawings mean. It is just

as difficult to decide whether or not other pictures, sculptures, abstract symbols, amulets, and similar objects were used to make magic in this and later periods.

Evolutionary development

Religion is always closely related to other realms of life, such as economic activities. These relations are partly direct and partly mediated by social forms. The latter are, on the one hand, at least partially dependent on economic conditions; on the other hand, social structures influence the formation of religious phenomena and often serve as models for their elaboration. In a negative sense, then, it is often possible to eliminate certain religious phenomena as inappropriate to a particular society. It is inconceivable, for example, that the religious conception of simple hunters and gatherers included an elaborately organized hierarchy of gods with detailed division of labour between the individual figures. Similarly, it is a mistake to attribute to hunters and gatherers conceptions that are bound up with agriculture and the fertility of fields. In a positive sense, however, certain economic and social conditions will encourage the development of certain corresponding religious conceptions. Animalistic notions will be especially effective in situations where animals play a large role as partners of humans. Nevertheless, the spiritual ties to animals will be considerably different among hunters or agrarian peoples who still find it necessary to rely heavily on hunting for their meat supply as compared with pastoral peoples. In fully agrarian cultures, on the other hand, ideas about the fertility of fields and cultivated plants play an important part; they are connected with other notions about fertility and influence other spheres of life.

STONE AGE CULTURES

Burial sites that indicate a belief in life after death

Lower or Early and Middle Paleolithic. The oldest burials that attest to a belief in life after death can be placed in the period between about 50,000 and 30,000 BC. The earliest evidence of human activity in any form, on the other hand, goes back more than 1,000,000 years. Yet, since religious conceptions are not always bound to material objects, and since there is evidence that truly human beings existed even during early Paleolithic times, it is inadmissible to infer that earliest man had no religion from the mere fact that no identifiable religious objects have been found.

A study of very simple hunters and gatherers of recent times shows that several religious conceptions generally considered to be especially "primitive" (*e.g.,* fetishism) hardly play an important part, but rather that, among other things, the supposedly "advanced" conception of a personal creator and preserver of the world does play an important part. Such a belief could never be discovered by examining archaeological sources—the material remains—and hence cannot be ruled out for the Early Paleolithic Period. Whether or not the sacrifices in that era involved divine creators or preservers or other beings can only be a matter of conjecture. Features of animalism, magic, and various other views and practices may have played a role, but probably less so than in later epochs.

Upper Paleolithic and Mesolithic. The animalistic features encountered in the art of the Upper Paleolithic Period were most likely only a part of the religion that existed at that time. Among present-day "primitives" the animalistic realm often occupies only a lower sphere of what can be considered religious, and beyond and above that sphere are still other notions about gods. Practices concerning the resurrection of animals and the preservation of species evidently also played an important part and were closely tied to animalistic conceptions. The corresponding rituals clearly took on a special significance in relation to bears and became the basis for the bear ceremonies that were later widely diffused. Although shamanism may have been initiated somewhat earlier, it was now evident, at least in some of its aspects.

The realm of hunting was primarily a masculine sphere; nevertheless, it also includes in religious phenomena the feminine aspect, as symbols of female fertility (and probably also of female deities) demonstrate.

Proto-Neolithic and Neolithic. The characteristics of early religion were continued but transformed in the proto-Neolithic and Neolithic periods. Shamanism developed, especially among the pastoralists of central and northern Asia. Animals, viewed as the hypostases (essences) of higher beings, especially the eagle or falcon and the raven, became highly significant in shamanism. Animalistic conceptions continued and often assumed the proportions of a true animal cult. Hoofed animals, especially sheep and oxen, played an important part as sacrifices, and bulls particularly assumed a leading role; they seem to have been relegated to the masculine sphere. Horses appear as domesticated animals and as sacrifices only toward the end of the Neolithic Period. They may have been connected with a heavenly divinity, as later evidence suggests.

In the early period of agriculture, before the full development of the Neolithic Period, deposits of human skulls appear that suggest the presence of ancestor cults. A spiritual identification between humans and plants apparently played a predominant part in conceptions connected with headhunting and cannibalism. The death of a god was often considered a prerequisite to the appearance and prospering of the plants, and this mythical event was repeated through human sacrifice that was either accompanied by or replaced by animal sacrifice.

Dualism of masculine and feminine elements

At an early stage, in addition to an agricultural connection with the earlier feminine aspects, the masculine aspect appears in the form of portrayals of sexual union and, perhaps, of the "holy wedding," or sacred coupling, as well as in portrayals of couples and families. Among the material remains, however, the direct representation of the male element recedes sharply, yet perhaps the symbol of the axe and probably also that of the bull may indicate the male element. This dualism of the masculine and feminine aspects can possibly be interpreted in terms of father sky and mother earth, and in their union as a couple by which they become parents of the world. In the early civilizations, the conception of a supreme being or a heavenly god (which cannot clearly be recognized either in pictures or in other material objects) plays a minor role. That does not mean, however, that such a conception is necessarily of recent origin but rather that it probably existed at an early period in places where there was no literate tradition (predominantly among pastoral cultures).

CIVILIZATIONS

The decisive factors that brought about the early civilizations were the new kinds of economic and social organization, the large-scale exploitation of human energy, the formation of ruling classes, hierarchical organization, and the administrative division of labour. Under such conditions polytheism, which had undoubtedly been nascent before, could develop fully. The social order is mirrored in the conception of city and state gods and of a hierarchically organized "state of gods" with a division of labour. The concentration of power and people in one place, in contrast with the wandering of earlier nomadic cultures, enabled fixed central shrines to become influential. Yet the old traditions continued, and not least among them, that of animalism, in the form of conceptions about a ruler of the animals, animal cults, and similar phenomena. Female fertility figures remain generally prominent, such as the Great Mother and the Earth Mother. (K.J.N./Ed.)

BIBLIOGRAPHY

General works. IAN TATTERSALL, ERIC DELSON, and JOHN VAN COUVERING (eds.), *Encyclopedia of Human Evolution and Prehistory* (1988), is a comprehensive reference source. ROBERT J. WENKE, *Patterns in Prehistory*, 3rd ed. (1990), studies 3,000,000 years of cultural, agricultural, and urban development. JOHN GOWLETT, *Ascent to Civilization: The Archaeology of Early Humans*, 2nd ed. (1992), examines archaeological evidence and what it can tell us about the human past. BRIAN M. FAGAN, *People of the Earth: An Introduction to World Prehistory*, 8th ed. (1995), a survey, includes coverage of the transition from hunting and gathering to farming and the development of cities and civilization. Two popular works are DEREK ROE, *Prehistory: An Introduction* (1970), emphasizing the prehistory of Great Britain; and ROBERT J. BRAIDWOOD, *Prehistoric Men*, 8th ed. (1975), an introduction to many aspects of prehistoric life. NEIL ROBERTS, *The Holocene* (1989), examines the changes, both

environmental and human, that have occurred over the past 10,000 years.

Cultural history of the Stone Ages. A sampling of the many regional studies includes, on Africa, C. GARTH SAMPSON, *The Stone Age Archaeology of Southern Africa* (1974), an archaeological sourcebook surveying 2,000,000 years of human prehistory; on Asia, ROBERT STIGLER (ed.), *The Old World: Early Man to the Development of Agriculture* (1974), a brief introduction to Paleolithic culture and the beginnings of the major civilizations in South Asia and the Middle East; on Europe, SARUNAS MILISAUSKAS, *European Prehistory* (1978), an anthropological treatment from the first settlements to the Roman Empire, tracing economies, settlements, social organization, trade, and ideology in the Neolithic and subsequent periods; TIMOTHY CHAMPION *et al., Prehistoric Europe* (1984), a comprehensive introduction; and BARRY CUNLIFFE (ed.), *The Oxford Illustrated Prehistory of Europe* (1994), through the decline of the Roman Empire; on the Americas, SHIRLEY GORENSTEIN (ed.), *Prehispanic America* (1974), a discussion of Paleo-Americans, Meso-Americans, and the rise of civilization in South America; ROBERT F. SPENCER *et al., The Native Americans,* 2nd ed. (1977), a scholarly study of traditional North American Indian cultures; JESSE D. JENNINGS (ed.), *Ancient North Americans* (1983), and *Ancient South Americans* (1983); BRIAN M. FAGAN, *The Great Journey: The Peopling of Ancient America* (1987), and *Ancient North America* (1991), for the general reader; and JESSE D. JENNINGS, *Prehistory of North America,* 3rd ed. (1989), a survey of the earliest cultures of North America; and on Oceania, J. ALLEN, J. GOLSON, and R. JONES (eds.), *Sunda and Sahul: Prehistoric Studies in Southeast Asia, Melanesia, and Australia* (1977), essays exploring biology, agriculture, ethnography, biogeography, and other aspects; J. PETER WHITE, *A Prehistory of Australia, New Guinea, and Sahul* (1982), a scholarly overview; and JOSEPHINE FLOOD, *Archaeology of the Dreamtime: The Story of Prehistoric Australia and Its People,* rev. ed. (1990), for the general reader.

Civilizations. CARROLL L. RILEY, *The Origins of Civilization* (1969), for the general reader, surveys the development of 11 civilizations, including those in North China, the Americas, the western Mediterranean, and others from Greece to the Indus Valley. C.D. DARLINGTON, *The Evolution of Man and Society* (1969), gives a biological interpretation of the growth and spread of culture and civilization from the Neolithic Period to the 20th century. *The Cambridge Ancient History,* vol. 1 and 2, 3rd ed. (1970–75), and vol. 3, part 1, 2nd ed. (1982), surveys prehistory in Europe, western Asia, and North Africa and studies the rise of civilizations in the Middle East and the Aegean. JOHN E. PFEIFFER, *The Emergence of Society: A Pre-history of the Establishment* (1977), offers a popular account of society's evolution during the past 10,000 years. RUTH WHITEHOUSE and JOHN WILKINS, *The Making of Civilization* (1986), a richly illustrated work, focuses on archaeological evidence.

Works treating the rise of civilization in the Middle East specifically include WALTHER HINZ, *The Lost World of Elam: Re-Creation of a Vanished Civilization* (1972; originally published in German, 1964), an account of art, religion, and social mores in the Elamite empire; CHESTER G. STARR, *Early Man: Prehistory and the Civilizations of the Ancient Near East* (1973), a popular overview of the development of literacy, urban living, art, literature, and religion; CHARLES L. REDMAN, *The Rise of Civilization: From Early Farmers to Urban Society in the Ancient Near East* (1978), on agricultural development, urbanization, and cultural change from 10,000 to 2000 BC; JACK FINEGAN, *Archaeological History of the Ancient Middle East* (1979, reissued 1986), a reference work covering history, mythology, and art between 10,000 and 330 BC; and DONALD O. HENRY, *From Foraging to Agriculture* (1989), a discussion of the origins of agriculture in the Middle East and surrounding area.

Various aspects of early civilizations are considered in E.O. JAMES, *Prehistoric Religion: A Study in Prehistoric Archaeology* (1957, reissued 1963), a collection of materials on the Paleolithic through the Neolithic periods; GRAHAME CLARK, *Aspects of Prehistory* (1970), emphasizing the recency of literacy and urban life and the exegetical worth of natural selection in understanding cultural development; THEODORE A. WERTIME and JAMES D. MUHLY (eds.), *The Coming of the Age of Iron* (1980), a scholarly study of metallurgy and the origins of ironworking in several cultures; JOHN E. PFEIFFER, *The Creative Explosion: An Inquiry into the Origins of Art and Religion* (1982), for the general reader; and MARGARET EHRENBERG, *Women in Prehistory* (1989), an examination of archaeological evidence for the roles of women from the Neolithic to the Roman period in Europe. Science in particular is the focus of JOSEPH NEEDHAM, *Science and Civilisation in China* (1954–), a detailed study of the development of science in China and other Asian nations; DEBIPRASAD CHATTOPADHYAYA, *Science and Society in Ancient India* (1977), which suggests that the growth of science was furthered most by an interest in medicine; and G.E.R. LLOYD, *Magic, Reason, and Experience: Studies in the Origin and Development of Greek Science* (1979), with emphasis on mathematics, astronomy, and medicine. Prehistoric art is treated in H. BREUIL, *Four Hundred Centuries of Cave Art* (1952, reprinted 1979; originally published in French, 1952), a study by a pioneer in the field; ANN POWELL, *The Origins of Western Art* (1973), a survey from Paleolithic times to the Roman Empire, covering Mesopotamia, Egypt, and the Aegean; ANDRÉ LEROI-GOURHAN, *The Dawn of European Art: An Introduction to Palaeolithic Cave Painting* (1982; originally published in Italian, 1980), a well-illustrated technical discussion; A.R. WILLCOX, *The Rock Art of Africa* (1984), a comprehensive survey; and RICHARD F. TOWNSEND (ed.), *The Ancient Americas: Art from Sacred Landscapes* (1992), a book prepared to accompany a major exhibition of works from 12 cultures. WILFRID BONSER, *A Prehistoric Bibliography,* extended and ed. by JUNE TROY (1976), includes 9,000 books and periodical articles focusing on Great Britain. (Ed.)

Printing, Typography, and Photoengraving

Printing traditionally has been defined as a technique for applying under pressure a certain quantity of colouring agent onto a specified surface to form a body of text or an illustration. Certain recently developed processes for reproducing texts and illustrations, however, are no longer dependent on the mechanical concept of pressure or even on the material concept of colouring agent. Because these processes represent an important development that may ultimately replace the other processes, printing should probably now be defined as any of several techniques for reproducing texts and illustrations, in black and in colour, on a durable surface and in a desired number of identical copies. There is no reason why this broad definition should not be retained, for the whole history of printing is a progression away from those things that originally characterized it: lead, ink, and the press.

It is also true that, after five centuries during which printing has maintained a quasi-monopoly of the transmission or storage of information, this role is being seriously challenged by new audiovisual and information media. Printing, by the very magnitude of its contribution to the multiplication of knowledge, has helped engender radio, television, film, microfilm, tape recording, and other rival techniques. Nevertheless, its own field remains immense. Printing is used not merely for books and newspapers but also for textiles, plates, wallpaper, packaging, and billboards. It has even been used to manufacture miniature electronic circuits.

The invention of printing at the dawn of the age of the great discoveries was in part a response and in part a stimulus to the movement that, by transforming the economic, social, and ideological relations of civilization, would usher in the modern world. The economic world was marked by the high level of production and exchange attained by the Italian republics, as well as by the commercial upsurge of the Hanseatic League and the Flemish cities; social relations were marked by the decline of the landed aristocracy and the rise of the urban mercantile bourgeoisie; and the world of ideas reflected the aspirations of this bourgeoisie for a political role that would allow it to fulfill its economic ambitions. Ideas were affected by the religious crisis that would lead to the Protestant Reformation.

The first major role of the printed book was to spread literacy and then general knowledge among the new economic powers of society. In the beginning it was scorned by the princes. It is significant that the contents of the first books were often devoted to literary and scientific works as well as to religious texts, though printing was used to ensure the broad dissemination of religious material, first Catholic and, shortly, Protestant.

There is a material explanation for the fact that printing developed in Europe in the 15th century rather than in the Far East, even though the principle on which it is based had been known in the Orient long before. European writing was based on an alphabet composed of a limited number of abstract symbols. This simplifies the problems involved in developing techniques for the use of movable type manufactured in series. Chinese handwriting, with its vast number of ideograms requiring some 80,000 symbols, lends itself only poorly to the requirements of a typography. Partly for this reason, the unquestionably advanced Oriental civilization, of which the richness of their writing was evidence, underwent a slowing down of its evolution in comparison with the formerly more backward Western civilizations.

Printing participated in and gave impetus to the growth and accumulation of knowledge. In each succeeding era there were more people who were able to assimilate the knowledge handed to them and to augment it with their own contribution. From Diderot's encyclopaedia to the present profusion of publications printed throughout the world, there has been a constant acceleration of change, a process highlighted by the Industrial Revolution at the beginning of the 19th century and the scientific and technical revolution of the 20th.

At the same time, printing has facilitated the spread of ideas that have helped to shape alterations in social relations made possible by industrial development and economic transformations. By means of books, pamphlets, and the press, information of all kinds has reached all levels of society in most countries.

In view of the contemporary competition over some of its traditional functions, it has been suggested by some observers that printing is destined to disappear. On the other hand, this point of view has been condemned as unrealistic by those who argue that information in printed form offers particular advantages different from those of other audio or visual media. Radio scripts and television pictures report facts immediately but only fleetingly, while printed texts and documents, though they require a longer time to be produced, are permanently available and so permit reflection. Though films, microfilms, punch cards, punch tapes, tape recordings, holograms, and other devices preserve a large volume of information in small space, the information on them is available to human senses only through apparatus such as enlargers, readers, and amplifiers. Print, on the other hand, is directly accessible, a fact that may explain why the most common accessory to electronic calculators is a mechanism to print out the results of their operations in plain language. Far from being fated to disappear, printing seems more likely to experience an evolution marked by its increasingly close association with these various other means by which information is placed at the disposal of humankind.

For coverage of related topics in the *Macropædia* and the *Micropædia*, see the *Propædia*, section 629, and the *Index*.

The article is divided into the following sections:

History of printing 72
  Origins in China 72
    Invention of movable type (11th century)
    Transmission of paper to Europe (12th century)
  The invention of printing 72
    Xylography
    Metallographic printing (1430?)
    The invention of typography—Gutenberg (1450?)
    The Gutenberg press
  Improvements after Gutenberg 74
    The metal press (1795)
    Stereotypy and stereography (late 18th century)
    Koenig's mechanical press (early 19th century)
    Attempts to mechanize composition (mid-19th century)
    Typecasting compositors (1880s)
  19th-century innovations 75
    Reproduction of illustrations
    Lithography: Senefelder (1796)
    Photosensitivity: Niepce (1820s)
    Gravure and rotogravure (1890s)
  The 20th century 76
    Discovery of offset (early 20th century)
    Dry offset (1920)
    Colour printing
    Automation of composition (after 1929)

Programmed composition (1950s)
Photocomposition
Toward direct impression
Serigraphy and collotype: a renaissance
Three-dimensional printing (1960s)
Office printing
Modern printing techniques 78
Composition and typesetting 78
Mechanical composition and typesetting
Phototypesetting
Makeup of letterpress copy
Printing (press operation) 85
Colour printing
Letterpress printing
Rotogravure
Offset
Other printing processes
Printing inks
Typography 92
The nature of typography 92
Typography as a useful art
Aesthetic qualities of the typographic page
History of typography 95
Type, from Gutenberg to the 18th century
Type and book design since the 19th century
Photoengraving 105
History of photoengraving 105
Early etched plates
Wet-collodion photography
The halftone process
Basis for selection of screen ruling
Contact screens
The benday process
Special effects
Process developments
Chemical etching—traditional and powderless processes
Electromechanical plate making
Colour scanners
Modern photoengraving techniques 107
Basic production processes
Colourplate production
Production specifications
Engraving techniques applied to intaglio processes
Bibliography 110

History of printing

ORIGINS IN CHINA

By the end of the 2nd century AD, the Chinese apparently had discovered printing; certainly they then had at their disposal the three elements necessary for printing: (1) paper, the techniques for the manufacture of which they had known for several decades; (2) ink, whose basic formula they had known for 25 centuries; and (3) surfaces bearing texts carved in relief. Some of the texts were classics of Buddhist thought inscribed on marble pillars, to which pilgrims applied sheets of damp paper, daubing the surface with ink so that the parts that stood out in relief showed up; some were religious seals used to transfer pictures and texts of prayers to paper. It was probably this use of seals that led in the 4th or 5th century to the development of ink of a good consistency for printing.

A substitute for these two kinds of surfaces, the marble pillars and the seals, that was more practical with regard both to manageability and to size, appeared perhaps by the 6th century in the wood block. First, the text was written in ink on a sheet of fine paper; then the written side of the sheet was applied to the smooth surface of a block of wood, coated with a rice paste that retained the ink of the text; third, an engraver cut away the uninked areas so that the text stood out in relief and in reverse.

To make a print, the wood block was inked with a paintbrush, a sheet of paper spread on it, and the back of the sheet rubbed with a brush. Only one side of the sheet could be printed.

Early wood-block works

The oldest known printed works were made by this technique: in Japan about 764–770, Buddhist incantations ordered by Empress Shōtoku; in China in 868, the first known book, the *Diamond Sūtra*; and, beginning in 932, a collection of Chinese classics in 130 volumes, at the initiative of Fong Tao, a Chinese minister.

Invention of movable type (11th century). About 1041–48 a Chinese alchemist named Pi Sheng appears to have conceived of movable type made of an amalgam of clay and glue hardened by baking. He composed texts by placing the types side by side on an iron plate coated with a mixture of resin, wax, and paper ash. Gently heating this plate and then letting the plate cool solidified the type. Once the impression had been made, the type could be detached by reheating the plate. It would thus appear that Pi Sheng had found an overall solution to the many problems of typography: the manufacture, the assembling, and the recovery of indefinitely reusable type.

In about 1313 a magistrate named Wang Chen seems to have had a craftsman carve more than 60,000 characters on movable wooden blocks so that a treatise on the history of technology could be published. To him is also attributed the invention of horizontal compartmented cases that revolved about a vertical axis to permit easier handling of the type. But Wang Chen's innovation, like that of Pi Sheng, was not followed up in China.

In Korea, on the contrary, typography, which had appeared by the first half of the 13th century, was extensively developed under the stimulus of King Htai Tjong, who, in 1403, ordered the first set of 100,000 pieces of type to be cast in bronze. Nine other fonts followed from then to 1516; two of them were made in 1420 and 1434, before Europe in its turn discovered typography.

Transmission of paper to Europe (12th century). Paper, the production of which was known only to the Chinese, followed the caravan routes of Central Asia to the markets at Samarkand, whence it was distributed as a commodity across the entire Arab world.

The transmission of the techniques of papermaking appears to have followed the same route; Chinese taken prisoner at the Battle of Talas, near Samarkand, in 751 gave the secret to the Arabs. Paper mills proliferated from the end of the 8th century to the 13th century, from Baghdad and then on to Spain, then under Arab domination. Paper first penetrated Europe as a commodity from the 12th century onward through Italian ports that had active commercial relations with the Arab world and also, doubtless, by the overland route from Spain to France. Papermaking techniques apparently were rediscovered by Europeans through an examination of the material from which the imported commodity was made; possibly the secret was brought back in the mid-13th century by returning crusaders or merchants in the Eastern trade. Papermaking centres grew up in Italy after 1275 and in France and Germany in the course of the 14th century.

But knowledge of the typographic process does not seem to have succeeded, as papermaking techniques had, in reaching Europe from China. It would seem that typography was assimilated by the Uighurs who lived on the borders of Mongolia and Turkistan, since a set of Uighur typefaces, carved on wooden cubes, has been found that date from the early 14th century. It would be surprising if the Uighurs, a nomadic people usually considered to have been the educators of other Turco-Mongolian peoples, had not spread the knowledge of typography as far as Egypt. There it may have encountered an obstacle to its progress toward Europe, namely, that, even though the Islāmic religion had accepted paper in order to record the word of Allah, it may have refused to permit the word of Allah to be reproduced by artificial means.

Contribution of the Uighurs

THE INVENTION OF PRINTING

Thus, the essential elements of the printing process collected slowly in western Europe, where a favourable cultural and economic climate had formed.

Xylography. Xylography, the art of printing from wood carving, the existence and importance of which in China was never suspected by Marco Polo, appeared in Europe no earlier than the last quarter of the 14th century, spontaneously and presumably as a result of the use of paper. It had been observed that paper was better suited than rough-surfaced parchment for making the impres-

sions from wood reliefs that manuscript copyists used to reproduce the outline of ornamental initial capital letters.

The process was extended to the making of religious pictures. These at first appeared alone and later were accompanied by a brief text. As engravers became more skillful, the text finally became more important than the illustration, and in the first half of the 15th century small, genuine books of several pages, religious works or compendiums of Latin grammar by Aelius Donatus and called *donats,* were published by a method identical to that of the Chinese. Given the Western alphabet, it would seem reasonable that the next step taken might have been to carve blocks of writing that, instead of texts, would simply contain a large number of letters of the alphabet; such blocks could then be cut up into type, usable and reusable.

It is possible that experiments were in fact made along these lines, perhaps in 1423 or 1437 by a Dutchman from Haarlem, Laurens Janszoon, known as Coster. The encouraging results obtained with large type demonstrated the validity of the idea of typographic composition.

But the results were disappointing with regard to type destined for use for text of the usual size. The letters of the roman alphabet were smaller than Chinese ideograms, and cutting them from wood was a delicate operation. Moreover, type made in this way was fragile, and it wore out at least as quickly as blocks carved with a whole text. Further, since the letters were individually carved, no two copies of the same letter were identical any more than when the text was engraved directly on a wood block. The process, thus, represented no advance in ease of production, durability, or quality.

Metallographic printing (1430?). Metallographic impression is more likely to turn out to be the direct ancestor of typography, although the record is far from clear. Several medieval craft guilds, notably the metal founders, the die-cutters, and goldsmiths and silversmiths, were familiar with the technique of using dies. Masters of this technique apparently realized that it could be applied to a process that would enable texts to be set in relief more quickly than by carving wood blocks, probably in three steps: (1) a set of dies, each bearing a letter of the alphabet, was engraved in brass or bronze; (2) using these dies, the text was struck letter by letter to form a mold on the surface of a matrix of clay or of a soft metal such as lead; (3) lead was then poured over the surface to form a small plate that, once hardened, would bear the text in relief.

Steps in metal relief printing

The theoretical advantages of this process were that only one engraving per letter, that of the die, was required to make the letter as often as desired, and any two examples of the same letter would be identical, since they came from a single die; sinking the matrix and casting the lead were rapid operations; the lead had better durability than wood; and by casting several plates from the same matrix the number of copies printed could be rapidly increased.

Metallographic printing appears to have been practiced in Holland around 1430 and next in the Rhineland. Gutenberg used it in Strassburg (now Strasbourg, France) between 1434 and 1439.

But the experiments were not followed up because of problems created by the cast plates. It was difficult to strike each letter die with the same force and to keep a regular alignment, and, worse, each strike tended to deform the adjacent letter. It may well be that the major value of metallographic printing was that it associated the idea of the die, the matrix, and cast lead.

The invention of typography—Gutenberg (1450?). This association of die, matrix, and lead in the production of durable typefaces in large numbers and with each letter strictly identical, was one of the two necessary elements in the invention of typographic printing in Europe. The second necessary element was the concept of the printing press itself, an idea that had never been conceived in the Far East.

Johannes Gutenberg is generally credited with the simultaneous discovery of both these elements, though there is some uncertainty about it, and disputes arose early to cloud the honour.

It is true that his signature does not appear on any printed work. If masterpieces such as the Forty-two-Line Bible of 1455 rather than the imperfect products of a nascent typography such as the *donats* of 1445 or the "Astronomic Calendar" of 1447–48 are attributed to him, this is because of deduction and historical and technical cross-checking. The basic assumption is that, since Gutenberg was by profession a silversmith, he would have retained the role of designer in an association set up at Mainz, Germany, with the businessman Johann Fust and Fust's future son-in-law the calligrapher Peter Schöffer. The assumption is based solely on the interpretation of obscure aspects of a lawsuit that Gutenberg lost against his associates in 1455.

Apart from chronicles, all published after his death, that attributed the invention of printing to him, probably the most convincing argument in favour of Gutenberg comes from his chief detractor, Johann Schöffer, the son of Peter Schöffer and grandson of Johann Fust. Though Schöffer claimed from 1509 on that the invention was solely his father's and grandfather's, the fact is that in 1505 he had written in a preface to an edition of Livy that "the admirable art of typography was invented by the ingenious Johan Gutenberg at Mainz in 1450." It is assumed that he had inherited this certainty from his father, and it is hard to see how a new element could have persuaded him to the contrary after 1505, since Johann Fust died in 1466 and Peter Schöffer in 1502.

Evidence supporting Gutenberg's claim

The first pieces of type appear to have been made in the following steps: a letter die was carved in a soft metal such as brass or bronze; lead was poured around the die to form a matrix and a mold into which an alloy, which was to form the type itself, was poured.

Spectroscopic analyses of early type pieces reveal that the alloy used was a mix of lead, tin, and antimony—the same components used today: tin, because lead alone would have oxidized rapidly and in casting would have deteriorated the lead mold matrices; antimony, because lead and tin alone would have lacked durability.

It was probably Peter Schöffer who, around 1475, thought of replacing the soft-metal dies with steel dies, in order to produce copper letter matrices that would be reliably identical. Until the middle of the 19th century, type generally continued to be made by craftsmen in this way.

The typographer's work was from the beginning characterized by four operations: (1) taking the type pieces letter by letter from a typecase; (2) arranging them side by side in a composing "stick," a strip of wood with corners, held in the hand; (3) justifying the line; that is to say, spacing the letters in each line out to a uniform length by using little blank pieces of lead between words; and (4), after printing, distributing the type, letter by letter, back in the compartments of the typecase.

The Gutenberg press. Documents of the period, including those relating to a 1439 lawsuit in connection with Gutenberg's activities at Strassburg, leave scarcely any doubt that the press has been used since the beginning of printing.

Perhaps the printing press was first just a simple adaptation of the binding press, with a fixed, level lower surface (the bed) and a movable, level upper surface (the platen), moved vertically by means of a small bar on a worm screw. The composed type, after being locked by ligatures or screwed tight into a right metal frame (the form), was inked, covered with a sheet of paper to be printed, and then the whole pressed in the vise formed by the two surfaces.

This process was superior to the brushing technique used in wood-block printing in Europe and China because it was possible to obtain a sharp impression and to print both sides of a sheet. Nevertheless, there were deficiencies: it was difficult to pass the leather pad used for inking between the platen and the form; and, since several turns of the screw were necessary to exert the required pressure, the bar had to be removed and replaced several times to raise the platen sufficiently to insert the sheet of paper.

It is generally thought that the printing press acquired its principal functional characteristics very early, probably before 1470. The first of these may have been the mobile bed, either on runners or on a sliding mechanism, that permitted the form to be withdrawn and inked after each sheet was printed.

Next, the single thread of the worm screw was replaced with three or four parallel threads with a sharply inclined pitch so that the platen could be raised by a slight movement of the bar. This resulted in a decrease in the pressure exerted by the platen, which was corrected by breaking up the printing operation so that the form was pushed under the press by the movable bed so that first one half and then the other half of the form was utilized. This was the principle of printing "in two turns," which would remain in use for three centuries.

IMPROVEMENTS AFTER GUTENBERG

Several of the many improvements in the screw printing press over the next 350 years were of significance. About 1550 the wooden screw was replaced by iron. Twenty years later, innovators added a double-hinged chase consisting of a frisket, a piece of parchment cut out to expose only the actual text itself and so to prevent ink spotting the nonprinted areas of the paper, and a tympan, a layer of a soft, thick fabric to improve the regularity of the pressure despite irregularities in the height of the type. An engraving of an early printing shop is shown in Figure 1.

By courtesy of Archiv fur Buchgewerbe

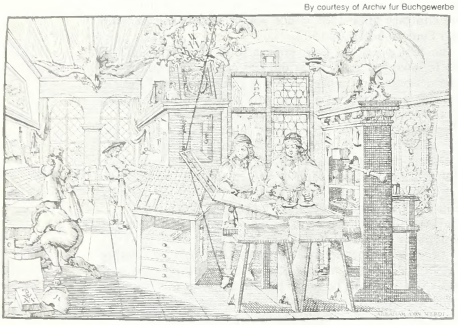


Figure 1: An early printing shop with lightly constructed wooden press; 17th-century engraving by Abraham van Weerdt.

The Dutch press

About 1620 Willem Janszoon Blaeu in Amsterdam added a counterweight to the pressure bar in order to make the platen rise automatically; this was the so-called Dutch press, a copy of which was to be the first press introduced into North America, by Stephen Daye at Cambridge, Massachusetts in 1639.

About 1790 an English scientist and inventor, William Nicholson, devised a method of inking using a cylinder covered with leather (later with a composition of gelatin, glue, and molasses), the first introduction of rotary movement into the printing process.

The metal press (1795). The first all-metal press was constructed in England in about 1795. Some years later a mechanic in the United States built a metal press in which the action of the screw was replaced by that of a series of metal joints. This was the "Columbian," which was followed by the "Washington" of Samuel Rust, the apogee of the screw press inherited from Gutenberg; its printing capacity was about 250 copies an hour.

Stereotypy and stereography (late 18th century). An increasing demand for printed matter stimulated the search for greater speed and volume. The concepts of stereotypy and stereography were explored. Stereotypy, used with notable success around 1790 in Paris, consisted in making an impression on text blocks of type in clay or soft metal in order to make lead molds of the whole. The stereotyped plates thus obtained made it economically possible to print the same text on several presses at the same time. The plates left the pieces of type in the form immediately available for further use and thus increased the rate at which they could be recycled.

A variation of stereotypy was the application, after 1848, of galvanoplastic metallization, in which process plates of thin metal lined with a base of lead alloy were made by electrolytic deposition of a coat of copper on a wax mold of the typeform.

Stereography aimed at bypassing the composition of the type in making the mold. Attempts to perfect the old metallographic method of preparing a clay matrix by stamping with dies brought no better results. In 1797 a variation was tried in which sets of copper matrices of each letter were made in large numbers. The matrices were then assembled according to the wording of the text, so that they covered the whole surface of the bottom of a mold in which the lead plate was then cast. Once the cast had been made, the matrices were available for further use.

Koenig's mechanical press (early 19th century). The prospect of using steam power in printing prompted research into means by which the different operations of the printing process could be joined together in a single cycle.

In 1803, in Germany, Friedrich Koenig envisaged a press in which the raising and lowering of the platen, the to-and-fro movement of the bed, and the inking of the form by a series of rollers were controlled by a system of gear wheels (see Figure 2, left). Early trials in London in 1811 were unsuccessful.

Presses with a mechanized platen produced satisfactory results after the perfection, in the United States, of the "Liberty" (1857), in which the action of a pedal caused the platen to be held against the bed by the arms of a clamp.

Though Nicholson very early took out patents for a printing process using a cylinder to which the composed type pieces were attached, he was never able to develop the necessary technology involved.

The cylinder was in fact the most logical geometric form to use in a cyclical process. It was also the one capable of providing the greatest output. Given an equal amount of energy, the pressure exerted by a platen had to be spread over the whole of the surface to be printed, whereas the pressure exerted by a cylinder could be concentrated on

By courtesy of Koenig and Bauer AG

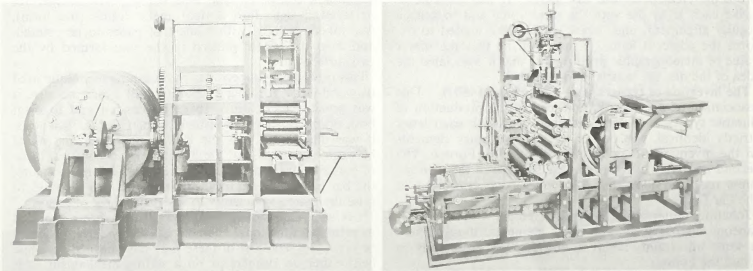


Figure 2: (Left) Friedrich Koenig's mechanical platen press, 1811, and (right) the first stop-cylinder printing machine, 1811, built by Koenig and Andreas Bauer.

the strip of surface actually in contact with the cylinder at any one instant.

A limited demonstration of the efficiency of the cylinder had been made as early as 1784 on a French press for books for the blind.

In 1811 Koenig and an associate, Andreas Bauer, in another approach to the rotary principle, designed a cylinder as a platen bearing the sheet of paper and pressing it against the typeform placed on a flatbed that moved to and fro. The rotation of the cylinder was linked to the forward movement of the bed but was disengaged when the bed moved back to go under the inking rollers (see Figure 2, right).

The London *Times* press of 1814

In 1814 the first stop-cylinder press of this kind to be driven by a steam engine was put into service at the *Times* of London. It had two cylinders, which revolved one after the other according to the to-and-fro motion of the bed so as to double the number of copies printed; a speed of 1,100 sheets per hour was achieved.

In 1818 Koenig and Bauer designed a double press in which a sheet of paper printed on one side under one of the cylinders passed to the other cylinder, to be printed on the other side. This was called a perfecting machine. In 1824 William Church added grippers to the cylinder to pick up, hold, and then automatically release the sheet of paper.

The to-and-fro movement of the bed that was retained in these early cylinder presses constituted an element of discontinuity; to make the cycle completely continuous, not only would the platen have to be cylindrical but the typeform also. In 1844 Richard Hoe in the United States patented his type revolving press, the first rotary to be based on this principle. It consisted of a cylinder of large diameter, bearing columns of type bracketed together on its outer surface; pressure was provided by several small cylinders, each of which was fed sheets of paper by hand. This system gave speeds of more than 8,000 copies per hour; its only drawback was its fragility; faulty locking up of the forms caused the type to fall out of the cylinder.

This defect was remedied by applying stereotypy to the production; that is, forming curved plates by making an impression of the typeform on strong pasteboard, the flong, or mat, which was fixed against the inside surface of a rounded mold, which was injected with lead alloy. In France, from 1849 onward, experiments were conducted with this process; it was regularly used in London by the *Times* from 1856 onward and after 1858 was in general use.

But feeding the press with paper still remained outside the mechanized cycle. Mechanization of this step was accomplished by the use of a continuous roll of paper supplied on reels instead of sheets. Techniques for producing paper in a continuous roll had been known since the beginning of the century. The first roll-fed rotary press was made by William Bullock of the United States in 1865. It included a device for cutting the paper after printing and produced 12,000 complete newspapers per hour. Automatic folding devices, the first of which were designed by Bullock and Hoe, were incorporated into rotaries after 1870.

Later, numerous other types of curved stereotype plates were used on rotary presses. These included electrotype plates that are curved before being backed; rubber or plastic plates made by molding or by a photomechanical process; and metal wraparound plates made by photoengraving or electronic engraving.

Attempts to mechanize composition (mid-19th century). Unlike the mechanization of the printing process, mechanization of the composition process was difficult to achieve in the 19th century. The invention of a compression mold in 1806 opened prospects for the mechanization of the production of type. In 1822 William Church of Boston patented a typesetting machine consisting of a keyboard on which each key released a piece of type of the corresponding letter stored in channels in a magazine. The pieces of type thus obtained had to be assembled by hand and the line justified. Church had avoided the problem of distribution and shown an intuition as to its solution by annexing to the magazine a device for constantly casting new pieces of type.

Numerous machines based on the same principle and with the further addition of a mechanism that placed the type pieces selected the right way round appeared in the course of the next 50 years. On one of these the more than 10,000 pages of the ninth edition of *Encyclopædia Britannica* were composed. These machines produced type at the rate of 5,000 to 12,000 pieces per hour, as opposed to about 1,500 by hand composition. But in all of them the type was simply delivered in a continuous row, which had to be divided into lines and justified.

These machines were completed by the introduction of a mechanical distributor, which was a sort of reverse compositor: pieces of type from lines that had been used passed before the operator, who pressed the corresponding key on his keyboard for the appropriate channel in the magazine to be opened up. The speed of mechanized distribution did not exceed 5,000 pieces of type per hour and was, thus, no faster than hand distribution.

Mechanization of letterpress composition faced two difficulties: first, justification, which required intelligent estimation of the size of spaces to be provided between words; and, second, the time taken during which the pieces of type were used for printing, which delay kept composition and distribution from being integrated into one cycle.

Typecasting compositors (1880s). Finally, in the 1880s in the United States, German-born Ottmar Mergenthaler invented the Linotype, a typecasting compositor that cast a solid one-piece line, or slug, from movable matrices of each letter. Each of the matrices was individually notched so that it could return only to its proper slot in the magazine after use. Justification was carried out by inserting wedged spacebands between groups of matrices immediately after making up the words of a given line. Here the matrices rather than type pieces went through the four basic operations of letterpress composition; cast lead was used for printing. The Linotype can produce the equivalent of 5,000 to 7,000 pieces of type per hour.

Mergenthaler's Linotype machine

In 1885, also in the United States, Tolbert Lanston invented the Monotype, which casts individual pieces of type for a line and justifies each line by a system of counting in units the width of the spaces taken up by the pieces of type. The matrices are indefinitely reusable, and the pieces of type, which are used only for the impressions, are returned to the caster. The contemporary Monotype typecaster is controlled by a ribbon of paper perforated on a separate keyboard. It can produce 10,000 to 12,000 pieces of type per hour.

In 1911 the American Washington I. Ludlow perfected a typecasting machine for the large display type that bears his name. The matrices are assembled by hand in a composing stick, which is then inserted above the opening of a mold; the matrices are also distributed by hand.

19TH-CENTURY INNOVATIONS

In the course of the 19th century several important innovations laid the foundation for a number of printing techniques that were not directly related to Gutenberg's invention.

Reproduction of illustrations. The first process for reproducing illustrations was xylography, using woodcuts that printed in relief and that therefore could be combined with letterpress, the picture blocks and the pieces of type for texts being locked into the same form. As early as the second half of the 15th century, xylography faced competition from engraving on metal that printed by intaglio; the metal plate (copper, sometimes brass, zinc, and even steel after 1806), engraved with a tool (burin) or etched with acid, was inked and carefully wiped so that ink remained only in the incisions and was transferred to paper under pressure in a cylinder press derived from the rolling mill. Since the intaglio method of printing was not compatible with woodcut printing, sheets of text and of illustrations for the same book had to be printed separately.

Presses for printing curved intaglio-engraved plates were perfected during the 19th century with mechanized inking with the use of rollers and wiping with the use of revolving cloth bands or rotating disks covered with calico. Their printing capacity was limited.

As early as the end of the 18th century, however, intaglio

printing had inspired a method for continuous printing of textiles by passing them under an engraved and inked cylinder from which excess ink had been removed by a scraper. In France in 1860 this technique was applied to printing paper for school-book covers. A solid copper cylinder was engraved not with continuous lines but with a multiplicity of tiny cavities in such a way that they retained the ink uniformly despite gravity, centrifugal force, and the action of the scraper. The process was suitable only for simple graphics.

Lithography: Senefelder (1796). A third printing process that had undergone significant development was lithography, neither relief nor intaglio printing but based on the principle that water and grease will not mix. In 1796 Aloys Senefelder of Prague investigated the properties of a stone with a calcium carbonate base and a fine, homogeneous, porous surface. A design drawn on its surface with greasy ink, wetted with water and then brushed with ordinary ink, retained the ink only on the design. This could consequently be reproduced on a sheet of paper pressed against the stone. Senefelder also established that a design drawn on such a stone and printed on paper could be transferred to another stone in as many identical copies as desired, side by side, which made it possible to obtain several copies at a time by printing a single large sheet. He further established that a metal such as zinc had the same properties.

Senefelder envisaged a press in which the stone, secured to an undercarriage, was inked, covered with the sheet of paper with a sheet of pasteboard above it, and submitted to pressure. By 1850 the first mechanized lithographic press with a cylinder, flannel-covered rollers for wetting, and rollers for inking was perfected.

The fact that it was possible to replace the stone by a zinc plate that could be curved made it possible to build rotary presses (the first in 1868) in which the paper passed between the plate-bearing cylinder and the impression cylinder.

Photosensitivity: Niepce (1820s). While searching for a means of automatically inscribing an image on a lithographic stone, then on a tin plate, in order to engrave it in intaglio, Joseph-Nicephore Niepce in the 1820s established that certain chemical compounds are sensitive to light. This marked the origins of photogravure (see *Photoengraving* below) and led to both the invention of photography (between 1829 and 1838) and the use of photographic processes for the printed reproduction of photographs.

In 1852 William Henry Fox Talbot, a British scientist and inventor, placed a piece of black cloth (tulle) between the object he wanted to reproduce (the leaf of a tree) and the photosensitive coating spread on a steel plate and obtained a picture that retained the fine mesh of the tulle. Consequently, etching with acid resulted not in an extensive and uniform erosion of an area but in tiny juxtaposed pits all over the photosensitive coating and varying in depth according to the degree of exposure. Talbot simultaneously had invented the screen and also had opened the way for a new development in intaglio printing: rotogravure.

The screen was perfected in the 1880s by substituting for the cloth two sheets of glass with uniform parallel lines that crossed perpendicularly. The screen made possible letterpress and lithographic reproduction of the full range of tones of a photographic document by using the effect of the diffusion of light through the mesh of its grid and converting the different intensity of tones into the different thicknesses of the printing surface.

Gravure and rotogravure (1890s). The circular mechanization of intaglio engraving, meanwhile, came up against two associated difficulties: the need to engrave an infinite number of tiny cells and the need to engrave them directly onto a cylinder. There were problems, because the rubbing of the squeegee to remove excess ink excluded the use of a curved plate that would not have provided a uniform surface in the area in which it was attached, and it was not possible to get photosensitive solutions to adhere to a cylinder.

In 1862–64 J.W. Swan of Britain invented carbon tissue, paper coated with gelatin that can be rendered photosensitive and exposed to light before being applied to a metal surface of any shape.

The Klietsch process

In 1878 a Czech, Karl Klič (also spelled Klietsch), thought of copying a grid screen directly onto carbon tissue, which could be used to transfer the cells necessary for intaglio printing to a cylinder at the same time as the image to be reproduced. In 1895 Klič, with English colleagues, founded the Rembrandt Intaglio Printing Company, which published reproductions of pictures, on paper, by rotogravure. They kept their process a secret.

In a parallel way, patents for a slightly different process, in which the image to be reproduced was screened before making the impression on the carbon tissue, were taken out in Germany and the United States. But a workman from the Rembrandt Intaglio Printing Company emigrated to the United States in 1903 and there revealed Klič's secret, and rotogravure, using his method, became widespread.

THE 20TH CENTURY

Beginning with the invention of the offset technique, the 20th century saw the steady development of innovations in the direction of mass production, speed, and economy.

Discovery of offset (early 20th century). At the same time, lithography was undergoing a new evolution. After the first mechanical presses had been perfected, this process had developed along two lines: (1) printing on thin sheets of metal (for example, tinplate for packaging canned foods) using a transfer process (1878) in which the impression cylinder carrying the metal sheet to be printed did not come in contact with the stone but did with an intermediary cylinder covered with rubber, the blanket, which transferred the image from the stone to the metal; and (2) printing on paper, which was done only comparatively infrequently in the last years of the 19th century, on cylinder or rotary presses.

In 1904 at Nutley, New Jersey, an American printer, Ira W. Rubel, discovered that an image accidentally transferred from the plate cylinder of his rotary to the rubber blanket of the impression cylinder during a paper-feed stoppage could itself be used for printing and in fact produced a superior impression. Rubel and an associate constructed a three-cylinder press, the first offset press, the term since used to describe this increasingly popular printing device.

Dry offset (1920). A few years later a problem arose in connection with printing the background of checks with a water-soluble ink to prevent forgeries. It was proposed that the lithographic plate of the plate cylinder be replaced with a stereotype plate or with a letterpress wraparound plate. This process, which combines the relief of letterpress, which does not require wetting, with the transfer of offset, is known as dry offset, or letterset. Its area of application is not limited to check backgrounds but is used in all areas of conventional printing.

Since 1950 another process has been developed, particularly in the United States. It combines rotogravure with the transfer of offset for printing wallpapers, plastic floor coverings, paper plates, and other products.

Colour printing. As early as 1457 a psalter, which Peter Schöffer signed but which some researchers now attribute to Gutenberg, included, in imitation of the contemporary illuminated manuscripts, paragraphs beginning with ornamental capital letters printed in two colours. This was accomplished by the use of two wood blocks that fitted one inside the other and could be separately inked.

Early colour experiments

Experiments to reproduce pictures in several colours on wood blocks were made in Germany in the 16th century. In the 17th century, different inks were applied to the different parts of the same engraved metal plate in such a way that all the inks were transferred to the paper in a single pressing. In 1719 a painter, Jacques-Christophe Le Blond, took out a patent in England for a process that used the three primary colours, blue, yellow, and red, and black for outlining shapes. Using a dense grid, he engraved four metal plates, bringing out on each plate the relative importance of the colour involved. The same sheet of paper then went through four successive impressions, each in a different colour.

In the 19th century the scientific definition of the principles of trichromatism, the enunciation of the fundamental theories of three-colour analysis and synthesis of colours by photography, the perfecting of coatings selectively sensitive to colours, and finally the use of the screen, instead of Le Blond's hand-drawn grid, established the modern trichromatic technique (which becomes quadrichromatic when black is also used).

Automation of composition (after 1929). The search for maximum efficiency had from the very beginning posed the problem of both the mechanization and the automation of composition. The Monotype system, with its separation of keyboard and caster, had constituted one approach to the solution, since the same caster could work at full speed when fed with perforated tapes produced on several keyboards.

The perfection of teletypesetter remote-control composing equipment in the United States by about 1929 permitted widespread application of the principle of separation of human function on the one hand and mechanized function on the other. The operator produces a tape on which each letter, symbol, and space is represented by a combination of perforations. A translator device reads the tape and, according to each combination of holes, orders the release of the necessary matrices for letters, signs, and justifying spaces. Machines casting one-piece fully spaced lines or slugs are able to produce more than 20,000 characters per hour.

Programmed composition (1950s). The production of the perforated tape remained relatively slow, however, because of the time taken by the operator to decide where to make the division in a word at the end of a line. Eventually, in the second half of the 20th century, electronics provided the means of automatically making this decision.

In the 1950s the BBR system, named by the initials of three inventors in France, introduced programmed composition. Starting with a perforated tape continuously produced by the operator, a computer takes over the task of determining the length of lines, the places where words are to be divided according to grammatical rules and typographic usage, the integration of corrections, and even the presentation of the text according to the layout. The speed at which a final tape bearing this information can be produced is limited only by the performance of the perforator, which is the outlet device of the computer. Operating speeds have exceeded 300,000 characters per hour, or 10 times the capacity of the most modern slug-casting machines.

Introduction of magnetic tape

During the 1960s, perforated tape began to be replaced by magnetic tape, which is even more rapidly made, at a rate of about 1,000 characters per second, or 3,600,000 per hour. Although magnetic tape is useless for mechanical composers casting pieces of type or lines in lead, such speed is practical for other kinds of machines not burdened with the weight of lead and the inertia of their mechanical components.

Photocomposition. In preparing cylinders for rotogravure, offset plates, or letterpress wraparound plates, it is illogical to use the vast weight of lead in letterpress composition to produce a reproduction proof that will then be photographed. Before the end of the 19th century this circumstance led to consideration of machines for composing headings by photographing the images of the letters in succession. In 1915 the Photoline, a photographic equivalent of the Ludlow, assembled matrices of transparent letters in a composing stick in order to film each line of the heading.

First generation of phototypesetters: mechanical. The next idea to be tried involved the adaptation of existing typesetters by replacing the metal matrices with matrices carrying the image of the letters and replacing the caster with a photographic unit. The industrial application of this idea resulted in the Fotosetter (1947), a phototypesetter, and its variant the Fotomatic (1963), controlled by a perforated tape, both derived from the Intertype slugcasting machine; the Linofilm (1950), derived from the Linotype; and the Monophoto (1957), derived from the Monotype. Retaining the mechanical limitations of machines intended to shape lead, they could not achieve appreciably higher rates of performance. Photocomposition had to be rethought in functional terms. This approach was explored in Germany as early as the 1920s with the Uher typesetter, which had photographic matrices attached to a rotating disk.

Second generation of phototypesetters: functional. The second generation of phototypesetters is characterized by the maximum elimination of factors of inertia, the number of moving parts having generally been reduced to only two: a steadily revolving disk or drum carrying the photographic matrices and an optical device of prisms or mirrors allowing directional action on the beam of light provided by an electronic flashtube.

The first revolutionary application of this notion was the Lumitype, invented as the Lithomat in 1949 by two Frenchmen, René Higonnet and Louis Moyroud. Executed by phototypesetting, *The Marvelous World of Insects* was done on their machine in 1953. The first model had an attached keyboard. Later models with a separate keyboard printed more than 28,000 characters per hour.

A new Linofilm (1954), functional and electronic, was fitted with a device for selecting matrices by the action of the blades on the photographic shutter, producing 12 characters per second, or 43,200 per hour. The model that succeeded it (1965) is equipped with a drum and is capable of double the output. The Photon-Lumitype 713 (1957) also performs at the rate of 70,000 to 80,000 characters per hour. But at this speed the technique of using a rotary matrix case reaches its limit because of the problems posed by centrifugal force. The Lumizip 900 (1959) introduced a further revolutionary change by retaining as moving parts only the lens, which scans in a single movement the fixed series of light matrices so as to photograph at one time the whole line of 20 to 60 letters. Output reaches 200 to 600 characters per second, or more than 2,000,000 per hour; this machine required magnetic tape.

The first book composed with a Lumizip, the *Index Medicus* (1964), was as much of a landmark in the history of phototypesetting as the Forty-two-Line Bible had been in printing. Its more than 600 pages were completed in 12 hours. To produce the same work on a typecasting machine would have taken almost a year.

Third generation of phototypesetters: electronic. Magnetic tape was still faster than the fastest phototypesetters. To narrow this gap, a third generation of phototypesetters appeared in the 1960s, in which all mechanical moving parts were eliminated by omitting the use of light and therefore omitting the moving optical device responsible for operating in its field.

Cathode-ray-tube phototypesetters (RCA, Linotron, etc.) operate on a principle analogous to that of television: a narrow pencil of electrons analyzes an image matrix of each letter and commands the modulation of another pencil of electrons on a luminescent screen, which leaves an impression on photographic film. Performance exceeds 500 characters per second and even approaches 1,000, or more than 3,000,000 per hour.

Digiset, a German development that appeared in 1965, pushed the use of the electron to its logical conclusion by suppressing even the image matrix of the character, simply keeping the binary analysis of its design available in its magnetic memory; this is, in fact, all that is needed to modulate the pencil of electrons on the final screen. Phototypesetters of this kind (called alphanumerical) have theoretical performance rates exceeding 3,000 characters per second, or more than 10,000,000 per hour, and should be able to approach 30,000,000. Speeds such as these exceed the production rate even of magnetic tape. Consequently, to work at its most efficient output, such a typesetter must be directly connected to a computer with a similarly high rate of output.

Speed of phototypesetters

Toward direct impression. The number of characters thus composed now approaches the number of characters printed on a press in the same time. The narrower this gap becomes, the more a still further revolutionary possibility looms—that of eliminating the press itself, since the typesetter can be made to deliver each page as quickly as the press would have. To accomplish this it would be necessary only to replace the photographic film that the

photographer imprints when it is conventionally used with an inexpensive carrier capable of receiving an image in the same way without pressure.

Several pressureless printing processes have already been perfected. In 1923 an electrostatic onset system drew the ink of a cylindrical typeform to the paper by means of an electrical charge. In 1948 two Americans conceived another type of electrostatic printing in which the colouring agent is not ink carried on a typeform but a powder or a solution sensitive to the pull of an electric charge inscribed in a plate. This technique gave birth to xerocopy in office duplicating (see below *Office printing*) and, at the industrial level, to xerography for producing posters and maps.

Printing without pressure can also be accomplished on papers impregnated with photosensitive preparations and passed in front of a cathode-ray screen of a phototypesetter. The first experiment using this facsimile printing process was carried out in Japan in 1964 by the *Mainichi shimbun*, a Tokyo daily newspaper. The image of the newspaper page formed on the cathode-ray screen was transmitted by radio waves, as in television. It was reproduced using the electrostatic system, which does not require chemical treatment of the paper after its exposure.

The Tokyo newspaper experiment

Serigraphy and collotype: a renaissance. Parallel to the evolution of the three major printing processes, letterpress, offset, and lithography, various other techniques have experienced a similar evolution, which has allowed them to survive or to establish themselves in the course of the 20th century and to preserve or win a place in printing.

The art of reproducing a design by forcing ink through the mesh of a silk screen partly blanked out with a stencil plate (serigraphy) had been practiced by the Chinese and Japanese long before the invention of letterpress. In the 19th century the textile manufacturers of Lyon adopted it for printing textiles. In the 1930s in Great Britain and the United States the most varied materials (glass, wood, plastic) and even the most varied shapes (round objects, for example) were printed by serigraphy, which from a handcraft progressed to an industrial technique, with the screen prepared by photosensitization and printing carried out by semiautomatic or automatic machines.

Another process, patented in France in 1855 under the name Photocollography, was modified in 1865 under the name Phototypy (still used in France) and in Germany in 1868 under the name Albertypy (still used in Germany). This process used photosensitive substances not as agents in making plates for printing but to serve directly as the effective surface of such plates. Known elsewhere as the collotype process, the technique was in great favour between 1880 and 1914, was then neglected, and has recently been revived and mechanized for printing posters and transparencies in black and in colour.

Flexography is a letterpress process using rubber plates on the plate cylinder; it occupies a special place in printing on account of the fluidity of its inks. It was first patented in England in 1890, and it was perfected in Strassburg a few years later. Flexographic printing is particularly suited to relatively coarse surfaces (pasteboard, wrapping paper, plastic or metal film) but has also been adapted to newspaper and magazine printing. It can be carried out by sheet-fed machines but is chiefly used on powerful rotaries.

Three-dimensional printing (1960s). In the 1960s a three-dimensional print was developed, essentially an illustration bearing two views, superimposed, of the same image taken from slightly different angles, on a transparent mount striped with a multitude of imperceptible parallel strips (Xograph process). On account of these strips, each eye, looking at the print from a different angle, sees only one image. The three-dimensional illusion is produced when this binocular vision is interpreted by the brain.

Office printing. The development of industry and commerce, in the 19th and 20th centuries, accompanied by an increase in administrative activity, created a demand for an abundance of printed information at various levels. In the field of office printing the first tool was the typewriter, perfected in 1867. Thereafter, machines appeared that would reproduce large or small numbers of copies of typewritten texts and, later, texts or illustrations of every kind. Some of these machines rely on techniques very close to those of conventional printing; others turned to original techniques that were in turn extended into modern printing. In 1881 in England appeared the stencil duplicator, basically employing the serigraphic technique. In 1900 a photocopying machine invented in France opened the way to facsimile printing. The offset printing process spread into the area of business printing with small offset duplicating machines; the simplified methods used for preparing plates for these machines eventually were adopted by industrial offset printers.

Stencil, facsimile, and xerography

The application of the electrostatic printing process to xerocopy, perfected in 1938, has since been taken over by industry.

All the various processes of duplication and reproduction of documents make up reprography, a name bestowed during the first congress devoted to these techniques, which was organized at Cologne in 1963. Though its boundaries with conventional printing are poorly delimited, to the extent that reprography can compete with conventional printing when a medium number of copies are concerned, reprography nevertheless represents an original field. In response to the increased need for quality reprography, the typewriter has been improved since the 1950s and given the capability of providing justified composition suitable for conventional printing.

Modern printing techniques

COMPOSITION AND TYPESETTING

Mechanical composition and typesetting. In the first decades of the 20th century all type was set and composed into columns and pages by hand or by mechanical means. These methods are still widely used.

Letterpress composition by hand. The font, which constitutes a complete set of characters of a given typeface, with duplicate numbers of each letter in proportion to the frequency with which each is used, is stored in the compartments of a case; capital letters, proportionately less frequently called for, are in the upper compartments, whence their name, uppercase, and the small letters in the lower compartments, which are more easily accessible and whence their name, lowercase.

The typographer works standing in front of the case. His principal tools are the composing stick, a metal angle iron with one fixed end and a "knee" with a screw or lever for locking; the line gauge, a ruler graduated in units of typographic measurement; and tweezers.

He locks the knee of the composing stick at the justification; that is, at the length of the line to be composed. Against the inside edge of the stick he places a lead, a strip of nonprinting lead alloy that later enables him, using a second lead, to grip the finished line in order to remove it from the composing stick. Holding the composing stick in one hand, he uses the other to select the individual type characters from the case. He can tell by touching which way up they should go, thanks to a nick indicating the top or bottom of the body (the bottom in English-speaking countries and Germany; elsewhere, the top), and he places them side by side in the composing stick. Having completed the proper number of characters to fill the length of the line with a whole word or at the correct division in a word, he adds as necessary to the nonprinting pieces already in place to mark the spaces between the words until the exact justification is obtained.

Having composed and justified the line, the typographer takes it, gripped by its two leads between the thumb and forefinger of both hands, to place it in a galley, a wooden or metal tray with a raised edge on two or three of its sides.

Semimechanized composition. The Ludlow is considered a combination machine; though it automatically casts slugs, it is related to hand composition by the way the matrices are assembled. The matrices are bronze blocks bearing the letter or sign engraved in intaglio on their lower side and with two shoulders on their upper side.

Operations of a Ludlow caster

The composer gathers them individually from the case, which is one of the drawers of a desk, and arranges them side by side in a special composing stick. This steel composing stick is hollowed out in the middle to receive the matrices supported on their shoulders with an adjustable

stopscrew for fixing the length of the line. Justification is ensured by blank unengraved matrices in various sizes equally distributed between the words.

The caster resembles a steel workbench with a hollowed-out slot on its surface in which the composing stick is inserted with the matrices face down. A lever starts the casting process by turning on an electric motor. A mold with an opening rises and positions itself under the aligned matrices; a plunger in the melting pot containing the molten alloy forces enough alloy into the mold to cast one line; casting is completed in less than 10 seconds, the mold withdraws and releases the solidified line, and the lever, which releases the composing stick, rises automatically.

Since the body size of the font is a uniform size, the upper part of characters whose body size exceeds its measurement projects beyond each side and has to be supported, when it is being used, with leads.

Since the width of the slugs is also uniform, when shorter lines are being cast the composing stick is furnished with thick, blank matrices; once cast, the line is clipped off to the proper length. For longer lines, composing sticks are used with justifications in multiples of that of the mold. Fractions of the line are cast one after another and fit together exactly. The Ludlow is used especially for casting lines of large type for use as titles and subtitles, using typefaces varying from 12 to 144 points (one point equals $\frac{1}{72}$, or 0.0138, inch).

The Ludlow caster is complemented by an Elrod caster. This automatically casts nonprinting leads and rules, narrow pieces of nonprinting alloy; both items come in various thicknesses.

Another type of mixed typecaster with manual assembly of the matrices is represented by the All-Purpose Linotype, a sort of Linotype from which only the casting part has been retained (see below *Linotype*). It is used primarily in United States printing establishments. An Italian equivalent, the Nebitype, is used, though less widely, in Europe.

Mechanical composition: slugcasting typesetters. The Linotype and Intertype slugcasting typesetters produce lines of letterpress composition in a single operation, starting with the assembling of the movable matrices. The letter matrices are thin, brass 19 × 32-millimetre (0.7 × 1.3-inch) plates, with two ears and a system of 14 notches arranged in a V on the upper surface and two heels in their lower part. The letter is engraved in intaglio on the face surface; usually two copies of the same letter are superimposed (duplex matrices)—one normal, or roman, the other a variant, either italic (sloping design) or boldface (stronger design). Thus, their thickness varies according to the letter and the body of the character.

The set of matrices is stored in a magazine, a flat, trapezoidal metal box consisting of 90 channels in which the matrices are aligned one behind the other, duplicates of 20 or 24 for each letter or sign, lying face down, resting on an ear and a heel.

Blanks are introduced into the line in two ways: either by using unengraved blank matrices, included in the magazine in three standard sizes, or by using spacebands designed to ensure justification.

The operator sits in front of a keyboard with 90 keys, corresponding to the channels in the magazine, on the left the lowercase letters, on the right the uppercase letters, in the middle the small capitals, numbers, and various symbols. A special bar operates the release of spacebands.

The nine steps in slugcasting

Slugcasting typesetters function as follows: (1) Touching a key releases the matrices, which are brought in proper order on a conveyor belt to a composing stick made of slide-bars and held by their ears. The spacebands, which are stored directly above the composing stick, fall into place between the words. (2) When the matrices and spacebands in the composing stick visibly take up the amount of space planned for the length of the line, the operator completes the line, either with a whole word or by dividing the last word, and pushes a lever to move the line. Since the remaining operations are done automatically he can go on to set the next line. (3) The assembled matrices and spacebands are moved three times in succession: vertically upward on the composing stick; sideways to the left on a transfer slide rest; vertically downward on an elevator that puts them in front of the opening of a mold mounted on a cogwheel called a mold wheel, connected to an electric melting pot containing the molten lead alloy. (4) A justifying hammer forces the long pieces of the spacebands upward, forcing them to separate by equal spaces until all the matrices and spacebands are locked between two steel jaws fixed at the precise justification of the line (see Figure 3). A piston plunges into the melting pot and forces the alloy into the mold to cast the line. (5) While the mold wheel rotates three-quarters of a revolution and the solidified line is finished to its exact letterpress height before it is ejected into a galley, the matrices and spacebands are again moved upward by the elevator. (6) They are pushed to the right toward a triangular bar bearing 14 grooves corresponding to the 14 notches in the matrices. (7) Raised by a catcher arm, this bar removes the matrices, which are caught by their notches; the unnotched spacebands are released and immediately return to the place where they are stored. (8) When the catcher arm is at its highest position, the matrices are pushed to the right toward another triangular bar with 14 grooves along its length and flush with the top part of the magazine; this is the distributor bar. (9) The matrices move along the distributor bar until at a certain point the arrangement of grooves ceases to provide support for the notches, which of course are different for each letter or sign. Each letter's matrix is then released at the opening of its own channel in the magazine.

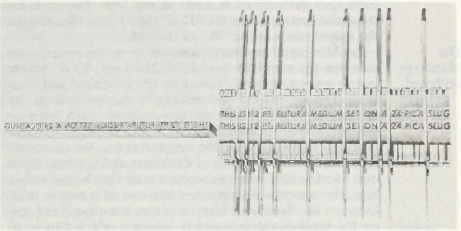


Figure 3: Lineup of matrices and justifying spacers (spacebands) and a cast line of type in a linecasting operation (see text).
By courtesy of Harris-Intertype

The automatic cycle of the typesetter is controlled by several large cams mounted on a single shaft driven by an electric motor.

Modern typesetting machines are equipped with several magazines of varying type sizes that can be used alternately. Some so-called double-distribution machines permit two magazines to be used at once by pressing a supplementary key.

The performance of recent models has been improved by accelerating the revolution of the matrices, intensifying the cooling system of the mold, and increasing the number of molds on the mold wheel to six.

The slugcasting typesetter, which furnishes solid, easy to handle, composed type, is particularly suited to printing newspapers. It has the disadvantage that to correct any error, however trivial, the whole line must be recomposed.

The All-Purpose Linotype is a combination manual and automatic machine that retains only the casting part of the Linotype. Special matrices, solidly rectangular or with notches, ears, and heels, are assembled by hand in a composing stick. Justification is done with blank matrices of various sizes. The line of matrices, held by the composing stick, is placed against two set squares fastened to the bedplate of the machine and manually pushed on a slide rest, which takes it to the elevator. The elevator places the matrices in front of the opening to the mold for the casting operation, which delivers the slug. The matrices are then distributed by hand.

Typesetter casting aligned characters. The operation of the Monotype typesetter, which casts individual aligned characters, is based on a system of measuring the width of characters, called the set. In each font, letters and symbols

have sizes determined in units of set, from five units for the narrowest, such as the "i" or the "l," to 18 units for the largest, such as the "W" or the "M."

The Monotype keyboard

The Monotype keyboard, separate from the caster, consists, on the standard model, of 274 keys, 30 of which, in two rows numbered from 1 to 15, are called justifying keys. Typing out the text results in the perforation, by an automatic punch, of one or two holes, each letter or symbol having its own pattern, in the width of a paper tape that allows 31 possible arrangements. An automatic calculator adds the widths of the letters and symbols typed out. Moving his forefinger across a scale that he has before him, the operator knows when the end of a line is near. Once he has finished a line, he places his other forefinger on the justifying drum, which indicates which two justifying keys he must now press. This move results in the perforation of one or two holes whose position indicates the quotient of the number of units lacking for the line to be completed by the number of spaces between the words of this line, plus a third hole in a position specifically fixed for the justification process.

The typesetter is composed of an electric melting pot containing molten alloy situated under a mold in the shape of a vertical chimney, the internal dimensions of which can vary according to the measurement in units of set of the characters or spaces to be cast.

The matrices are in the form of small bronze cubic blocks measuring five millimetres (0.2 inch) square, arranged in a steel frame nine centimetres (3.5 inches) square containing 15 rows of 15, enough for five complete alphabets, typically uppercase and lowercase in roman, italic, and boldface, and small capitals, as well as double or triple letters, numbers, and punctuation marks. Each row includes only matrices of letters and symbols of the same unit of set, from the smallest (five units) in the first row to the largest (18 units) in the back row.

The frame can slide horizontally in either direction to place any matrix from any row above the opening to the mold.

The process of setting and casting type on the Monotype is as follows: the roll of perforated paper tape is placed in the pneumatic tower of the typesetter—a row of 31 pipes distributing compressed air. As it unrolls, the tape prevents the compressed air from entering all pipes except those corresponding to the perforations.

The tape unrolls in the direction opposite to that in which it was rolled up—that is, in the opposite order to the way it was typed, the last line appearing first and the justifying perforations being inserted into the pneumatic tower before those for the letters and symbols. The compressed air that the perforations allow to pass into three (or two) of the pipes causes pieces of metal (justifying quoins) to fall into position in such a way as to control the internal measurement of the mold each time that the spaces between the words of the next line are to be cast.

The perforations for each letter or symbol allow the compressed air to pass into two (or only one) of the pipes connected to two blocks, each of which also contains a series of graded pins. Compressed air raises a pin in each block and halts the movement of the matrix frame in either of its sideways movements. In this way the matrix's row and the matrix's place in its row are selected. Selecting the row is the same as selecting a measurement given in units of set. Positioning the row is automatically linked to setting a piece of metal (the set quoin) whose position regulates the dimensions of the mold for casting the letter or symbol.

For the casting, a centring device places the selected matrix precisely against the opening of the mold, and a plunger in the melting pot forces the alloy up to cast the character or the space (that is, the unengraved matrix).

The composed line emerges from the machine completely assembled and justified and is placed in a galley.

The Monotype can cast type ranging in body size from five to 24 points (with a special mold for each). The addition of a speed-reducing device enables it to cast in 48 points. The maximum width of lines assembled is 60 picas.

In the early 1970s Monotype models could be equipped with a frame carrying 15 rows of 17 matrices (255) or 16 rows of 17 (272), with six or seven complete alphabets. The keyboard then has 310 keys.

A special model of the keyboard permits simultaneous perforation of two tapes for composing the same text in identical or different kinds of type and lengths of line.

Advantages of Monotype

The advantages of the Monotype system are the quality of its composition and the ease with which corrections can be made without having to reset the whole line. It is not well suited to newspaper printing because of the difficulty of handling lines of movable type and because, since typecasting begins with the end of the tape, composition must wait until all the type has been cast.

Automatic composition (perforated tape). The Teletypesetter (TTS) system extends to slugcasting machines the principle of separation of function originally characteristic of the Monotype: it enables Linotype or Intertype machines to be controlled by a perforated tape produced on a separate keyboard, even situated in a different city, since the combination of the perforations on the tape can be sent telegraphically.

The Teletypesetter tape is six-channelled; that is, it contains six possible positions for perforations across its width. This allows 64 different combinations of from one to six perforations.

This limited capacity, less than the number of keys on the keyboard of the typesetter, is corrected by an arrangement whereby each combination of perforations may have two different uses (for example, the uppercase and the lowercase of the same letter) according to whether it follows one or other of two special signals (themselves represented by combinations of perforations) that control passage of one or other of these uses.

The Teletypesetter keyboard

The keyboard for preparing the Teletypesetter tape looks like a typewriter with, in addition to the usual 44 keys and the space bar, 20 special keys. Striking each key establishes contact with the electric circuit or circuits that operate the perforators and at the same time acts on a calculating mechanism: a needle moving across a screen warns the operator of the end of each line.

Usually, as the tape is perforated the text is also typed out on a sheet of paper, which allows the work to be checked, reread, and corrected. For use with a Teletypesetter, the typesetting machine is equipped with a mechanism that translates the tape. In this mechanism the tape passes under six sensors that register electric contacts as the perforations pass. In accordance with the combination of electric contacts thus established, relays control the action of the keys or of the bar that causes the spacebands to drop and, at the end of each line, the starting of the casting cycle.

The most recent typesetters specially designed for use with the Teletypesetter offer such technical refinements as the elimination of the composing stick and immediate dispatch of the line to the elevator, simplifying the path taken by the matrices; and electromagnetic, rather than mechanical coupling, which speeds the starting up of the casting cycle.

Programmed composition (prepared by computer). The use of a computer eliminates manual intervention in preparing the perforated tape, in assessing the length of the lines, and even in deciding how to end them; *i.e.,* whether by completing or dividing a word.

Normally, the operator types out a continuous tape called idiot tape in the United States (kilometre tape in France) without concern for the length or division of lines. This band is inserted into the computer's input device, a tape scanner, which operates by means of either electric sensors or photoelectric cells and converts letters, signs, and orders into combinations of electric impulses. The computer semi-automatically or automatically processes this raw information in accordance with its programmed instructions and immediately communicates the result to an electromagnetic perforator, the output device, that produces a second tape, like the first but that also bears, in the proper places, perforations ordering the ends of lines.

A general program establishes the operation of the computer in its application to the work of composition. Individual programs adapt it to the machinery of the company concerned (models of typesetters, available magazines of matrices) and to the kind of work carried out (usual length

of line, method of indenting paragraphs, etc). Finally, special instructions punched on the tape by the operator at the same time as the text can interrupt the execution of the programs registered in the computer with directions valid for this text alone, in its entirety or in certain parts: choice of typeface among those available, transition from one kind of typeface to another available for the same typesetter, length of line and changes in the length, alignment to right or left, squaring of lines, indentations for ornamental capitals, spaces for borders or illustrations, and other details.

Having identified the combinations of the perforations on the tape and separately retained the service signals addressed directly to it, the computer proceeds to estimate the amount of the space occupied in a line by its letters and symbols, referring to instructions registered in its memory regarding each. In the same way, it determines a justification zone in which a division in the line is necessary and possible, the minimum and maximum limits of this zone being fixed by the limits of expansion of the spacebands on the typesetter.

If the end of a whole word comes within the justification zone, the computer itself signals the end of a line after this word and suppresses the space that would normally follow it. Otherwise, the last word must be divided. The process is said to be semiautomatic if a special operator, seated at a keyboard linked to the computer, must intervene to decide where to place the division in the word submitted to him on, for example, a cathode-ray viewing screen. The process is automatic if the computer is designed and programmed to make the decision itself; the operation then is carried out by starting a subprogram in which all the divisions possible in the word considered are listed (after a prefix, between syllables). This list is tested against prohibited divisions (according to the rules of etymology, phonetics, typography) stored in the rapid-access memory of the computer. From among the positions that are not eliminated during this test, the computer chooses the one situated nearest the end of the word. It automatically inserts the signal for the hyphen and orders the end of the line.

Computerized division of words

The computer can also carry out the correction of mistakes before composition. Various methods are possible, of which two will be described briefly. In one, the perforation of the justified tape delivered from the computer includes the introduction at the beginning of each line of a numbered signal and is accompanied by a proofing copy of the text with a corresponding reference number for each line. When the mistakes have been corrected on the proofing copy, an operator prepares another, much shorter correction tape, which consists of the corrections preceded by the reference to the line on which they occur. The justified tape and the correction tape are jointly introduced into a double reader, the mixer, which determines anew the length of the line and where the division should occur, as well as for such succeeding lines as need to be modified, before producing a final tape.

In the second method, the proofing copy can be typed out or shown on a cathode-ray viewing screen with the lines numbered but without the tape. At a keyboard connected to the computer, an operator types out the corrections, preceded by their line reference. If a viewing screen is used, the text reappears immediately in its corrected form, and the output perforator immediately delivers a justified and corrected tape.

The computer is usually programmed to sort out and correct even mistakes or anomalies in typing, such as the presence of two consecutive spaces, in which case it cancels one. It can, if its capacity allows, receive a makeup program independent of the tape of the text; following the specifications of the layout (positioning and size of headings, text, and illustrations) coded in binary language, the computer itself introduces onto the perforated tape the special instructions concerning kinds of typeface, length of lines, changes in lines, etc.

Because of the quantity of information needed for composition, the six-channel Teletypesetter tape is being increasingly replaced by seven- and eight-channel tape.

Computer processing using a continuously typed tape can be applied equally well to the Monotype system. The programmed operation for dividing lines is in this case carried out by the automatic calculation of the width of the spaces between words and by the perforation, before the end-of-line signal, of a signal signifying the appropriate position of the justifying quoins. To enable the text to be read by the pneumatic tower of the typesetter, a converter transcribes the perforations from a narrow six-, seven-, or eight-channel conventionally perforated tape to the wide tape of the Monotype system.

The use of computers is now widespread in preparing photocomposition jobs, with programs adapted to the specifications. The computer's output device can produce magnetic tape instead of perforated paper tape.

Optical scanning devices

One intake device no longer reads perforated tape but is an optical mechanism for scanning a typewritten text. The Retina reader, for example, is a sort of artificial retina made up of a group of photosensitive units able to identify each letter typed by a special typewriter, using only three data: height, width, and gray value; that is, the surface area occupied by the outline of its design.

Cold type. Cold type is the expression used, particularly in the United States, to describe a simple and economic method of preparing text by machines resembling ordinary typewriters but capable of producing justified lines in type that varies in width according to the letter involved. Justification is achieved in several ways by different versions of the machine. In the IBM Multipoint, a first typing calculates the total measurement of the type pieces up to the beginning of the justification zone and causes a coded sign to appear. A button is set in position over the coded sign thus assigned to each line before a second, final typing is done. The position of this button determines the automatic adjustment of the spaces between the words to the amount needed to obtain justification.

In the Justowriter, the keyboard on which the uncoded, unjustified proofing copy is typed simultaneously perforates a paper tape with the code for the letters, as well as, for each line, the code for the amount of space between the words as indicated by a calculator. The tape then controls, on a second unit of the machine, the electric typing of the final justified copy.

In the IBM Multipoint with magnetic tapes, a magnetic tape produced at the keyboard is processed by a computer for justification and, if necessary, for corrections. The final tape delivered by the computer controls the action of an output unit, which carries out the final typing.

If the copy thus produced on paper is to be photographed to prepare printing plates by photogravure, cold type cannot be directly incorporated into photocomposition because of the intermediate operation.

Optype is a hybrid process that simultaneously carries out the operation of justifying a text typed directly in cold type and transmits it to photographic film. By means of optical distortion, each line is stretched to the exact length of line projected on the film. The same mechanism also enables the line to be magnified or reduced or set in italics.

Phototypesetting. Using phototypesetting, a direct image of the text is obtained, positive or negative, according to need, on a photosensitive, usually transparent surface by exposing the surface to light through transparent matrices, negative or positive, of the letters and symbols.

Manual phototypesetters. Several small machines permit phototypesetting of short texts and titles in conditions to a greater or lesser degree short of automation. Among them are the following:

Dantype uses separate transparent plastic matrices, which are assembled in a composing stick and placed in direct contact with the photosensitive film inside the machine.

Typro makes use of letters and symbols on a negative film that moves to and fro to place the desired type piece in contact with the photosensitive film.

Headliner incorporates letters and symbols that appear in negative on an interchangeable plastic disk whose position is controlled from outside. The film is exposed by contact.

Hadego uses plastic matrices assembled in a composing stick, exposure taking place through an adjustable photographic lens that permits enlarging or reducing. With just two series of 350 matrices, one with a 20-point body, the

other with a 48-point body, all sizes of type from eight to 110 points can be obtained.

The Starlettograph, comparable to an ordinary photographic enlarger, can be used only in a darkroom. The type, inscribed on a semirigid plastic tape, is set in position one piece at a time, using red light that does not affect the photosensitive film.

Letterphot works on the same principle as the photographic enlarger but on a luminous table. A first projection is made of all the characters of a line without the sensitive surface. Then the sensitive surface is placed on the luminous image of the line, which appears transparently and cannot therefore make an impression. Letters are successively printed in a two-part operation. First the letter is projected in normal light to cause it to coincide with its luminous image; the normal light does not make an impression on the sensitive surface, because the latter has a special composition. After this adjustment has been made, the letter is projected in actinic (photographically active) light, which exposes the sensitive surface.

Diatyp and the Monotype photoheadliner (as well as the Varityper, which is similar in composition) are more elaborate phototypesetters, easier to operate and permitting production speeds of nearly one character per second. The image of each character on the matrix disk is controlled by a symbol that is read by photoelectric cells and which automatically moves the film forward the same amount as the space taken up in the line by the character. A totalizing calculator informs the operator of the rate at which the line is being completed, and justification can be achieved by a first typing without having the source of light in operation; in a second typing, the spaces between the words are adjusted the necessary amount. Adjusting the lens of the Diatyp produces characters ranging in size from four to 36 points and, using the Monotype, from five to 84 points.

Automatic phototypesetting. The first Linofilm was a direct adaptation of the Linotype. Its photographic matrices were the normal Linotype matrices, the only difference being that, instead of bearing an intaglio engraving of the character on their face, they bore its outline in black on a white background. Lines were composed in exactly the same way as on the typesetter, justification being carried out by expanding the spacebands. The justified line is then passed a single time in front of a lens to be photographed.

Fotosetter operations

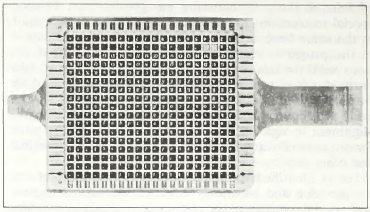
The Fotosetter is an adaptation of the Intertype machine but with functional differences. The matrices resemble matrices used for casting; they have the same notching and different thicknesses, depending on the character. But the outline of the character, instead of being inscribed on the face, is a transparency (*i.e.*, a photographic negative), in a capsule set into the level surface of the matrix. These special matrices are called fotomats. In place of spacebands there are space fotomats of different thicknesses.

The Fotosetter is equipped with magazines of 117 channels, 27 more than the typesetters, with an enlarged keyboard of 114 keys.

Once the line has been assembled and justified, using space fotomats of the necessary sizes, the fotomats move inside an optical apparatus that sends a brief flash of light toward the sensitive film. After each exposure, the support of this film is moved slightly sideways by a rack-and-pinion system commanded by the withdrawal of the next fotomat from its alignment; the matrix moves in proportion to the thickness of this fotomat. When all the type pieces in a line have been photographed, the film unwinds the correct amount to present a clean surface ready for the phototypesetting of a new line, while the fotomats are carried off to the distribution bar.

Equipped with a turret of 14 different lenses, the optical apparatus produces 14 sizes of type from three to 72 points, from the same set of fotomats of uniform 12-point size.

The Monophoto is a direct adaptation of the Monotype system with, on the one hand, an independent keyboard that produces a wide perforated tape in the Monotype code and, on the other, a phototypesetter operated by inserting this tape. The type pieces are chosen by positioning a frame, which carries 17 rows of 20 cubelike matrices in



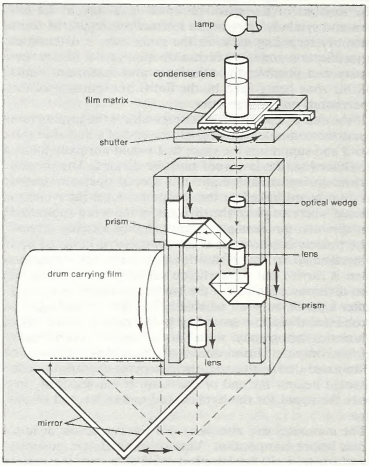


Figure 4: Basic operation of the optical system used in Monotype photocomposition (see text). A film matrix is shown at top.

By courtesy of Monotype

which the letter or symbol appears as a transparency, in negative, in the path of a beam of light (see Figure 4). This beam, after proper processing, is directed toward the sensitive film, on which it makes an impression. It first travels through a combination of magnifying glasses and prisms whose position in relation to each other is adjusted to obtain the desired ratio of enlargement or reduction. The sensitive film remains stationary on the drum carrying it as a composed line, while the element that enables the beam of light to move from letter to letter is a set of two mirrors placed face to face at a 90° angle and mounted on a mobile carriage. Before each exposure, this set of mirrors shifts, parallel to the direction of the sensitive film, the same amount of space as the width of the character about to be composed. This amount of space depends also on the number of units of set of the letter or symbol and on the ratio of photographic enlargement or reduction. The movement of the mirrors is thus subjected to the command mechanisms of two factors: the position of the frame, since the matrices are arranged in rows of the same units of set, and the adjustment of the combination of magnifying glasses and prisms.

Justification is accomplished, as on the typesetter, by predetermining the width of the spaces between the words. Since the justification perforations appear before those for the type pieces, they establish for the line to come the amount of space the set of mirrors has to shift at each space command punched in the perforated tape.

After all of the type in a line has been photographed, the set of mirrors returns to its original position, and the drum bearing the sensitive film turns the amount necessary to

continue on to the composition of the next line according to the degree of line spacing (leading) chosen.

Using matrices of a single eight-point size, the Monophoto makes available the whole range or type size from six to 24 points. For perfect photographic reproduction it is usually found preferable to use two or three sizes of matrices to cover this range. Given the quality of its production, the Monophoto, sometimes linked to a unit programmed to prepare the tape, is popular for work that demands careful composition.

Functional phototypesetters. A second generation of phototypesetters consists of functional machines that are analogous neither in structure nor in operation to typesetters using lead. Outwardly they resemble metal chests comparable to office furniture. Their design aims at reducing mechanical parts, inertia, and friction to the minimum. Their technical characteristics vary according to model, use, and cost.

The keyboard, which is hardly more complex than that of an ordinary typewriter, can be attached or separated; in the latter case, information regarding the text to be composed is inserted by perforated tape. Computer units can be integrated, either merely to direct the machine's operations or to ensure completion of the justification process, division of words, and correction, whether from the adjoining keyboard or from a continuous perforated tape. Selection of the matrix image of each character can be done either by using a mobile support for the matrices (plastic tape, disk, drum) in front of a fixed source for the beam of light or by using a mobile beam of light in front of a fixed support for the matrices (glass or plastic plate); alignment of type pieces is carried out in either case by mobile prisms or mirrors.

Versatile capabilities of the photo-typesetter

In addition to enlargement or reduction, the optical apparatus can be designed to carry out special effects, such as converting roman to italic type or stretching a line to make it longer or higher. The photographically composed text can be delivered either on paper or on film, in positive or negative, or in straight reading or reverse reading. The source of light is usually an electronic flash the intensity of which can, if necessary, be made proportionate relative to enlargement or reduction.

Some of the characteristics of phototypesetters are outlined below.

Linofilm (new method): The matrices of the 88 characters in a set are inscribed on a plate of glass that remains stationary during composition. The character is chosen by the shutter of the photographic lens. This shutter consists (as in a commercial camera) of very thin, overlapping metal blades, eight in number. Instead of always opening at the same point at the moment of exposure, it opens facing the desired character, each being set in position by an electromagnet so as to obtain this arrangement. After passing through the matrix of the character thus chosen, the beam of light is taken over by one of the 88 small lenses arranged behind the plate of glass and its trajectory directed towards a mirror mounted on an undercarriage, which carries out the alignment on the sensitive film.

Using this very light electromagnetic mechanism, the Linofilm can produce up to 12 exposures per second, or 43,000 symbols per hour. Eighteen matrix plates arranged in a turret magazine are instantaneously usable, producing 1,584 characters. Three matrix plates are enough to photograph the same type face in 16 sizes, from six to 36 points.

Diatronic, a phototypesetter made in Germany with an adjoining keyboard, uses matrix plates with 126 symbols. Selection is made after the beam of light has passed through all the symbols on the plate, through prisms which take up the position necessary to retain only the light coming from the matrix of the chosen character.

Continuous high-speed photo-typesetting

Photon-Lumitype was the first phototypesetter to introduce the selection and photographing of the character in a rapid circular movement without interrupting continuity.

The matrices are inscribed in concentric circles on a disk that revolves continuously at 10 revolutions per second in front of an electronic flashtube whose light lasts a few millionths of a second for each character. Selection is by means of a system of rotary contact makers, controlled by the telegraph system. A nylon drum is integrated with the matrix disk and turns with it in the same movement; the drum is encircled with as many tracks as there are channels in the binary code used to define the characters. These tracks are the transmitting and isolating elements that pass under a row of electric sensors. A special combination of transmitting and isolating elements corresponds to each character matrix positioned ready to be photographed.

Whether by striking a keyboard or by perforating a tape, selecting a given character consists of the precise formulation of the combination that establishes the electrical contact and initiates the flash of light. This selection can be acted on only at the precise moment when, as the disk revolves, the matrix of the desired character moves into position to be photographed.

Whether textual information is fed into the machine on an adjoining keyboard (as on the early Linotypes) or on a keyboard directly connected to the photographic unit or whether it is done on perforated tape, this information is in every instance preserved, line after line, in a memory, formerly mechanical but magnetic on the later models, which at the same time permits calculation of the size of the spaces between the words and ensures that the character's binary signal is presented during the 1/10-second period of time available.

It is possible to attain a production speed of 10 symbols per second, or, theoretically, about 36,000 per hour; in practice the figure averages less.

Each of the eight concentric circles on the matrix disk contains two complete sets of 90 characters, which can be filmed in 12 sizes, from five to 72 points. In other words, a total of 17,280 characters are immediately available.

Another Photon-Lumitype model operates on the same principle, but the disk is replaced by a drum revolving 30 times per second around an axis that coincides with the source of light. The type matrices are inscribed in negative on two films carried on the surface of the drum, and the source of light consists of two electronic flashtubes, one for the upper, the other for the lower half of the drum (see Figure 5, right). The lens and mirror system for producing images of the matrices on a moving light-sensitive film is shown at the left in Figure 5.

By courtesy of International Photon Corp.

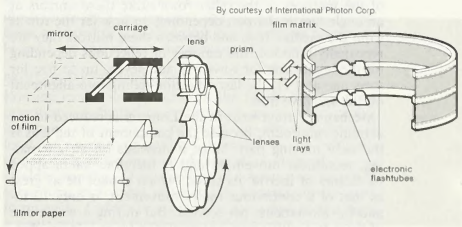


Figure 5: Basic operating principle of the Photon 713-10.

This model's total capacity in characters (four complete sets and eight ratios of enlargement or reduction) is three times smaller, but its speed of composition is three times faster: it can attain 80,000 symbols per hour.

By putting the same films of type matrices in the upper and the lower part of the drum—that is, by cutting in half the number of characters stored—speed of production can be raised to 120,000 symbols per hour.

Europa-Linofilm is similar in design to the Photon-Lumitype just described, with a permanently revolving drum but with an electric type-matrix selection system. These matrices are small individual plates bearing not only the negative image of the letter or symbol but also a series of transparent marks whose arrangement constitutes its binary identification code. In the revolving drum this coded part of the plates passes in front of a scanner made up of a series of photoelectric cells. As soon as the scanner picks up a coincidence between the code of the type matrix and the code of the character selected for composition, it activates the shutter release of the electronic flash.

The Europa-Linofilm drum is composed of four superimposed levels, each containing 120 duplex type matrices—

for example, with the same letter in both roman and italic—easily interchangeable in order, since their identification is not linked to their place.

The Photon-Lumizip

Photon-Lumizip is based on a different principle. The performance speed of drum phototypesetters can hardly be increased because of the technical problems posed by the rapid rotation of the drum. To increase speed, the Lumizip abandoned rotary movement. The type matrices are stationary and are aligned in negative on a large-sized plate. There is an individual electronic flash behind each type matrix. The sensitive film is stationary while a line is being composed. The only moving element is the component of the lens placed between the plate and the film, which carries out a rectilinear to-and-fro movement, parallel to them. The flashtube situated behind a given matrix emits its beam of light each time that the component of the lens finds itself, in the course of its to-and-fro movement, in the axis that joins this matrix to the position where the letter to which it corresponds is to appear in the line to be composed on the sensitive film. Thus, the order in which the characters are photographed is neither that in which they appear in the text nor that of the matrices on their plate but that determined by an angular relationship between them both.

A computer is built into the Lumizip. As soon as the coded signals for a line to be composed reach the computer, it determines the order and the exact moment when each flashtube is to operate, synchronized with the movements of the component of the lens.

In practice, the type matrices are aligned on the plate not in a single row but in 11 horizontal rows. The component of the lens always moves in the same horizontal plane, which is the same as that of the line to be composed and also that of the median row, the sixth, of the matrix plate. Alignment of characters from the other rows is achieved by means of two level, horizontal mirrors placed on either side of this horizontal plane, parallel to it and face to face at a small distance from one another. The beams of light coming from the flashtubes of the matrices of the median row pass between these two mirrors without touching them. But the beams of light coming from the flash tubes of the matrices of the other rows strike these mirrors at an angle that is sharper, depending on how far the row is from the median row, and between these mirrors they are repeatedly reflected, the number of reflections depending on the angle: one for rows five and seven, up to five for rows one and 11, the last reflection falling into alignment on the sensitive film.

Mechanical movement in the Lumizip is reduced to the extreme minimum, because the component of the lens is the only moving part. Since it depends on an alternating, rectilinear movement and is therefore handicapped by factors of inertia, its speed, which cannot be as great as that of a continuous rotary movement, is only 10 to-and-fro movements per second. But during a single one of these to-and-fro movements all the several dozen characters for one line are photographed. Thus, the Lumizip attains a performance rate perhaps 20 times superior to that of the Lumitype. Theoretically, it can perform at a rate exceeding 2,000,000 symbols per hour and in practice has produced over 1,000,000.

Electronic phototypesetters. In phototypesetters of the third generation, the beam of light is replaced by a flow of electrons, which offers the advantage that the electrons can be deflected by means of magnetic fields without the intervention of mechanical parts such as mirrors and lenses. Television systems are based on this characteristic, and an early type of electronic phototypesetter is structurally comparable to a closed-circuit television system. A reading device analyzes, by fine scanning, the outline of the matrix of the letter to be composed and converts the luminous information it obtains into electronic signals. The cathode-ray screen of an output device reconstitutes, in accordance with these signals and by a fine scanning device synchronized with the reading device, a luminous image of the letter, which makes an impression, through an optical reducing device, on a photosensitive surface. A computer, depending on the text to be composed, directs the position of the reading device's scanner towards that part of the matrix plate that bears the matrix of the letter selected, and simultaneously directs the position of the output device's scanner toward that part of the screen that corresponds to the position of this letter in the line being composed.

On some models, the scanning device consists of the equivalent of a television camera whose electron beam is selectively deflected towards the chosen matrix and directly analyzes the luminous information coming from it. On others, a cathode-ray tube takes the place of the emitter of a regular beam of light by scanning behind a plate on which the matrices appear in transparency. On the other side of the plate, photoelectric cells collect this beam at the moment it passes through the matrices and react by emitting electronic signals directed to the output device.

Electronic matrix selection

For matrix selection, the face of the emission tube is divided into 16 square sections (four by four), of which only one is illuminated at a time by selective scanning directed towards it. There are 16 photoelectric cells arranged in a square (four by four), only one of which is in operation at a time. Thus, there are 256 (16 by 16) possible arrangements of the chosen section and of the chosen cell. Each combination corresponds to an optical trajectory belonging to one or the other; that is to say, to the precise positioning of one of the matrices over the plate.

On the screen of the output device, the letters have a definition of 650 lines per inch for ordinary work and 1,300 lines per inch for quality work. The line structure is invisible after the letters have been reduced for photographic reproduction.

A more complex model of the Linotron scans, on the screen of the output device, the surface of a whole page, composing as it goes all examples of the same letter in all the places where it occurs on the page. Composition of the page is completed after all the matrices have been exposed once. The average speed of production is on the order of 1,100 symbols per second, or almost 4,000,000 per hour.

Carrying the system of electronic composition to its logical conclusion, designers have replaced the matrices, whose outline had to be repeatedly electronically analyzed, by the results of analyses previously carried out and preserved in binary form in a magnetic rapid-access memory, setting up for each letter the output program for its luminous image on the cathode-ray screen when it is selected for composition. Electronic phototypesetters of this kind are called alphanumeric.

Hell-Digiset carries out a preliminary analysis by inscribing the outline of each letter on a very dense grid of 3,000 to 6,000 small squares, according to the body size of letter envisaged. Those squares covered by the outline are assigned the symbol 1 of the binary code; the others are assigned the symbol 0. The result of the analysis is first inscribed in perforations on an eight-channel tape. Tape containing perforations for an entire set of type in a given style is inserted into a special Digiset reader to instruct the magnetic memory, in a few dozen seconds, concerning type production. All that is necessary to change the style of type is insertion of the tape belonging to another set.

The Digiset 50 T 2 can reach a production speed of 3,000 characters per second, or more than 10,000,000 per hour. One Digiset is designed to permit a whole newspaper page to be composed photographically in a single scanning operation; not only the words but also the illustrations are analyzed in binary code.

Fototronic-CRT and APS (Alphanumeric photocomposition system) reduce the amount of coded information by interpreting each letter as a series of closely packed adjacent vertical lines whose distinguishing parameters are their height and their position. Vertical scanning on the screen of the photographic output device reproduces these lines one after another according to these parameters (see Figure 6).

The number of lines varies from about 50 to 90, depending on the width of the letters, and the number of units calculating the measurement of parameters of height can go up to 80, which amounts to a definition perceptibly as fine as the Digiset grid, or 800 lines per inch in two dimensions on the screen of the output mechanism.

The APS electronic phototypesetter has a production

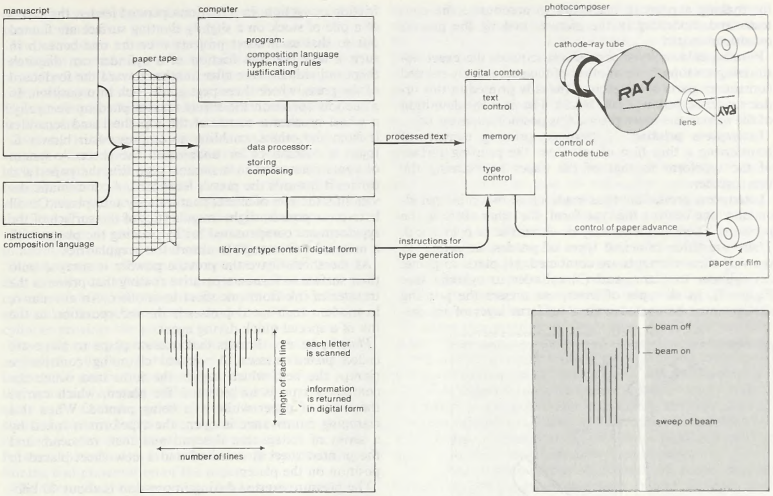


Figure 6: Basic components of the Alphanumeric photocomposition system (see text).
By courtesy of Alphanumeric Inc.

speed of 3,000 to 10,000 characters per second, the latter figure amounting to 36,000,000 per hour.

Makeup of letterpress copy. Preparing a form suitable for use in printing from letterpress copy, whether in individual type pieces or in lines of lead alloy, is an operation called makeup. This is preceded, if the same form is to include several smaller pages to be printed together, such as a book, by an operation called imposition, which consists in laying out the pages in the form so that they are in their numerical order after the printed sheet has been folded into a signature of eight, 16, or 32 pages.

Newspaper makeup

In the case of a daily newspaper, one form is generally used for each page. When the manuscript or typescript copy containing information regarding the type and the justification required arrives at the printing plant, it is divided among several Linotype machines. The titles are composed, depending on their body size, in movable type on a Ludlow or Linotype machine. When corrected from galley proofs, the titles and columns of text arranged in galleys are brought to the compositor, together with, where relevant, plates of illustrations or advertising matter mounted on lead blocks the same height as the type. Standing at a level casting table, the stone, and following the instructions on a layout, the compositor arranges the elements inside a rectangular steel frame, the chase, usually equipped with a locking system using quoins that slide along the inside of two of its adjacent sides.

The compositor inserts leading between paragraphs to bring each column on the page to its proper height. He separates columns and articles using leading or rules cut to the desired dimensions. Once the chase has been locked, proofs are taken to check the page for final corrections before it is pressed on to a metal frame to be brought to the press.

Mounting composition on film. Composition on film consists of a mounting operation carried out on a luminous table. Films of the text and titles and positives or negatives of the screened or unscreened photographs, depending on the printing method to be used, are arranged on a sheet of transparent plastic of the dimensions indicated on the layout, placed under the plastic sheet and lighted from beneath. The film is glued or fixed using pieces of transparent adhesive.

Conversion systems. One composition process can be converted to the other. A page composed on film, in negative, can be used for photogravure plates or for metal or engraved plates intended for letterpress printing (see below *Printing*). Contrariwise, a page composed in type can be converted to a positive or negative, direct or inverted transparency by any of a number of techniques.

The whole page, including the screened plates or photographic illustrations, can be subjected to conversion, or the text can be taken alone, leaving the positioning of illustrations until the positives or negatives of the illustrations, screened or not, according to the printing process to be used, are included in the makeup at a later stage.

PRINTING (PRESS OPERATION)

Printing in the second sense of the word—that is, press operation—is the technique by which ink or other colouring agent, transferred to paper or any other material, is localized on printing surfaces delimited by the composition of texts or the making of illustration material.

Colour printing. Juxtaposition of colours is achieved by submitting each sheet to successive impressions by typeforms each of which prints only on areas designed to carry a single colour and inked only in that colour.

There are sufficient wavelengths in the three primary colours, blue, red, and green, to reconstitute all the colours of the spectrum. The colour perceived from an inked impression results from the fact that the ink reflects some waves of the range of colours of the spectrum and absorbs the others—that is, prevents them from being seen. Thus, three colours of ink can reconstitute the visual effect of all the range of colours by combining them appropriately. Yellow totally absorbs all waves of colour that are basically blue and reflects waves of colour that are basically red and green; magenta (deep crimson) totally absorbs green and reflects red and blue; cyan (close to turquoise) absorbs red and reflects blue and green.

Effects of combining coloured inks

When two of these inks combine, each annuls in the other the capacity to reflect that one of the two primary colours that it cannot itself reflect; the eye then perceives only the one primary colour that both inks reflect; for example, red, by combining yellow and magenta inks. All three inks combined no longer reflect any of the three primary colours but appear as black.

In trichromatic printing the screened plates necessary for each of the three colours of ink are prepared by selecting the colours through filters. A fourth plate is usually used

for making a print in black ink to accentuate the contours and modelling in the picture, making the process quadrichromatic.

Printing colours by superposition requires the exact positioning on top of one another of the successively printed constituent parts of the picture (usually printed in this order: magenta, yellow, cyan, black). The finer the definition of the screen, the more precise this positioning must be.

Letterpress printing. Letterpress printing consists of transferring a thin film of ink from the printing surface of the typeform to that of the paper by pressing the two together.

Letterpress presses are thus made up of two principal elements, one bearing the type form, the other exerting the pressure. These elements may be either flat or cylindrical. There are three principal types of presses, according to the way these elements are combined: (1) plane to plane; (2) cylinder to plane; and (3) cylinder to cylinder (see Figure 7). In all types of letterpress presses the printing surface must be coated with a uniform layer of ink before receiving the paper. This inking is carried out by a mechanism composed of up to 20 rollers: take-up rollers carry the paste ink from the ink supply to the distributing and sliding rollers, which operate with a longitudinal to-and-fro movement to crush and spread the ink in an even layer on a level or cylindrical metal surface and finally contact rollers that transfer the ink to the printing surface of the typeform.

From J.S. Mertle and G.L. Monsen, *Photomechanics and Printing*

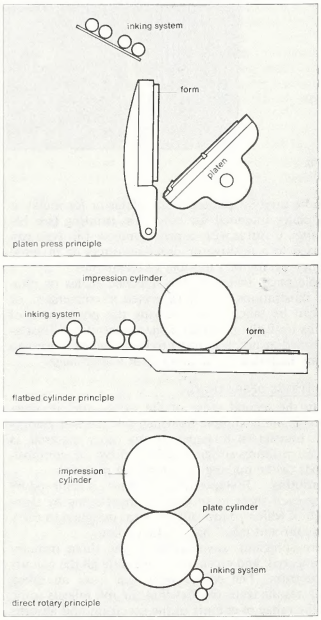


Figure 7: Operating principles of the three principal types of letterpress presses (see text).

The first two kinds of presses are sheet fed. The third can be either sheet fed or roll fed (web fed), depending on the model and type of work. Sheets of paper must be fed into the press in a sequence synchronized with the movement of the press. Most modern presses are equipped with an automatic feeder that functions in conjunction with the movement of the press. The feeders operate by either friction or suction. In a friction-operated feeder, the sheets of a pile of stock on a slightly slanting surface are fanned out so that each sheet projects over the one beneath in such a way that the friction of a cylinder can dispatch them individually one after another toward the feedboard of the press, where three pegs guide each into position. In a suction operation the sheets remain piled up vertically; a wheel brushes a corner of the top sheet and separates it from the others, enabling a compressed-air blower to inject a cushion of air underneath the sheet. A system of vents connected to a suction pipe lifts the paper and carries it towards the press's feedboard. An automatic device lifts the pile of sheets continuously into place. On all letterpress presses, slight irregularities of the surface of the typeform are compensated for by packing the platen with a material soft enough to absorb the irregularities.

As the sheets leave the press, a powder is sprayed onto their surface to form a separative coating that prevents the transfer of ink from one sheet to another. An alternative in modern high-speed presses is the incorporation in the ink of a special quick-drying agent.

Platen presses. Presses that operate plane to plane are called platen presses. A vertical clamping contrivance clamps the bed, which carries the form into which the composed type is locked, and the platen, which carries the sheet of paper while it is being printed. When this clamping contrivance is open, the typeform is inked by a series of rollers that descend and then reascend, and the printed sheet is removed and a new sheet placed in position on the platen.

The pressure exerted during impression is about 40 kilograms per square centimetre (570 pounds per square inch). A platen press can reach a production speed of 5,000 sheets per hour.

Cylinder presses. In presses that operate cylinder to plane, called flatbed presses, a cylinder provides the pressure while the typeform retains its flat surface, generally in a horizontal position. Generally, too, the bed is mobile to allow the typeform, as it moves back and forth, both to pass under the rollers of the inking system and to pass under the impression cylinder around whose outer surface the sheet of paper is wrapped, attached by a set of clamps. Flatbed presses fall into various categories, depending on the cylinder's operation.

The stop-cylinder press

In the stop-cylinder press, a toothed rack incorporated in the bed engages a cogwheel incorporated in the cylinder while the bed is moving forward. As it moves back again, the cogs disengage. A shallow cavity in the cylinder makes it possible for the typeform to be slid underneath. Printing speeds can reach 5,000 sheets per hour.

In the two-revolution press, the cylinder never stops revolving but is raised on its bearings as the bed moves back again in order not to touch the form. In its lowered position the cylinder's cogwheel engages a low-toothed rack incorporated in the bed; in its raised position the cogwheel engages a parallel high-toothed rack, which enables the cylinder to continue revolving in the same direction. Printing takes place during the first revolution; during the second revolution the cylinder runs free. Printing speeds are about the same as for the stop-cylinder press. But, by avoiding the mechanical jerkiness due to the stopping of the cylinder, the two-revolution press has a much smoother, more regular, and quieter action.

On a single revolution press, the cylinder does not stop revolving but must be raised while the bed moves back. Its diameter is twice that of the two-revolution cylinder press, but one half of its surface is hollowed out in order not to touch the form. Printing, then, takes place in the first half of the cylinder's revolution.

A perfecting press is a combination of two two-revolution presses bracketed together. It has two cylinders and a single bed bearing two different type forms. Thus, as it moves back and forth the bed prints two impressions, one for each form. The same sheet of paper is passed from one cylinder to the other and is thus printed on both sides. The padding on the second cylinder is constantly cleaned by a kerosene-coated roller to prevent a transfer from the first side of the paper to be printed. In fact, the second impression is better than the first, and the part

of the job requiring a better quality is reserved for the second cylinder.

The two-colour press also combines two bracketed two-revolution presses, but a small auxiliary drum is interposed between the two cylinders to ensure that a single sheet of paper presents the same side to be printed twice. The typeforms, which are designed to complement one another, are each inked with a different colour.

Contrary to the general design of cylinder presses, the vertical cylinder press is composed of a vertical bed, and both bed and cylinder move vertically with a reciprocating motion, each in an opposite direction. The cylinder revolves only while it is moving up and down, which makes this kind of press similar to the stop-cylinder press.

Printing speeds exceed 5,000 sheets an hour for paper up to about 2,000 square centimetres (300 square inches).

Rotary presses. Presses that operate cylinder to cylinder, usually described as rotary presses, consist of two cylinders turning in opposite directions. The typeform is attached to the surface of one cylinder, and the impression cylinder provides the pressure.

Sheet-fed rotary presses produce the same kind of work as flatbed cylinders, but the size of paper can be slightly larger, and production speed is three times higher using the same size of paper. The inking system is much the same as in flatbed cylinder presses. In both models the sheet of paper is held to the impression cylinder by a set of clamps. On the largest models passage of the paper between the cylinders is controlled only by precise positioning and presentation of the paper.

The two-colour rotary press combines two plate cylinders, each bearing a different typeform and each provided with its own inking system against a single impression system, in a so-called satellite arrangement. Thus, the same side of the same sheet of paper receives two successive impressions of two different colours in a single revolution of the impression cylinder.

The rotary perfecting press uses the same arrangement as the previous machine, but a second impression cylinder of smaller dimensions is interposed between the first impression cylinder and one of the plate cylinders. Thus, the sheet changes sides between one impression and the other and so is printed on both sides.

The two-colour cylinder and flatbed press combines a rotary and a single-revolution press by using the same impression cylinder. Clamped around this cylinder, the sheet of paper first moves against the curved form on a plate cylinder inked by the first inking system; then it moves against a flat form on a mobile bed inked in another colour by a second inking system.

The planetary colour press

Polychrome rotaries permit three, four, or five colours to be printed on the same sheet without handling the pile of paper. Some of these are designed on the planetary principle, using as many plate cylinders as colours, each being fed by its own inking system and grouped around a single impression cylinder. This type of machine is much more popular in North America than in Europe. Others are designed as a row of identical units each printing one colour, the paper moving from one unit to another by means of a transmission drum or a conveyor contrivance.

Roll-fed rotaries are exceptionally large, with high rates of production, and are used almost exclusively for printing daily newspapers. Their principle is extremely simple; the continuous roll of paper drawn from a reel moves between a plate cylinder and an impresssion cylinder. In practice, the cylinders of some have a circumference twice the height of a page of newsprint. Thus, each revolution of the cylinder prints two copies of the same page. On others the circumference of the cylinders corresponds to the width of four pages of newsprint side by side; each revolution prints eight copies of the same page.

In another type, the basic unit of the rotary, called the group, is composed of a symmetrical arrangement of two plate cylinders with their impression cylinders. Within the same group the roll of paper moves from one of the plate cylinders, where it is printed on one side, to the other cylinder, where it is printed on the other side. Each revolution of the machine thus produces a group of two times eight pages of newspaper printed on both sides. Inking is effected by distributing, sliding, and contact rollers that receive ink from dozens of openings distributed across the width of the cylinder, each of which can be precisely adjusted. In each group paper is fed in by a barrel-like device with three axes for supporting the reels, which makes it possible to move over from a finished roll to a new one by a simple gluing operation and a 120° revolution. The reels weigh up to 600 kilograms (about 1,300 pounds). The rotary press is made up of a certain number of identical groups in alignment.

Folding and cutting operations

After printing, the roll of paper is automatically folded down the middle so that on both sides only two pages remain face to face. Moving along a triangle with rounded sides and then passing between rollers, each half roll is folded in the middle to form a newspaper. A cutting mechanism, synchronized with the action of the rotary, separates each newspaper from its neighbour.

Depending on the number of pages of the newspaper, various arrangements are possible: rolls coming from different groups can be added to by accumulation to produce two issues in a multiple of four pages. A roll half as wide as the others coming from one of the groups can be added to the middle of one or more of the usual rolls, producing two issues in multiples of four pages, plus two. Two parallel rollers, called turning bars, arranged at an angle of 45° in relation to the flow of the roll, enable the roll, after it has been printed and folded in the middle, to be folded in half again so that each of the two issues produced by the group is formed into a signature of eight pages.

For colour printing, the same roll of paper moves through several groups in succession, in each of which the plate cylinders carry typeforms appropriate for each of the colours necessary.

Modern rotaries can revolve at the rate of 35,000 revolutions per hour (500 metres [1,600 feet] of paper per minute), or at a theoretical production rate of 140,000 newspapers an hour in two issues from the final group. In practice, the average rate of production is about half this figure.

At speeds such as this, inspection and safety precautions are more reliably carried out by electromagnetic devices based primarily on the use of photoelectric cells; for example, a series of cells on the track, over which a roll of paper moves, react to tearing by stopping the machine.

In colour printing, photoelectric cells ensure matching by selectively picking up the guide marks printed in each colour as they go by and by reacting to any irregularities in the distance between these guide marks. Any error is automatically corrected by modifying either the speed of one group or the pressure of the rollers that control the tension of the paper between one group and the next. Matching up across the width of the paper is controlled by cells that react to any lateral digressions as the paper moves along by controlling a lateral-shifting mechanism.

Quality control of colour reproduction is also carried out by photoelectric cells that emit a current whose strength is proportional to the intensity of impression of the guide marks for each colour. By comparing this intensity with a colour scale, a computer determines the continual adjustments needed in the composition of the inks and controls the opening of valves to add pigment if the inks are too light or colourless varnish if they are too dark.

Preparing stereotypes and plates. The curved shapes with which the printing cylinders of the rotaries are fitted are the stereotypes, or plates. These may be produced from ordinary flat typeforms by reproducing the relief surface of the type copy and of the plates for illustrations made from halftone photoengravings and line engravings. Alternatively, screened photographic illustrations are used to make plates from mounted positives of transparencies by means of photoengraving.

Casting stereotype plates

In making stereotype plates, a flong, or mat, a thin sheet of pasteboard, pliant enough to register an impression and sufficiently heat-resistant to tolerate the molten type metal, is placed on the type form with paper and cotton packing. It is subjected to heavy pressures in a press at a moderately high temperature to ensure that it dries; it retains an intaglio impression of the relief surface of the type form and is placed against the inside wall of a curved

casting box into which a lead alloy is injected. The cooled stereotype plate resembles a rigid shell, solid or ribbed according to its thickness.

The plate is mechanically finished to ensure uniform thickness and to provide properly bevelled edges; metal from nonprinting areas is routed out to avoid ink smudges. Finally, the plate is electroplated with a thin layer of nickel to increase resistance to wear.

Though it is rarely done, stereotype plates can equally well be cast flat, then curved while heated after being finished.

The stereotype process is the fastest and most economical process for obtaining curved plates, but such plates are not suited to the precise matching up necessary in colour printing because of the irregular behaviour of the mat; its performance depends heavily on humidity and temperature conditions.

In making electrotype plates, an impression is made of the typeform using a substance that is a conductor of electricity or can be made conductive by treating with black lead or dusting with a powdered metal such as silver. Any of the following may be used: a sheet of wax submitted to heavy pressures in the press; a sheet of lead submitted to extra heavy pressure; a sheet of Tenaplate, a kind of vulcanized plastic covered with a film of black-lead wax, submitted to pressure only slightly less than that for lead; a sheet of celluloid or a sheet of plastic (Vinylite, tenalite composed of a layer of aluminum between two layers of Vinylite, etc.).

After metallizing to ensure conductivity, these molds are electroplated with a thin copper shell, delicately reproducing the relief surface of the type form. This shell is stripped from the mold and reinforced with a backing of lead alloy poured over its underside. Nickel plating can again be employed to increase its wear resistance.

Electrotype plates can be curved while cold after they have been backed or while hot, during leading, when the lead has not yet solidified. There are also impressions that can be curved before electroplating to obtain curved copper shells. The plate is then reinforced by spraying molten metal against the shell while it is fixed against the inside wall of a rapidly spinning drum.

Making electrotype plates, especially in a curved shape, is a costly process, but they produce the best quality print. An impression made with a sheet of lead is best suited to printing in colour because it is least sensitive to variations in humidity and temperature. Metal-shell plates are attached to the plate cylinder mechanically.

Stereoplastic plates are made in two successive moldings. A hot mold is made in the press from the typeform in a thermosetting material; that is, one that can be melted only once and can thereafter tolerate high temperatures without damage, such as Bakelite. This first molding itself serves as the mold into which is pressed the hot material for the plate—usually cellulose acetate, vinyl resin in sheets or in powder to which a plasticizer has been added to make it more durable, or a sheet of rubber gum that will be vulcanized when pressed. New liquid plastics can also be employed using a method similar to casting.

The plates are trued up by milling or filing to the desired thickness and glued either to the plate cylinder of the rotary or to a metal plate wrapped around it.

Stereoplastic plates are light and easy to use, particularly on small rotaries. Their printing quality is good enough for texts and line illustrations, although they are ill suited for fine-screened halftone illustrations.

Whether wraparound plates (wrapround plates in Great Britain) are made of metal or of plastic, they utilize photosensitive materials in their preparation.

Metal wraparound plates may be of copper, magnesium, or zinc. Only one type of zinc, microzinc, whose molecular structure permits finer prints than ordinary zinc, is used. The plates are covered with a layer of photosensitive material and processed like photoengravings, using negatives of the pages on which the photographic illustrations have already been screened. Because the engraving is only half as deep as letterpress engraving, ink rollers of a larger diameter are used.

The plate can be curved after engraving, but curving is usually done beforehand, both to avoid breaks in the metal and to obtain an absolutely uniform degree of bend in the various plates for colour printing.

Plastic wraparound plates

Plastic wraparound plates utilize the property of photosensitive polymers to lose their solubility in certain solvents when exposed to light. Exposure to light through the negative of a page fixes the insolubility of the polymer and limits it to the areas that are to constitute the printing surfaces. A suitable solvent then eliminates all the nonprinting areas and sets off the type in relief.

New polymers with this property and with new qualities are constantly being perfected. Among the better known examples are nylon, Dycril, and KRP. Nylon is sensitized in bulk by immersion in a solution of acetone containing the sensitizing agent. The plate is exposed to ultraviolet light, and the nonprinting areas are dissolved by a bath of methyl and ethyl alcohol. It takes 24 hours for the plate to attain its maximum hardness. Dycril is sensitized by immersing it for 24 hours in a carbon dioxide atmosphere. The nonprinting areas are removed by sprinkling with a solution of sodium hydroxide. Because it is preferable for the plate to be curved before being engraved, exposure takes place on a rotary drum turning in front of an arc lamp and the bath on a rotary drum turning in a trough. The total amount of time required to make a plate is about 45 minutes.

KRP (Kodak Relief Plate) is a sheet of cellulose acetate that is superficially sensitized by the deposit of a thin coat of photographic emulsion. After exposure to light, this emulsion remains only on the printing areas, which it protects from the action of the solvent. Engraving the KRP plate can also take place on a rotary drum.

The polymer of the plastic wraparound plates is usually mounted on a base consisting of a metal sheet. The depth of the engraving can equal the actual thickness of the polymer. The type therefore stands out in sharp relief. Whether metal or plastic, the wraparound plates are easily and rapidly attached by means of register hooks, which ensure perfect tension of the plate on the surface of the rotary's plate cylinder.

Scope of letterpress. Letterpress printing is characterized by the sharpness and strength of bite of the type. It produces good reproduction of illustrations on sheet-fed machines, with these two reservations, that screening prevents the reproduction of pure white, and it is not possible to use more than four colours without risking a speckled moiré pattern. Roll-fed printing still preserves a good sharpness in the text but produces only an acceptable or average photographic illustration, depending on the quality of paper, and then only in black. Colour printing of illustrations on roll-fed rotaries can hardly rise above the mediocre level, even with the use of special rotaries in which the printing groups are drawn as close together as possible.

Rotogravure. Rotogravure is a system of intaglio printing. It consists of transferring to paper fluid ink contained in the cells of the printing cylinder, while the projecting nonprinting areas on the surface of this cylinder are kept free of ink by constant wiping.

The density of the print at each point depends on the depth of the cell at that point and the quantity of ink it contains, rather than on the printing surface, as in the letterpress process. The screen no longer plays an optical role. It is used to establish the partitions that separate all the cells of the honeycomb from each other and that form a surface of uniform height, while the cells are all of different depths, so that the ink is taken up on the engraved surface in an exactly defined quantity. The screen also prevents the wiping mechanism from penetrating the cells of the cylinder and withdrawing the ink. For this reason, line drawings and even the text type must be screened, as well as the photographic illustrations.

Rotogravure rotaries are thus composed of two principal elements: the cylinder bearing the printing form and the impression cylinder; the paper moves between these two cylinders either in sheets or from a roll.

The nature of the ink used in rotogravure and the operation of the wiping mechanism make the use of plates technically more delicate because of the indentation that

would be necessary in the places where they were affixed to the cylinder. Plates are used only on certain sheet-fed machines, because they are easier to prepare and stock. Generally speaking, the image is etched directly onto the printing cylinder, which must therefore be easily handled.

Ink of sufficient fluidity is used to allow direct inking of the engraved cylinder without the use of a roller to spread and distribute it. While the paper tangential to its upper surface is being printed, the lower part of the cylinder is bathing in the pool of ink. In fast-working machines the ink can be poured on through a spout or sprayed onto the surface of the cylinder, to avoid flying ink. On the plate machines, inking must be done by roller to avoid filling the hollows of the clamping system with ink.

The operation of the doctor blade

Between inking and printing, the wiping mechanism comes into action. It consists of a thin blade of soft steel, the scraper, or doctor blade, which moves slowly to and fro lengthwise. By rubbing against the cylinder with a precisely regulated degree of pressure, it causes the excess ink to drop off before the cylinder moves over the paper.

In sheet-fed machines the impression cylinder is like that of a letterpress press, with a diameter large enough to carry the sheet wrapped around its surface, gripped and released by a system of articulated clamps. Production speed can attain 6,000 sheets per hour.

Roll-fed rotaries generally consist of several printing units in line (up to as many as 18), each composed of a printing cylinder, an inking system, and a hard-rubber impression cylinder of small diameter. Each unit can turn in either direction to allow the paper to move through in any combination.

When printing both sides, the same roll of paper must pass through two successive printing units. In colour printing the same roll must move through as many printing units as there are colours used on each side of the paper. Several rolls coming from different printing units can be added together by accumulation.

Rotogravure rotaries normally include the same folding and cutting systems as letterpress rotaries, along with the electronic devices necessary for controlling the operation. Because of the fluidity of the ink, all rotogravure machines, whether sheet-fed or roll-fed, require a drying process. The paper passes over a heated drum or over sources of infrared rays or through a compartment ventilated with hot or cold air.

Preparing rotogravure cylinders. The printing cylinders or plates are prepared from positives of pages on which neither text nor illustrations are screened. The photosensitive substance necessary for carrying out the photoengraving operation, carbon tissue, is separately packaged to permit it to be treated before being affixed to the metal. Carbon tissue is paper, in sheets or on rolls, coated with a layer of gelatin. This layer is sensitized before use by being plunged into a solution of potassium bichromate. The carbon tissue is first exposed to intense light through a glass plate bearing a transparent screen on an opaque background and then exposed a second time through a positive of the pages. During these two exposures the gelatin is hardened by the light, completely in the white and screened areas, more or less deeply in the halftones of the photographic illustrations, and not at all in the areas covered by text and lines.

The carbon tissue is then applied to the copper surface of the cylinder (or plate). The tissue is then peeled off, leaving the gelatin film fused with the metal. The cylinder (or plate) is plunged into a warm-water bath, which dissolves the gelatin in inverse proportion to the degree to which it was hardened: totally in areas where it received no light like the text and line drawings, more or less deeply in the halftone photographic illustrations, and not at all in the screened areas.

The etching of the copper is begun by sprinkling it with a solution of ferric chloride, which attacks the metal in the areas where the gelatin is totally dissolved and bites into it more or less deeply depending on the thickness of the surviving gelatin that it must penetrate. After being etched, the printing surface can be reinforced by chromium plating.

Improvements on this classic method of preparing the rotogravure plate are continually being made. The carbon tissue may be replaced with silver emulsions on a plastic base; a direct copy without carbon tissue may be made by dusting the cylinder with a photosensitive substance and projecting the image on the surface of the cylinder by optical means or by electronic engraving.

Plates for rotogravure are made of solid copper. Rotogravure cylinders consist of a steel mandrel on which a layer of copper has been deposited by electroplating. After printing, the etching is removed by grinding and a thin film of copper deposited to restore the cylinder to its original diameter. Restoration is simplified if the film can be prevented from adhering to the cylinder so that it can be ripped off after printing. This can be done by coating the surface of the cylinder with a copper–mercury amalgam before electroplating. After plating, the new copper film is then polished. Some electroplating baths produce a shiny copper finish, dispensing with the polishing operation.

Scope of rotogravure. Even in long runs, rotogravure produces quality illustrations with rich colours. It is less suited to printing small typefaces because they are cut up by the screen.

Offset. The first characteristic of offset printing is that the printing and the nonprinting elements are parts of a single continuous surface. The difference is simply that the printing parts of the surface repel water when moistened but absorb the ink with which they are coated, whereas the nonprinting parts absorb the water and repel the ink. The form used in offset printing is a metal plate the surface of which has been prepared to divide it into the parts with these opposite properties.

The second characteristic of offset printing is that the ink is not transferred directly from the printing form to the paper but is first transferred to a rubber blanket. It is the ink retained by the blanket that is transferred to the paper.

The three-cylinder offset press

Offset presses utilize three principal elements: the plate cylinder; the blanket cylinder; and the impression cylinder, which applies the paper, in sheets or a roll, against the blanket cylinder.

The plate cylinder is like the cylinder for wraparound plates on a letterpress rotary, with a groove in which the attaching and tension mechanisms are housed. A moistening system and an inking system are connected to this cylinder. The inking system is similar to that on letterpress presses, with a series of alternating hard and soft rollers that grind, spread, and evenly distribute the ink in a uniform thickness.

The wetting system, which dampens the plate before it is inked, consists of a series of rollers that transmit regularly controlled quantities of water taken up by puddling in a tank or sprayed from rotary brushes grazing its surface.

The blanket cylinder also has a groove containing the mechanism for attaching and controlling the tension of the blanket, which consists of several layers of fabric and rubber. During printing the groove of the blanket cylinder and that of the plate cylinder coincide at each revolution.

On sheet-fed machines the impression cylinder has a cavity in which a system of articulated grippers is recessed. The movement of the machine is synchronized so that at each revolution the slightly projecting grippers enter into the groove in the blanket cylinder in such a way as to cause no damage to the latter. This problem does not arise on the impression cylinders of roll-fed rotaries, and their movement does not therefore have to be synchronized with that of the other cylinders.

In sheet-fed presses the three cylinders, usually placed side by side, are of the same diameter and turn at the same speed. The movement of the cylinders is synchronized by meshing sprocket wheels. Paper is fed in by a feeder similar to that of letterpress presses.

Offset machines can be used to print very large sheets. The fastest can produce 10,000 sheets per hour.

In printing two or more colours, during the offset process the dampness of the plate is partly transferred, via the blanket, to the paper. This slightly alters the paper's dimensions, causing difficulties in register. The separate colours are therefore printed as nearly simultaneously as possible.

On some two-colour machines a single impression cylin-

der applies each sheet successively against two blanket cylinders, each of which receives a transfer from its own plate cylinder.

More generally, sheet-fed offset machines printing in two or more colours are designed as a series of printing units each consisting of three cylinders with their wetting and inking systems, the sheet being transferred from one unit to the next on drums of large diameter with a system of articulated grippers. Size of paper and speed are the same as for the one-colour machines.

Blanket-to-blanket machines

In the offset process an unusual arrangement, without impression cylinders, is possible. Simultaneous printing is done of both sides of the sheet as it passes between two juxtaposed blanket cylinders. Each receives a transfer from its own plate cylinder and prints one side of the sheet while at the same time acting as an impression cylinder for printing the other side of the sheet. Thus, the blanket-to-blanket machine consists of a total of only four cylinders.

The various kinds of machines can be used in combination: for example, three one-colour printing units and one blanket-to-blanket machine can print one side of the sheet in four colours, the other side in black.

Roll-fed rotaries work on the same principle as sheet-fed presses, with two differences—the groove for the plate fasteners on the plate cylinder is very narrow, and the impression cylinder, lacking grippers, has an entirely level surface.

In-line rotaries are designed as a succession of one-colour printing units or, more frequently, blanket-to-blanket units. On these, as on the others, printing in several colours is done by sending the paper directly and horizontally through several successive units.

On blanket-to-blanket rotaries, the back is printed simultaneously with the front. On rotaries made up of one-colour units, a turning bar reverses the direction in which the roll of paper presents itself. Feeding the paper into and taking delivery of it from the machine can be combined so that the rolls are superimposed to form a signature.

Satellite rotaries (drum presses) work on the principle of a single impression cylinder of very large diameter that has arranged around its outer surface four, five, or sometimes six blanket cylinders. Each receives a transfer from its own individually moistened and inked plate cylinder. Thus, colour printing one side of a roll of paper is accomplished in a single revolution of the impression cylinder.

Drum cylinders are most often made up of two symmetrical groups printing one side after the other. For quality colour printing with a heavy thickness of ink, rapid drying is necessary to prevent smudging. The roll therefore moves horizontally into a drier heated by a ramp of gas burners or hot-air blowers and is cooled as it passes over metal drums refrigerated by circulating water.

Offset rotaries work at a speed of 15,000 to 20,000 revolutions per hour and have the same cutting and folding systems as letterpress rotaries.

Preparing the plates used in offset involves defining and fixing the joint presence of two mutually repellent materials, one water-receptive for the nonprinting areas, and the other ink-receptive for the printing areas. As in letterpress the screen is necessary to translate the halftones of photographic illustrations into surface densities.

Even though there is no question of creating reliefs, the preparation process is very similar to that of photoengraving. But, because the blanket intervenes between plate and paper, text and illustrations appear on the same offset plate in straight rather than in reverse reading as on letterpress plates.

Offset plate making. Depending on the work they are intended for, various kinds of plates and processes for preparing them exist.

Monometal plates are made of a metal with hydrophilic (water-receptive) properties, such as zinc or aluminum, the surface of which is treated to make it more porous. The plate is coated with a thin layer of a photosensitive substance, covered with a negative of the texts and previously screened photographic illustrations, and exposed to intense light. This hardens the photosensitive substance in the areas of the negative where the light passes through; that is, in those corresponding to the printing areas. The photosensitive substance is then washed away from the nonprinting areas, where the metal, now stripped, is wetted; the hardened photosensitive substance is inked.

Presensitized plates are monometal plates coated with a photosensitive layer that has a life-span, away from light, of up to six months. Certain presensitized plates for use in short runs can be made of paper or plastic.

Deep-etch plates are monometal plates prepared by inversion. The plate is coated with a photosensitive substance and then exposed to light through a positive of the text and screened photographic illustrations. This exposure hardens the photosensitive substance on the nonprinting areas. Washing eliminates the photosensitive substance on the printing areas, exposing the metal. The plate is then given a mild acid bath, which etches the metal of the printing areas to a shallow depth. An ink-receptive lacquer is spread over the surface of the plate; that is, over the hardened photosensitive substance wherever it still survives, as well as over the etched-metal areas. The hardened photosensitive substance and the lacquer are both removed at the same time by dissolving and brushing, exposing the metal of the nonprinting surfaces to be wetted. Lacquer remains in the cavities of the etched areas, ready to be inked. Plates treated by this process are suitable for longer runs (up to 250,000 copies).

Bimetal or trimetal plates are composed of two superimposed metals, one of which is hydrophilic (aluminum, stainless steel, chromium, nickel) and the other ink-receptive (copper, bronze). Whichever of the two covers the other in a microscopic film is partly eliminated by the photoengraving operation. In the case of a hydrophilic metal, photoengraving is carried out using positives of the text and illustrations. In the opposite case, negatives are used.

Among the combinations used in various commercial plates are chromium on copper or nickel on bronze for use with positives and copper on stainless steel or copper on aluminum for use with negatives. The combination of two metals can, in certain plates, take the form of a double film deposited on a third metal (steel or zinc), which simply serves as a base. Bimetal or trimetal plates are very durable and can be used to produce as many as 500,000 copies.

Alongside these various types of conventional offset plates, special processes are being developed that tend to further simplify the work of preparation. Electrostatic (or xerographic) transfer plates were developed from the principle that certain materials (such as selenium) that are insulators while in the dark become conductors of electricity when exposed to light. A selenium plate is given a positive electric charge in the dark and then exposed to light through the positive of the text and illustrations. In the areas touched by the light, the plate becomes a conductor, and the positive charge disperses. A fine, negatively charged powder is sprinkled over the plate and is attracted by the positive charge in those areas of the plate where it still remains. The image thus revealed is transferred to an aluminum plate, which is positioned on the selenium plate and then given a positive electrical charge. The powder transferred to the aluminum plate is fixed by heating to constitute the ink-receptive printing surface.

Innovations in offset plates

This whole operation, which is carried out automatically in a compact apparatus, takes only three minutes. Xerographic plates are used only on small machines.

"Immediate" offset plates consist of a polymer layer that is sensitive not to light but to heat, which makes it become hydrophilic. Those areas not touched by heat retain the opposite property. The plate is then ready for printing without further treatment.

Scope of offset. In offset printing the letters print a little less sharply and strongly than in letterpress, except on special qualities of paper. The reproduction of photographic illustrations, in black and in colour, also depends on the quality of the paper and may rival rotogravure reproduction when printed on offset rotaries with driers.

Other printing processes. *Letterset.* This process, also called indirect letterpress, or dry offset, is a combination of letterpress and offset. Like offset, it has a blanket cylinder as the transfer element between typeform and paper, but, like letterpress, it uses a relief typeform. Thus, letterset

presses use the same three cylinders as offset, but there is no dampening system.

Because of the use of a blanket, the text and illustrations appear on the typeform straight and not reversed. Consequently, the use of plates is excluded, since these would be made by making a mold of a conventional typeform. On the other hand, all types of metal or plastic wraparound plates, prepared using positive or negative transparencies, can be used. These plates are thinner than those used for letterpress, and the height of their relief is less. Rollers of a quite large diameter must be used for inking, and contact between plate and blanket cylinder must be extremely light.

All the combinations that the blanket allows on offset machines are equally possible on letterset machines: blanket-to-blanket and drum press. Machines are manufactured that can be used interchangeably, both for offset and for letterset; in the latter case, the action of the dampening system is suspended. The size of paper and performance rates are the same in letterset as in offset.

Serigraphy (screen printing). Serigraphic printing consists of forcing an ink, by pressing with a squeegee, through the mesh of a netting screen stretched on a frame, onto the object to be printed. The nonprinting areas of the screen are protected by a cutout stencil or by blocking up the mesh.

Silk and other screens

The screen is usually made of fine but strong silk gauze (available in different mesh sizes) but can also be made of synthetic gauzes (nylon, tergal) or of wire gauze (phosphor bronze, stainless steel, nickel) or of combinations (nylon–copper, nylon–bronze).

Preparation can be carried out by hand. The design to be reproduced is transferred to the screen by drawing with a benzene-soluble ink, spreading glue over the whole surface of the screen, then dissolving the ink so that the glue remains only on the nonprinting areas. The nonprinting areas can also be covered with glued paper or film, cut out in the shape of the desired image and attached either by heat or by a solvent.

Direct or indirect photomechanical processes are being more and more widely used, however. In direct photomechanical processing the screen is covered with a photosensitive layer and then exposed under a positive, the photographic illustrations having been screened previously. In indirect photomechanical processing, the printing and the nonprinting areas are prepared by exposure under a positive (screened, if necessary) of a photosensitive film (carbon tissue, the same as that used in rotogravure, or presensitized film), which is then bonded to the screen.

Screen printing is still done largely by hand, the frame being lifted up after each operation; however, semiautomatic and automatic machines that are driven mechanically or by compressed air can carry out the series of operations: positioning the object to be printed, moving the frame, inking, spreading the ink with the squeegee, delivering the printed sheet or object, and transferring it to a drying apparatus.

Serigraphic printing can be applied to a wide variety of surfaces—paper, cardboard, glass, wood, plastic, posters, bottles, electronic circuits, etc.—and to varied shapes: for printing cylindrical objects, the squeegee is stationary, while the screen revolves under it at the same time as the object to be printed.

Modern machines reach speeds of from 1,000 to 6,000 copies per hour.

Collotype. A photosensitive layer, spread on a plate of glass and exposed under a negative, hardens and loses its hydrophilic properties in the areas where it receives light and in proportion to the intensity of the light received through the halftones of the photographic documents. It is then able to retain both ink and water in all areas, the ink repelling the water in inverse proportion to the intensity of the exposure.

Collotype printing, which is based on this property of mutual repulsion, is thus related to lithography. But it is also related to rotogravure in the fact that the thickness of the film of ink is not uniform but in proportion to the shades of tone in the original image. Collotype, the only printing process that can reproduce photographic documents without a screen, is characterized by its fidelity of reproduction.

Collotype presses consist essentially of a bed bearing the glass plate prepared for printing, an impression cylinder carrying the sheet to be printed on its outer surface, and a system of ink rollers. The photosensitive preparation retains its own moisture, so no dampening system is necessary. Printing speed is low, rarely more than 200 copies per hour, and the useful life of the printing surface is very limited, from 2,000 to a maximum of 5,000 copies. Increasingly, the glass plate is being replaced by a cellophane film as the surface to which to apply the photosensitive layer. The prints can be cut out and, when glued onto wooden or metal blocks of type height, can be placed alongside a typeform for a limited number of copies to be printed from a flat surface.

Speed of collotype printing

Collotype is used for printing limited editions of works that require an excellent quality of photographic reproduction in one or more colours, such as reproductions of documents or pictures, posters, and transparent illustrations for advertising or various artistic uses.

Flexography. Based on letterpress-printing principles, flexographic presses are composed of the same basic elements as letterpress cylinder-to-cylinder presses; that is, the impression cylinder covered with a rubber packing and an inking system, which is simplified, owing to the fluidity of the ink used. Although several sheet-fed models exist, flexographic presses are usually roll-fed rotaries that can reach large dimensions and high printing speeds. They consist of a group of several identical units that print an equal number of colours (up to eight). Flexography provides economical printing, either of solid lines or with quite a coarse screen, on unfinished surfaces, wrapping paper, cardboard, plastic film, and it is also used for printing newspapers and magazines.

Electrostatic printing. Electrostatic printing is a process of printing without contact, without a typeform, and without ink. The paper is coated with a very thin layer of zinc oxide, which makes it an insulator while in the dark and a conductor of electricity when exposed to light. This paper is given a negative electrical charge in the dark. It is then exposed by light projected through a positive film of the document to be reproduced. The zinc oxide layer becomes conductive wherever it is illuminated, and the negative charge is dissipated in those areas that correspond to the blank surfaces of the document. Finally, the paper goes through a bath containing pigmented particles that are attracted by what remains of the negative charge and are fixed by drying.

Electrostatic machines have been designed for printing geographic maps. These are composed of five successive units, each carrying out the same complete cycle of processing the paper to produce an edition in five colours at speeds of about 2,000 copies per hour. Improvements in the bath of pigmented particles are making possible the application of the electrostatic process to printing small books.

Printing inks. Printing inks contain three components: the vehicle, the colouring ingredients, and the additives. The vehicle, responsible for transferring the colouring ingredients from the ink fountain to the typeform, can be either a vegetable base (linseed, rosin, or wood oils), which dries by penetration and oxidation and at the same time ensures fixation, or a solvent base derived from kerosene, in which case drying takes place by evaporation. The colouring ingredients come in several forms: pigments, which are fine, solid particles manufactured from chemicals, generally insoluble in water and only slightly soluble in solvents; agents made from chemicals but soluble both in water and in solvents; and lacquers, obtained by fixing a colouring agent on powdered aluminum. The additives stabilize the mixture and give the ink additional desirable characteristics. The nature and proportions of the ingredients vary according to the printing process to be used and to the material to be printed. The proportions must be checked and sometimes modified during printing.

Letterpress and offset use greasy inks. For printing on sheet-fed presses, thick greasy inks are used in which the vehicle is generally made of vegetable oils with the addition

of hard natural or synthetic resins dispersed in mineral oils. Roll-fed rotaries use fluid greasy inks in which the vehicle is made up of heavy mineral oils.

The colour black is generally obtained from an organic pigment, carbon black, derived from the incomplete combustion of oils or of natural gas. Coloured pigments are inorganic compounds of chromium (yellow, green, and orange), molybdenum (orange), cadmium (red and yellow), and iron (blue).

Inks for offset are more highly coloured than those used in letterpress, because they must be transferred to the blanket before they reach the paper. Furthermore, the pigments must resist being picked up by the water from the dampening system.

Special quality inks

Inks with various special qualities exist for both letterpress and offset. In high-gloss inks, the vehicle is not homogeneous, as with ordinary inks, but heterogeneous, based on synthetic resins dissolved in a solvent, with lead and cobalt additives. This ink glazes as it dries. When printing several colours, the whole series of operations must be finished before the ink has time to dry so that the inks can attach themselves to the surface.

Quick-setting inks utilize a vehicle that also has a base of resins dissolved in a quick-drying solvent.

Heat-set inks require the application of heat to facilitate both the oxidation process and the evaporation of the solvent, as well as the penetration of certain elements that had rendered the ink more fluid. Cold-set inks are hardened by chilling after printing, having been kept fluid by heat until they were applied to the typeform.

Moisture-set inks become fixed when they are applied directly to damp paper or upon exposure to a water-spray after they are applied to dry paper. In such inks, which are used more in the United States than they are in Europe, the vehicle is a solvent, soluble in water, that, on contact, penetrates the paper, leaving the pigment on its surface. Odourless moisture-set inks are used for printing food packaging.

Among other special-characteristic inks are metallic inks containing powdered copper, bronze, aluminum, or gold mixed with the pigment; magnetic inks, containing a powdered magnetized iron mixed with the pigment for "recognizing" the shape of printed characters as they pass before electronic reading equipment; and fluorescent inks.

Rotogravure uses fluid inks in which the colouring agent, fixed on a natural or synthetic resin, is integrated in a fluid solvent to which, just before printing, a second, extremely volatile solvent is added.

Flexography also uses fluid inks whose pigments or colouring agents are dissolved in pure alcohol, in alcohol solutions, or in water.

Serigraphy uses inks of extremely varied consistency, depending on the surface to which they are to be applied; some are little different from ordinary paint, except that their composition must not be such that rapid drying would clog the mesh of the screen. (Ro.L.)

Typography

Typography is concerned with the design, or selection, of letter forms to be organized into words and sentences to be disposed in blocks of type as printing upon a page. Typography and the typographer who practices it may also be concerned with other, related matters—the selection of paper, the choice of ink, the method of printing, the design of the binding if the product at hand is a book—but the word without modifier most usually denotes the activities and concerns of those most involved in and concerned with the determination of the appearance of the printed page.

Thus understood, there was by definition almost—but not quite—no typography before the invention of printing from movable type in the mid-15th century; and, thus understood, it is only by analogical extension that the term can be applied, if ever it can be, to "reading" in which the material at hand is something other than words that remain stationary on flat firm surfaces. The electronically created letter that lives out its brief life while moving across the face of a signboard or a cathode-ray tube is not a typographic item. Typography, then, exists somewhere between the extreme of manuscript writing, on the one hand, and the transient image on the electronic device, on the other hand. Whether the letter be made by metal type or photographic image is no longer important in defining the subject; whether the finished item is a book or a page influences its inclusion as typographic not one bit.

THE NATURE OF TYPOGRAPHY

Typography as a useful art. An overview of typography suggests that a number of generalized observations may be reasonable:

Relationship between typography and communication

First and most important, typography and printing, the mechanical processes by which the plans of the typographer are realized, are useful arts. Though there is indeed fine typography, typography is not a fine art. Books, the primary source of typographic examples, are written in the main by people with something to say; they are selected for printing in the main by publishers who see merit and hope for profit in disseminating the statements of the writers to an audience; properly they are edited and designed and printed in the main by craftsmen whose boundaries are fixed for them by considerations germane to the needs of the writers to communicate and the needs of the readers to understand and appreciate. The typographer exists not to express his own design preferences, his own aesthetic needs, but to provide a useful (because usable) connection between someone with something to say and someone to say it to.

But to say—as did the late Beatrice Warde, one of England's great typographic authorities—that printing ought to be invisible is not to say that the typographer has no contribution to make; to say that typography is a functional art and as such ought not to get between the writer and the reader is not to say that there is only one solution to every typographical problem, that aesthetics, taste, personal judgments, and imagination cannot find room for expression in the typographic studio.

Nonetheless, there are limitations to what the typographer may and may not do; for, in addition to being a useful art with the generally accepted first use of transmitting information, typography for at least three reasons is a secondary art.

First, it is secondary in that its basic materials, the alphabets or other similar notational systems with which it works, are not of its own invention. The influence of this fact on the art form is obvious. Generally speaking, Western writing, or printing, is accomplished by the use of a relatively small number of individual letters capable of being grouped in almost infinite numbers of meaningful permutations. Even in the face of language differences, there is a wide carry-over of letter shapes and typefaces from one language to another. Because the number of images (letters) to be designed is limited and entirely manageable, the type designer's job is made less difficult. The language carry-over makes possible the establishment of meaningful typologies, the evolution of international styles and conventions, and the development of criteria and traditions of taste by which typographers improve their work. As the result, it is fairly certain that in a little more than 500 years of printing history since Gutenberg, at least 8,000 and very probably 10,000 or 11,000 typefaces have been designed. The practicing typographer has, then, a vast number of types to choose from, and, because the best of those types have evolved within cosmopolitan traditions and have stood the test of judgment by many people in many places over many years, there are, within the several thousands of types available, many that are of unquestioned excellence.

By way of contrast, the Japanese method of writing and printing involves a combination of systems—some 3,000 *kanji* (symbols based on Chinese characters), *seicho* (based on the brush-written Kana), and two groups of phonetic symbols (*hiragana* and *katakana*), each of which consists of 46 separate symbols. The problem of individually designing some 3,000 symbols, some of them of incredible complexity, is not one that many designers are able to surmount in a lifetime. As a result, to all intents and purposes, Japanese typographers have had only two typefaces

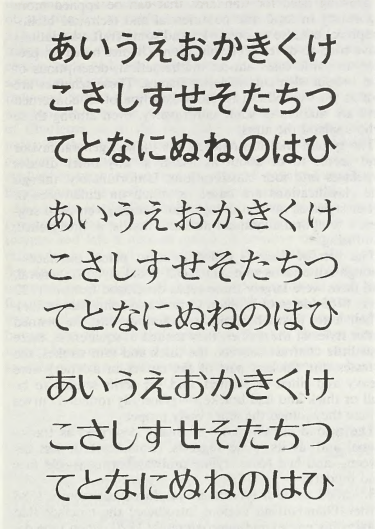


Figure 8: *Three Japanese typefaces.*
(Top) Gothic, (centre) *mincho*, and (bottom) Typos.
By courtesy of *Visible Language*, c/o The Cleveland Museum of Art, Ohio

to choose from—*mincho*, roughly equivalent to the West's roman, and Gothic, functionally a Japanese sans serif. In the 1960s a group of Japanese designers produced a third typeface called Typos (see Figure 8).

Reading conventions and typography

Second, the typographer is limited by reading conventions over which he has little or no effective control. The appearance of a book page, whether well or badly designed, is governed more by the fact that Western readers begin at the top left of a page and read right, a line at a time until they get to the bottom, than it is by the aesthetic desiderata of the designer. Many typographers have long been attracted to the clean and uncluttered look of so-called sans serif type (the two little bases on which the vertical elements of the lowercase "n" rest are serifs, as is the backward pointing slab atop the lowercase "i" or "l," and sans serif types are those in which such embellishments are lacking [*T 1*]). But the difficulty is that almost every study ever completed has indicated that sans serif type is less easy to read in text than is type with the serif. It may well be that if Western texts were printed vertically, from the bottom of a page to the top and read upward so that each letter occupied a separate line with no horizontal connection to those before and after it, the apparent advantage of serif types in this regard might disappear.

To consider another example of the restrictions put on the typographer by the necessity of working with the reading conventions, it is arguable that the appearance of the printed page would be changed and one of the petty annoyances of reading—"doubling," in which the eyes finish a line and then return to the left margin and begin the same line all over again—could be eliminated if people could be persuaded to accept the following reading pattern:

Typography as an art is concerned with the design, into organized be to forms letter of ,selection or words and sentences to be disposed in blocks of type .page a upon printing as

or

Typography as an art is concerned with the design, otni dezinagro eb ot smrof rettel fo ,noitceles ro words and sentences to be disposed in blocks of type .egap a nopu gnitnirp sa

But the fact, of course, is that the problems involved in winning acceptance of any change as fundamental as this one would appear to be so numerous and so substantial as to preclude its further consideration by any but writers of typographical journal articles or textbooks. Basic changes in the format of written text attributable to the typographer or, his earlier form, the printer-typographer, have been few, though occasionally dramatic, as in, for example, the practice of separating succeeding sentences with periods or of separating paragraphs (which in handwritten manuscripts were separated only by the insertion of the scribe's paragraph mark without the initiation of a new line or indentation).

Third, it would appear to be reasonable to call typography a secondary art because, just as the typographer uses letter forms and reading conventions over which he has had little control, so too what he contributes comes into being only through the intervention of a mechanical process that, as often as not, in the 20th century at least, has become the province of the printer, so that the typographer practices his art at least once removed from its final production. The extreme example of the consequence of such a situation may have been seen in the early years of computer-generated typefaces in which, many felt, most faces revealed quite clearly that they had been developed by specialists whose first capabilities were not in the field of typography. And when typographers were later introduced into the process, they found that they had to work through the electronics expert, even as, for many years, those unable to cut their own type had been forced to work through typefoundries.

Dependence upon a mechanical process

It will already have become apparent that there is, at the worst, some confusion and, at the least, some lack of uniformity involved in talking about typographers and typography. The words themselves are of relatively recent origin and have been used self-consciously in their contemporary sense only from about the mid-20th century. The difficulty is, of course, the matter of the process involved. Gutenberg was his own typographer. It may well be, in fact, that his major personal contribution to the invention of printing was the development of a way to cut and cast type so that after the shape of the letter had been fixed and the molds prepared each letter form might be replicated over and over again in one relatively simple process. He was also the publisher, who undertook to risk capital in the selection and preparation of material to be printed for sale; he was presumably the man who designed the layout of each page; he may have done whatever editing was required, and he certainly either printed or supervised an assistant in the printing of the finished product. In the course of years many of the functions at first performed by one man came to be divided among several. Quite early, some printers employed men to cut type to their design; others employed men to design and cut the type; some held their services out for hire to others who became publishers; editors were separated from the process, though not always from decision-making roles in the appearance of the final product. After the introduction of bound volumes, trends were initiated that led eventually to the creation of binding designers as separate artists; it became not uncommon to find persons performing services as book designers and, as such, responsible for coordinating and leading the work of type designers, layout artists, binding designers—all who were in any way responsible for the appearance of the book as a whole.

Function of the book designer

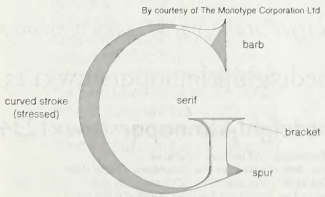


Figure 9: Typeface nomenclature.

The situation became further clouded by the great variety of practice as to the status of each of the persons who perform one or more or, in some cases, all of these functions. They may be professionals retained by printing concerns for a single project; they may be full-time members of a corporate printing staff; in some very few cases they may be a single artist-craftsman-patron carrying out all of the functions in operations (usually necessarily small) devoted self-consciously to the production of "fine books."

Parenthetically, it is significant to note that, in general, the major examples of really fine typography—the significant developments that have raised the possibilities for the improvement of the typographic arts and, in fact, a preponderance of the typographic examples held up as outstanding—have been produced by publisher, printer, and typographer all working within the normal day-to-day requirements of their regular operations. Such a statement must not, however, be taken as dismissing the outstanding services rendered by the best work of the so-called private presses and by the valued demonstration volumes produced in limited numbers by major presses such as the Cambridge University Press in England, in holding up to view the best that the craft is capable of and thus serving as models for the craft itself.

Aesthetic qualities of the typographic page. Confusion notwithstanding, the typographer as he is most generally understood is responsible—whether or not he does all of the work himself—for the appearance of the printed page, and his work is best seen in the several examples of the printed page that are used to illustrate the present article.

The typographic page may be considered in terms of two aesthetic qualities. The first of these has been called "atmosphere," "feel," "impress," "sense," and other similar terms. It is easier felt than defined, and it depends in large measure on such things as the size of the block of type, its placement on the page, the kinds of display letters used for titles, running heads, and subheads, and the size of the margins—all elements that in the hands of a competent typographer create an expectation regarding the contents (possibly even the purpose) of the page and lead to a sense of the time of its production, its seriousness, and its function.

The second aesthetic quality is that of colour, the darkness or lightness of the block of type sensed somehow as a whole rather than as a collection of individual letter forms with substantive meanings. Colour is the result of letter shapes, distances between letters and between words, the amount of space left between lines, the inking of the type, the printing process employed in making the impression on paper, and the paper itself.

Design of the letters

Of all elements the design of the letters is, in the dominant view, the most important. It is important that early typography was in fact overtly engaged in the explicit search for typefaces that would mechanically reproduce the written scripts in which, before the invention of printing, books had been prepared. For most of its life since the invention of printing with movable type in the mid-15th century, typography in the West has been dominated by three type families—roman, italic, and black letter (see below). All are easily recognizable as refined and regularized versions of letter styles first developed and standardized by scribes. The debt of sans serif, more a subclass than a family, is apparent but less unequivocal.

To divide the several thousand typefaces that have existed since Gutenberg into three major families is only the grossest type of classification, and historians of typography, like teachers of typography, have found it useful to set up other classifications. Unfortunately, but not unexpectedly, they have not found it possible to agree on a system of classification. The variety of proposals has been so bewildering as to defeat one of the first purposes of classification. Some have concentrated on the influence of seminal workers in the field and talk of "Caslon's," "Bodoni's," etc. Others have concentrated on the major uses or types of literature from which various dominant types have come and talk, for example, of Humanistic faces. Other systems have employed nomenclatures that emphasize the early manuscript models from which a face evolved or a national influence. In attempts to meet a growing need for standards that can be applied more rigorously in legal and commercial and technical bibliographical uses, some countries and some craft associations have tried to develop classification schemes based on precise, in some cases almost mathematical, descriptions of the various elements of letter shapes. These schemes are not as yet worked out to the satisfaction of all concerned and are matters of some controversy, even among those who applaud the aims.

The system that has enjoyed the longest general favour and seems to be about as useful as any other divides typefaces into four classifications. Unfortunately, though the classifications are based as much on differences in letter styles as on chronology, the names given each suggest a temporal alignment and one that is, at first glance, confusing.

The first of these is called in most places old face—though Americans sometimes call it old style. In general, old faces were largely those types developed from *c.* 1722 to *c.* 1763 (dates of William Caslon and John Baskerville). Their letter forms had marked affinities with the penned letter styles of the scribes: they tended to squareness, there was little contrast between the thick and thin strokes, the stresses (the thickest part of the curves in a letter) were heavy and tilted or slanted, and the serifs tended to be full or thick and had brackets—gracefully rounded curves where they joined the letter body proper.

The second classification is usually described as transitional and, as its name suggests, came more or less between—and had letter styling midway between—old face and modern.

So-called modern types were produced between *c.* 1788 (when Giambattista Bodoni introduced the typeface that retains his name) and approximately 1820, when type design almost everywhere went into a major decline. Modern faces in general share characteristics resulting from the engraver's tool rather than the pen. There is a marked contrast between the thick and thin strokes of the letter, the thins in particular being almost exaggeratedly thin, and the change from thick to thin is sudden and pronounced; serifs tend to be thinner and somewhat longer, and the stresses are less pronounced and are vertical rather than tilted. The effect is that of a letter less square, more up and down, than old-face letters, with lines more sharply defined.

Old style faces

The classification called old style or, in America, modernized old style is reserved for revivals of old faces undertaken by contemporary type designers, a practice typified by the important work of one of England's truly prestigious type designers, Stanley Morison. It is necessary to establish a special classification for such types because in each case the modern reworking, while it owed much to the original, was in fact sufficiently different to be a new creation in its own right.

It is worth pointing out once more that typefaces are assigned to one or the other classification on the basis of the style in which their letters are drawn and not—despite their time-oriented designations—by the year of their creation.

Finally, and although this is not the place for a detailed history of the evolution of the written letter forms on which printing depends, it is significant that models for all

(A) Quod cū audiſſet dauid: deſcendit in

(B) *"Old style" italic, one of many versions pre-1700*

(C) Aabcdefghijklmnopqrstuwx1234MQ

(D) Aabcdefghijklmnopqrstuwx1234MQ

Figure 10: *Traditional and modern typefaces.*
(A) Black face, line of type from the Gutenberg 42-line Bible, 1456. (B) "Old style" italic, pre-1700. (C) Roman by P.S. Fournier, revised as Monotype Fournier, 1925. (D) Gill Sans, E. Gill, 1928.

three major type families were in use before the invention of printing by movable type: in England and Germany, a handwriting that was symmetrical, elongated, spiky, magnificently decorative, and difficult to read, not unlike certain parts of letters called today Old English, German script, and Gothic, was to be the origin of black type; in Italy and Spain, a free, open, square, uncluttered writing, not too far from the letter forms regularized by a decree of Charlemagne in the late 8th century, is recognizable today as the seed source of roman type; and a slanted, cursive, more hurried form of the same—from which it evolved by chancery scribes whose work required speed in handwriting—is easily seen to be the origin of italic.

The ready availability of serviceable letter models freed typography from the necessity of creating its own prototypes and left it able to spend its creative impulses in other ways. So well did it succeed that, within the first few decades after Gutenberg, it had brought forth almost every major development that it was to contribute and, in fact, had established itself so well that it may be fair to say that, until the 20th century, the art was a static one for all but the first 50 years of its existence. Further, the fact that the art form could take its basic ingredients from existing sources gave it a stability that it would not otherwise have had. Since it was unnecessary to wait while various producers in various countries came to agreement on which letter shapes to adopt or what reading conventions to employ, the typographer was enabled to get on with the work of overseeing the printing of books for distribution on a scale never before envisioned. The influence on the Renaissance was of incalculable importance if, indeed, it was not one of cause and effect.

HISTORY OF TYPOGRAPHY

Type, from Gutenberg to the 18th century. Whatever else the typographer works with, he works with type, the letter that is the basic element of his trade. It has already been said that there have been but three major type families in the history of Western printing: (1) black letter, commonly and not quite rightly called Gothic by the English; (2) roman, in Germany still called by its historical name of Antiqua; and (3) italic. All had their origin in the scripts of the calligraphers whose work printing came ultimately to replace.

Calligraphy is dealt with at length in other articles (see further WRITING). It is necessary here only to provide a context for the evolution of the typefaces of the printer's font. The basic letter forms of the Latin alphabet were

By courtesy of the Newberry Library, Chicago

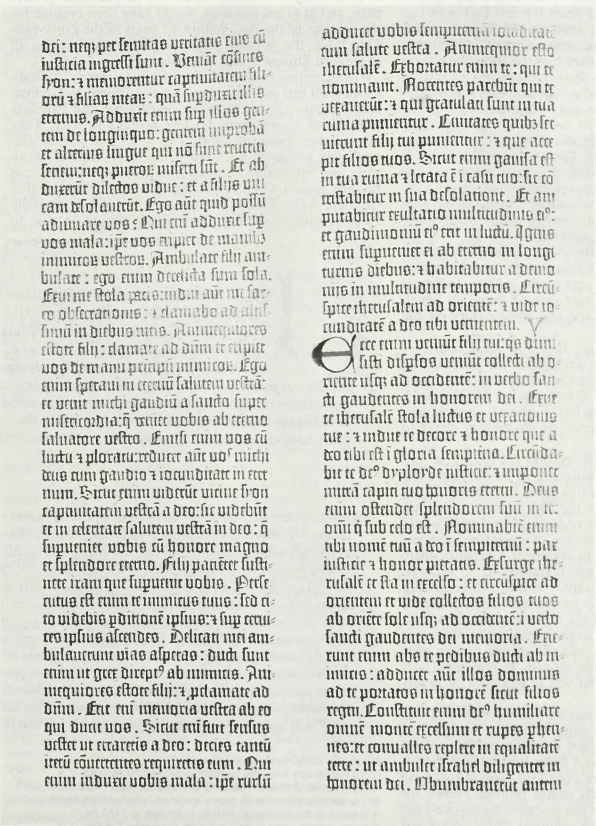


Figure 11: A page from the Gutenberg 42-line Bible, 1456.

established by the classical imperial capital letters of 1st-century Rome. Lowercase letters emerged only slowly, with their most vigorous development coming between the 6th and 8th centuries.

Charlemagne, in order to encourage standardization and discourage further experimentation, ordered his educational program for the Holy Roman Empire to be written in a script consisting of roman capitals and a specific form of minuscules (lowercase letters) known as Caroline. The uniformity thus achieved was short-lived. Under the impact of the national and regional styles of the scribes who worked with the alphabet, the letters—clear, simple, and somewhat broad by today's standards—were gradually compressed laterally, until, by the 11th century, the curves had been converted to points and angles, and the body of the letter had been made thinner while the strokes of which it was composed had been made thicker. This was black letter. By the 15th century it had completed its evolution into the formal, square-text Gothic letter.

First model for printer's type

It was this formal black letter that provided the first model for printer's type when printing was invented. It served well in Germany, but when printers in Italy, in part under the influence of the Humanist movement, turned to the printing of Latin texts, they found the pointed stateliness of the Gothic letter out of keeping with the spirit of Humanism. For these works, they went back in calligraphic history to a time when the text had been less open than the first Caroline alphabet but more rounded than the narrowed, blackened, and pointed Gothic that it had become. When the printers Konrad Sweynheim and Arnold Pannartz in Subiaco, Italy, brought out an edition of Cicero in 1465, they used a typeface that was explicitly intended to be, but was not, a printed copy of the text of Cicero's own time. To distinguish this type from the Gothic that was more "modern" in the 15th century, the Italians called it Antiqua. Known today as roman, it spread rapidly throughout western Europe except in Germany, where the Humanist movement was blocked by the counter-impulses of the Reformation. There, Gothic type was accepted almost as a national typeface until 1940, when its discontinuance was ordered.

It is notable that the majority of early printers continued for many years to use the Gothic type for non-Humanist texts, ecclesiastical writings, and works on law. In Spain, for example, Jacob Cromberger printed books in which the text was set in roman type and commentary on the text was set in Gothic.

Like the Gothic and roman, the third great family of types had its origins in the writings of the scribes. The italic and the Gothic Schwabacher, which serves as a kind of italic to Fraktur (as black letter is known in Germany), both had their genesis in the fast, informal, cursive, generally ligatured letters developed by chancellery clerks to speed their work.

The early years. The 11th edition (1910–11) of *Encyclopædia Britannica,* not uniquely in this day, gave the honour of inventing the printing press to Laurens Coster of Haarlem. Later research in the 20th century, which has more or less become common consent, gives it to Johannes Gutenberg. Actually, the amount of invention involved in the development is open to argument. Certainly, there was in the air at the time much interest in an artificial method of reproducing calligraphic scripts, and books had already been printed from blocks; the techniques necessary to the punching of type and the making of matrices from which to cast it were known to the metalsmiths; paper was replacing vellum; and wine, oil, and cheese presses were readily available as adaptable models. It remained only for someone to combine what was in existence or clearly capable of creation.

Gutenberg began his experiments around 1440 and was ready to put his method to commercial use by 1450. In that year, facing the need (not unknown to later printers) for financing, he borrowed from Johann Fust. About 1452 he borrowed once more from Fust, who at that time became his partner. The only extant printing known for certain to be Gutenberg's is the so-called Forty-two-Line (the number of lines in each column) Bible, completed in 1456, the year after Fust had foreclosed on his partner and turned the business over to his own future son-in-law, Peter Schöffer. Experts are generally agreed that the Bible displays a technical efficiency that was not substantially bettered before the 19th century. The Gothic type is majestic in appearance, medieval in feeling, and slightly less compressed and less pointed than other examples to appear shortly.

Germany: the Gutenberg Bible

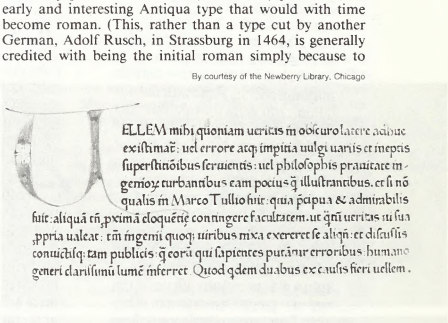
The Forty-two-Line Bible, like the other works of its day, had no title page, no page numbers, no innovations to distinguish it from the work of a manuscript copyist—this was presumably the way both Gutenberg and his customers wanted it.

Some five years later, also in Mainz and quite possibly from the re-established printshop of a refinanced Gutenberg, there appeared the *Catholicon,* notable among other reasons for its early use of a colophon, a tailpiece identifying the printer and place of printing, and for the slight condensation of its type—a move toward more economic use of space on the page and greater type variety in printing.

While not all early results of the printer's art were accepted in all quarters (in 1479 the cardinal who later became Pope Julius II ordered scribes to copy by hand a printed edition of Appian's *Civil Wars* as printed in 1472), they were generally well received by a basically conservative literate public that wanted reading matter in clear, legible, compact forms and in quantities greater than, and at prices less than, would have been possible for the copyists of the day. Within 15 years of the Forty-two-Line Bible, the printing press had been established in all of western Europe except Scandinavia.

When printing moved outward from Germany, it established itself first in Italy, where it was nurtured by German and German-trained craftsmen. Sweynheim and Pannartz (mentioned above) were the first printers in Italy. They opened their press in Subiaco in 1465 and almost immediately produced a Cicero (*De oratore*) printed in an early and interesting Antiqua type that would with time become roman. (This, rather than a type cut by another German, Adolf Rusch, in Strassburg in 1464, is generally credited with being the initial roman simply because to

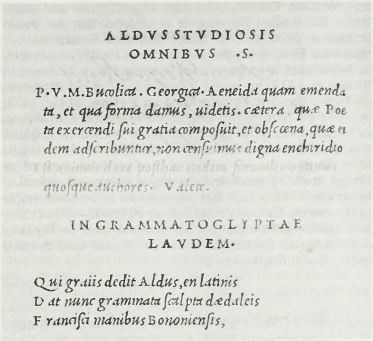
By courtesy of the Newberry Library, Chicago



ELLEM mihi quoniam ueritas in odium uenire exiſtimaꝫ: uel errore atq; impitia uulgi uaniſ et ineptis ſuperſtitionibus feruentis: uel philoſophis prauitate ingenioꝝ turbantibus eam potius q illuſtrantibus. et ſi nõ qualis in Marco Tullio fuit: qua papua & admirabilis fuit: aliquã tñ proximã eloquẽtię contingere facultatem. ut q̃ñ ueritas ui ſua ꝓpria ualeat: tñ ingenii quoq; uiribus mixta exerere ſe aliq̃n: et diſcuſſis conuictiſq; tam publicis q̃ corũ qui ſapientes putãtur erroribus humano generi clariſſimũ lumẽ inferret. Quod quidem duabus ex causis fieri uellem.

Pernitioſos autem dæmonas fraudulento & peſtes hominũ manifeſte uidentes: a quibus libe.i etiam ueſtri ad cædem petuntur non fugitis. hæc ille. Ego autẽ etiã Dionyſium halicarnaſſea uirũ romanæ hiſtoriæ acitalicæ pernitiſſimum huius ſceleris teſtem adducam: qui i primo de antiquitate italica iouem & appollinem quoniam decima hominum imolata nõ fuerat magnas italis calamitates iduxiſſe his uerbis ſcribit. Nullus in arboribus fructus inquit ad maturitatem uſq; permaſit: ſed immaturi omnes defluebant: nec ſpicæ femine replebantur: nec herbæ pecori ſufficientes germinabant: fontes quoq; ipſi alnibi nõ poterãt: alnãtalis tempore deficiebant: & aut mulieres abortum patiebantur: aut nati pueri mãci & diſcerpti erant. Cætera quoq; hoium multitudo argotatione ac morte crebrius q̃ ſolebant uexabantur: ſatisq; tantus uſq; ipis qd in deos cõmiſerũt quidq; facturi ut ab his peſtibus liberarent:

Figure 12: *Early roman face types.*
(Top) Paragraph from the Lactantius printed by Konrad Sweynheim and Arnold Pannartz at Subiaco, Italy, 1465, one of the earliest attempts to create a roman face type. (Bottom) A section of the Eusebius, printed in Venice in 1470 by Nicolas Jenson, who is credited with producing the first true roman type form.

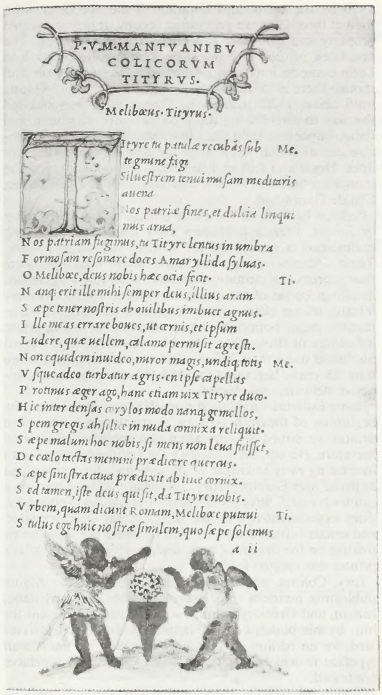


ALDVS STVDIOSIS
OMNIBVS ·S·

P·V.M.Bucolica. Georgica·Aeneida quam emenda
ta, et qua forma damus, uidetis. cætera quæ Poe
ta exercendi sui gratia composuit, et obscœna, quæ ei
dem adscribuntur, non censuimus digna enchiridio
est animus dare post hac iisdem formulis optimos
quosque authores· Valete.

IN GRAMMATOGLYPTAE
LAVDEM·

Qui graiis dedit Aldus, en latinis
Dat nunc grammata scalpta dædaleis
Francisci manibus Bononiensis,



P.V.M.MANTVANIBV
COLICORVM
TITYRVS.

Melibœus·Tityrus·

Tityre tu patulæ recubãs sub Me.
tegmine fagi
Siluestrem tenui musam meditaris
auena
Nos patriæ fines, et dulcia linqui
mus arua,
Nos patriam fugimus, tu Tityre lentus in umbra
Formosam resonare doces Amaryllida syluas.
O Melibœe, deus nobis hæc ocia fecit· Ti.
Nanq; erit ille mihi semper deus, illius aram
Sæpe tener nostris ab ouilibus imbuet agnus.
Ille meas errare boues, ut cernis, et ipsum
Ludere, quæ uellem, calamo permisit agresti.
Non equidem inuideo, miror magis, undiq; totis Me.
Vsque adeo turbatur agris·en ipse capellas
Protinus æger ago, hanc etiam uix Tityre duco·
Hic inter densas corylos modo nanq; gemellos,
Spem gregis ah silice in nuda connixa reliquit·
Sæpe malum hoc nobis, si mens non leua fuisset,
De cœlo tactas memini prædicere quercus·
Sæpe sinistra caua prædixit ab ilice cornix.
Sed tamen, iste deus qui sit, da Tityre nobis.
Vrbem, quam dicunt Romam, Melibœe putaui Ti.
Stulus ego huic nostræ similem, quo sæpe solemus
a ii

Figure 13: *Pages from the first book to incorporate italic typeface.*
(Above) Dedication and (right) first page from Virgil's *Opera*, printed by Aldus Manutius in Venice in 1501.
By courtesy of the Newberry Library, Chicago

most modern eyes its connection with the later face seems more clearly demonstrable, less tenuous. Indeed, more conservative theorists are not entirely convinced that even the Subiaco type was close enough to roman to be so called, except in the light of very informed hindsight.)

Italy: first press in Venice

The brothers Johann and Wendelin von Speyer (sometimes called da Spira and sometimes of Spire) opened the first press in Venice in 1469 and, until Johann died in 1470, had a one-year monopoly on all printing in that city. They used a clear and legible typeface that represented another step toward the contemporary roman. Whether or not these earlier types were really roman, there would seem to be no reason for putting the production of the first clearly recognizable roman any later than the work of a Frenchman, Nicolas Jenson, who had learned printing in Germany and set up business in Venice at about the time the von Speyer monopoly ran out. An excellent idealization of the roman typeform, Jenson's type was cut for an edition of Cicero's *Epistolae ad Brutum,* printed in 1470. It has been described by most modern critics as an elegant cutting, and one—Stanley Morison—called it perhaps the most perfect roman face ever cut. The expertness of the work may be attributed to Jenson's training as a medalist before becoming a printer. It is notable that Jenson never used his roman type for the printing of ecclesiastical or legal works—for which various versions of black letter were to remain standard.

By all measurement the commanding figure in the typography of the late 15th century was Aldus Manutius, who also was in Venice. Manutius established his business around 1490 and, by 1495, was issuing a series of Greek texts which were notable more for their editorial authority than for their typographical excellence. Manutius was his own editor. His type designer and cutter was Francesco Griffo of Bologna, who made two major contributions: he drew on pre-Caroline scripts as the inspiration for a more authentic roman type that soon displaced the Jenson version; and, for what was to become the most important series of books in its time, he cut the first example of the cursive type now known as italic. It was, in the opinion of some critics, not a very good italic face, and it has been described as more a slanted roman than an italic. Nevertheless, it was the first of a new family of typefaces. Interestingly, it was at first a combination of new-face lowercase letters with roman uppercases. Equally interesting, the entire text of the Aldine books for which it was used were set in the new type. Not until 1550 did it become what it is today, a special-function type.

The books for which this new type—based on a chancellery (*cancellaresca*) cursive—was first cut consisted of a spectacularly successful series of Latin texts initiated in 1501, with Virgil's works as the initial release. The series was planned deliberately to interest a market of new readers—Renaissance men who were hardly interested in liturgical writings or Greek classics but who had instead the Humanist's passion for the Latin writers, with whom, somehow, they associated themselves. To fill that market, Manutius projected a series of books compact enough to be carried easily, set in type that was both economical and highly readable, edited with scrupulous accuracy, and sold as inexpensively as possible. With Griffo's cursive type as the base, the problems of size and readability were both solved; and, by increasing the normal print run to 1,000 copies per edition, the economics were rendered more favourable. They were, indeed, the first pocketbook best-sellers, and they were what would today be called an instant success. The volumes were sought after throughout Europe, as much or more for their scholarly authority as for the excellence of their typography. New volumes were issued every two months for the next five years, and Manutius early had the honour, but dubious pleasure, of being pirated.

First pocketbook best-sellers

The continuity implicit in the work of Manutius and others during this period destroys the value of that older approach to the history of typography that isolated everything printed from 1455 to 1500 as incunabula. The year 1500 did not provide a genuine dividing point, and later historians have generally marked the end of the first valid "period" in typographic history at around 1540, after which the importance of experiments with typefaces tended to be ignored, if not disapproved of.

In Germany and in Italy, the many centres of printing grew up for the most part in the centres of commerce. But in France—where printing was from the first a sponsored activity—there were only two such centres: Lyon, from which significant printing largely disappeared after the Inquisition; and Paris, where it was established in about

1470 by the rector and librarian of the Sorbonne, who invited three German printers to occupy university-owned property and who later supervised all of their work. The first book printed in France—a manual of instruction in Latin composition—was printed in an Antiqua type; and though there is some history of the use of a mixed Gothic until about 1520, printers in France from the start led the way to establishing the predominance of roman and italic. Important influences in effecting the almost exclusive use of roman type were the printers Simon de Colines, Henri and Robert Estienne, Geoffroy Tory, and the man who was the world's first commercial typefounder, Claude Garamond.

Perhaps because of the quasi-official nature of printing in France, French publishers early established and long maintained a reputation for careful and elegant work. Their volumes, sumptuous more often than not, were characterized by minute attention to almost extravagant detailing. Books of the hours, introduced by one Antoine Vérard, whose tastes ran to illustrated and heavily ornamented pages bound in deluxe editions, were important influences in these directions. It is estimated that Vérard published more than 200 of these editions in a little more than 25 years, beginning in 1485. They are precise, mannered, delicate, and elegant.

France: the work of Henri Estienne

Henri Estienne established himself sometime around the beginning of the 16th century. A scholar, publisher, and printer, he gained his reputation as a publisher of classical literature. His edition of Galen's *De sectis medicorum* is an interesting early scientific work. Estienne, for a time, had as his adviser Geoffroy Tory, a scholar who later became a printer himself. Strongly influenced by Italian typography, Tory experimented with the use of floral ornamentation and ornate initial letters. In 1529 he wrote the first known treatise on the design of type, and in 1530 the title king's printer was created for him.

Tory, Colines, and a few others introduced the Aldine publishing methods into France. Colines designed italic, roman, and Greek type fonts, some of which were cut for him by his punch cutter, Garamond. In 1531 they created, for an edition of St. Augustine's *Sylvius,* the roman typeface to which all later so-called Garamond typefaces are traced.

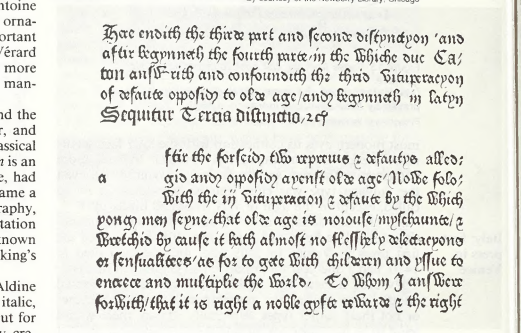
Garamond quickly became a major force in making well-designed and superbly cut types available to printers, including those who generally could not have afforded the services of capable cutters. Though Garamond's efforts with a Greek font were not notably successful, his French versions of the roman type of Manutius and an italic type of Ludovico degli Arrighi (an official in the apostolic chancellery who soon after 1522 had produced specimen pages of a type based on the cursive letters of the chancellery clerks) were of commanding importance in European typography until the end of the 16th century. In 1540, after years of experimentation, Garamond perfected a roman type that, though it had affinities with the lettering of scribes, was designed unmistakably for mechanical reproduction. It was sharply drawn, graceful and of good contrast, and it soon displaced most other typefaces then in use. This typeface ushered in the new era in which, for the first time, the typographic book was more common than the manuscript one.

From the middle of the 16th until well into the 18th century, if not later, the most notable type designers in Europe were important more for their refinements on Garamond's modifications of earlier faces than for innovations of their own. One of the very few who attempted new departures in type design was Robert Granjon, who, in addition to fashioning some notable versions of Garamond types, also tried—with his type called Civilité—to create a fourth major typeface to be different from and stand alongside roman, italic, and Gothic. He envisioned it as a national type for the use of French printers. Reminiscent of a cursive Gothic, it ultimately found its only acceptance as a display face and was not utilized in the printing of books.

Printing was introduced into England near the beginning of the last quarter of the 15th century by an Englishman who had traveled widely throughout Europe to study the art—William Caxton, who was a gentleman and dilettante. He studied printing, it is said, so that he would be able to print his own translation of a French work—Raoul Le Fèvre's *Recueil des histoires de Troye*—exactly as he wanted it to be printed. Setting up in business in Bruges in 1473, he issued *The Recuyell of the Historyes of Troye,* the first book printed in English, about 1474; in 1476 he returned to England and established a press in Westminster. The first dated book printed in England was the *Dictes and Sayenges of the Phylosophers,* issued from his press in 1477. Printed in black-letter type of an almost startling blackness, its pages command attention by means of a contrast too pronounced to be comfortable to the reader. Caxton printed some 90 books—70 of them in English—before turning his business over to Wynkyn de Worde, his former assistant. De Worde used the first italic type in England in 1524.

Introduction of printing to England by William Caxton

By courtesy of the Newberry Library, Chicago



Here endith the thirde part and seconde diſtynccyon and after begynneth the fourth parte in the whiche due Cau ſith anſwerith and confoundith the third Vitupercyon of defaute oppoſyd to old age and begynneth in latyn Sequitur Tercia diſtinctio, etc

for the forſeyd two reprefes & defauhys alled; a gid and oppoſyd ayenſt old age Nowe folowith the iij vituperacon & defaute by the which yonge men ſeyne that old age is noiouſe myſchaunce & wrechid by cauſe it hath almoſt no fleſſhly delectacyons or ſenſualitees as for to gete with children and yſſue to encrece and multiplie the world To whom I anſwere forwith that it is right a noble gyfte reward & the right

Figure 14: Portion of a page from William Caxton's edition of Cicero's *Desenectute*, printed at Westminster in 1481.[11]

Stanley Morison is authority for the statement that English typography in the first 100 years after the invention of printing was of a secondary order except for the work of Richard Pynson, a Norman who operated a press in London from 1490 to about 1530. Pynson, who used the first roman type in England in 1518, issued more than 400 works during his approximately 40 years of printing. Of these, a substantial number are legal handbooks and law codes, on the printing of which he enjoyed an effective monopoly.

Maturation of the printed book

Well before the end of the first century of typography, the printer had brought to the book the basic forms of nearly every element that he was to contribute. The styles of the three major typefaces had been formalized to the point at which little other than refinement remained to be added to them; most of the business and craft functions that were to mark the production of books down to the present had been identified and differentiated; the printed book had achieved an acceptance comparable to, and an audience far greater than, that of the manuscript volume; and publishing specialties had already emerged. Fully one-third of all of the books printed during the period of the incunabula—that is from the 1450s to 1500—were illustrated. The printing of music had become practical, and the practice of numbering the pages of a volume in sequence had been adopted.

The printer's mark, an identifying device, was used—though only briefly at first—in the typographic book from the very beginning. Almost as early, and probably more important, was the typographer's addition of the colophon, in which the printer-publisher recorded the place and date of publication, asserted his claim to credit for his role in the production of the work, advertised the merits of the enterprise, and, on occasion, attempted to protect his property from the depredations of rival printer-publishers. Indeed, Caxton turned the colophon into a short essay in which he included, in addition to the normal ele-

ments, an editor's preface and a dedication. Whether or not it is accurate to assert that the title page—the major nonmanuscript feature of the typographic book—emerged from the colophon, it is a fact that the title page took over some of the content of the colophon, which, however, continued to exist.

The first title page was probably used by Gutenberg's successor, Peter Schöffer, in 1463 on a papal bull. It was Schöffer's only known use of the device, and, like the other early versions that followed, it was really—in today's terms—a half title. The full title page did not appear until 1476, when one Erhard Ratdolt in Venice used it on an astronomical and astrological calendar. The device was well established by the end of the incunabula period. Continuing the tradition of relative anonymity of authorship of the manuscript books, the earliest pages never, and later ones only seldom, revaled the author of the work. The title page, apparently, was meant to provide, first, a protective cover for the text within and, second, an opportunity for advertising for the publisher-printer.

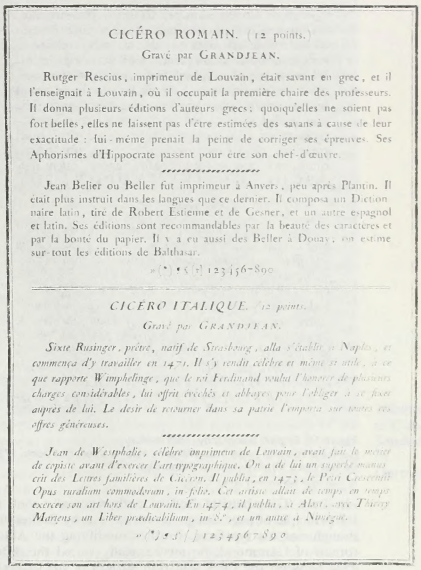
The middle years. The first really notable roman type had been cut by Jenson for a text by Cicero in 1470. It had been replaced in popularity and importance by the romans that Francesco Griffo cut for Manutius in the late 15th century. The first italic had been a Griffo design introduced by Manutius in his pocket editions early in the 16th century. These two faces had, in turn, been displaced in European typography by letters designed in the mid-16th century by Garamond in France: a roman based on Griffo's cutting and an italic based on a form put forth by Ludovico degli Arrighi. The Garamond versions of these faces were to be of prime importance in European typographical work until the end of the 16th century, during which time so many adaptations of them were produced that "Garamond type" came to be used as a generic term.

By the end of the 16th century, typography in Europe had, generally speaking, deteriorated in vigour and quality. In France, the first comeback step was taken in 1640 by Louis XIII, who, under the influence of Cardinal de Richelieu, established the Imprimerie Royale at the Louvre. In 1692 Louis XIV ordered the creation of a commission charged with developing the design of a new type to be composed of letters arrived at on "scientific" principles. The commission, whose deliberations were fully recorded, worked mathematically, drawing and redrawing each letter on squares divided into 2,304 equal parts. The approach was far removed from the style of the calligraphers, whose work had provided models for all of the important alphabets until then. It is probably fortunate that Philippe Grandjean, who was called on to do the punch cutting, did not feel himself to be under constraint to carry out his own work with the mathematical precision of the commission members who had drawn the patterns. Using the basic designs merely as suggestive, he cut a type that almost immediately drove the Garamond style from its favoured position. Known as Romain du Roi, it was used first (1702) in one of the *médaille* books that were then popular as commemorative devices. As might be expected, the type is notable for its regularity and precision; there is a good, though not exaggerated, contrast between the thick and thin strokes, and the addition of flat serifs on the lowercase letters was effective.

Grand-jean's Romain du Roi

Though intended for the exclusive use of the Imprimerie Royale, the new roman was immediately copied by other designers, one of the most active of whom was the founder Pierre-Simon Fournier, who is also remembered for his creation of a wide range of printers' devices that could be combined into festoons, borders, and headpieces and tailpieces for the heavily ornamented *éditions de luxe* that were popular in France then and that were to remain so until the Revolution.

It is reasonable to say—as did designer-theorist William Morris—that the Romain du Roi replaced the calligrapher with the engineer as a typographical influence. In general, the calligrapher was not to be reintroduced until Morris himself performed the operation as an ideological matter in the 19th century. Before that could happen, typography was to undergo further modifications under the influence of three great designers, two in England and one in Italy.



CICÉRO ROMAIN. (12 points.)

Gravé par GRANDJEAN.

Rutger Rescius, imprimeur de Louvain, était savant en grec, et il l'enseignait à Louvain, où il occupait la première chaire des professeurs. Il donna plusieurs éditions d'auteurs grecs: quoiqu'elles ne soient pas fort belles, elles ne laissent pas d'être estimées des savans à cause de leur exactitude: lui-même prenait la peine de corriger ses épreuves. Ses Aphorismes d'Hippocrate passent pour être son chef-d'œuvre.

Jean Belier ou Beller fut imprimeur à Anvers, peu après Plantin. Il était plus instruit dans les langues que ce dernier. Il composa un Dictionnaire latin, tiré de Robert Estienne et de Gesner, et un autre espagnol et latin. Ses éditions sont recommandables par la beauté des caractères et par la bonté du papier. Il y a eu aussi des Beller à Douay, qui estime sur-tout les éditions de Balthasar.

CICÉRO ITALIQUE. (12 points.)

Gravé par GRANDJEAN.

Sixte Rusinger, prêtre, natif de Strasbourg, alla s'établir à Naples, et commença d'y travailler en 1471. Il s'y rendit célebre et même si utile, à ce que rapporte Wimphelinge, que le roi Ferdinand voulut l'honorer de plusieurs charges considérables, lui offrit évêchés et abbayes pour l'obliger à se fixer auprès de lui. Le desir de retourner dans sa patrie l'emporta sur toutes ces offres généreuses.

Jean de Westphalie, célebre imprimeur de Louvain, avoit fait la moitié de cpiste avant d'exercer l'art typographique. On a de lui un superbe manuscrit des Lettres familières de Cicéron. Il publia, en 1472, le Petri Crescentii Opus ruralium commodorum, in-fol. Cet artiste allait de temps en temps exercer son art hors de Louvain. En 1474, il publia, à Alost, avec Thierry Martens, un Liber praedicabilium, in-8.°, et un autre à Nuremberg.

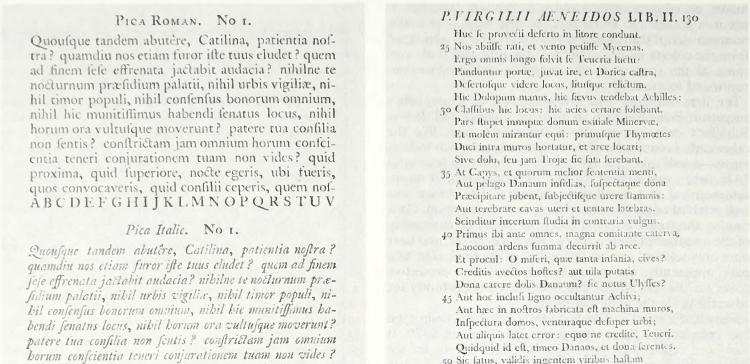
Figure 15: Philippe Grandjean's Romain du Roi type, from the specimen book of the Imprimerie Royale.
By courtesy of the Newberry Library, Chicago

William Caslon, who issued his first type-specimen sheet in 1734, made a number of refinements of the Garamond style and created faces that have become traditional and are still much in use. Caslon's refinement of the Garamond version of the Aldine roman was essentially straightforward and unmannered except for a slightly pronounced contrast between the thin strokes and the thick ones. The letters were graceful and well balanced. Serifs were bracketed (see above). They were well cut, and they made up into type blocks that were comfortable to read.

Caslon, Baskerville, and Bodoni

The type won wide acceptance and became well known in the American colonies, where it was introduced by Benjamin Franklin. It was the type in which a Baltimore printer issued the official copies of the United States Declaration of Independence.

Even more significant changes in typographical fashions were achieved about a quarter of a century later by John Baskerville in Birmingham. Baskerville, who taught calligraphy, introduced further variations in the spirit of Caslon. His letters suggest a greater concern for aesthetics. Their feeling of gracefulness is more pronounced. They were more original than Caslon's. His roman letters were open and legible; his italics tended to be spidery and quite pinched. Open and quite rounded, they are, perhaps, more self-consciously pleasing to the eye. As a book designer, Baskerville combined his new faces with exaggerated page margins and relatively wide spacings between letters to suggest new directions in style. By the use of special papers, improved press methods, and special inks, he achieved an effect of almost glaring contrast, an effect heightened by his preference for emphasizing the typographer rather than the illustrator or the engraver. Though his acknowledged masterpiece, a Cambridge Bible, was not printed until 1763, he was an important influence on English and European typography almost from the first printing of his Virgil in 1757.



PICA ROMAN. No 1.

Quouſque tandem abutère, Catilina, patientia noſtra? quamdiu nos etiam furor iſte tuus eludet? quem ad finem ſeſe effrenata jactabit audacia? nihilne te nocturnum præſidium palatii, nihil urbis vigiliæ, nihil timor populi, nihil conſenſus bonorum omnium, nihil hic munitiſſimus habendi ſenatus locus, nihil horum ora vultuſque moverunt? patere tua conſilia non ſentis? conſtrictam jam omnium horum conſcientia teneri conjurationem tuam non vides? quid proxima, quid ſuperiore, nocte egeris, ubi fueris, quos convocaveris, quid conſilii ceperis, quem noſ-

A B C D E F G H I J K L M N O P Q R S T U V

Pica Italic. No 1.

Quouſque tandem abutère, Catilina, patientia noſtra? quamdiu nos etiam furor iſte tuus eludet? quem ad finem ſeſe effrenata jactabit audacia? nihilne te nocturnum præſidium palatii, nihil urbis vigiliæ, nihil timor populi, nihil conſenſus bonorum omnium, nihil hic munitiſſimus habendi ſenatus locus, nihil horum ora vultuſque moverunt? patere tua conſilia non ſentis? conſtrictam jam omnium horum conſcientia teneri conjurationem tuam non vides? quid proxima, quid ſuperiore, nocte egeris, ubi fueris, quos convocaveris, quid conſilii ceperis, quem noſtrum ignorare arbitraris? O tempora, o mores! Senatus hoc intelligit,

A B C D E F G H I J K L M N O P Q R S T U V W

P. VIRGILII ÆNEIDOS LIB. II. 150

Huc ſe provecti deſerto in litore condunt.
Nos abiiſſe rati, et vento petiiſſe Mycenas.
Ergo omnis longo ſolvit ſe Teucria luctu:
Panduntur portæ, juvat ire, et Dorica caſtra,
Deſertoſque videre locos, littuſque relictum.
Hic Dolopum manus, hic ſævus tendebat Achilles:
Claſſibus hic locus: hic acies certare ſolebant.
Pars ſtupet innuptæ donum exitiale Minervæ,
Et molem mirantur equi: primuſque Thymœtes
Duci intra muros hortatur, et arce locari;
Sive dolo, ſeu jam Trojæ ſic fata ferebant.
At Capys, et quorum melior ſententia menti,
Aut pelago Danaum inſidias, ſuſpectaque dona
Præcipitare jubent, ſubjectuſque urere flammis:
Aut terebrare cavas uteri et tentare latebras.
Scinditur incertum ſtudia in contraria vulgus.
Primus ibi ante omnes, magna comitante caterva,
Laocoon ardens ſummâ decurrit ab arce.
Et procul: O miſeri, quæ tanta inſania, cives?
Creditis avectos hoſtes? aut ulla putatis
Dona carere dolis Danaum? ſic notus Ulyſſes?
At hoc incluſi ligno occultantur Achivi;
Aut hæc in noſtros fabricata eſt machina muros,
Inſpectura domos, venturaque deſuper urbi;
Aut aliquis latet error: equo ne credite, Teucri.
Quidquid id eſt, timeo Danaos, et dona ferentes.
Sic fatus, validis ingentem viribus haſtam
In latus, inque feri curvam compagibus alvum
Contorſit. ſtetit illa tremens, uteroque recuſſo
Inſonuere cavæ gemitumque dedere cavernæ.
Et, ſi fata Deum, ſi mens non læva fuiſſet,

R 2 55 Impulerat

Figure 16: *English typography, 18th century.*
(Left) Portion of a page from William Caslon's specimen book, 1785. (Right) A page from John Baskerville's Virgil, printed in Cambridge in 1757.
By courtesy of the Newberry Library, Chicago

In Italy, Giambattista Bodoni enthusiastically took up the principle of page design as worked out by Baskerville, though not his typefaces. Further modifying the Aldine roman of Garamond, he mechanically varied the difference between the thick and thin strokes of his letters to achieve the ultimate contrast possible in that direction. His letters are rather narrower than those of either Caslon or Baskerville. He exaggerated his thick lines and reduced the thin ones almost—it seems at times—to the point of disappearance. Like Baskerville, he used opulent papers and inks blended for special brilliance. His pages were not easy to read, but he became, in the words of Stanley Morison, the typographical idol of the man of taste, and his "plain"—though deliberately and artfully contrived—designs were an important factor in the decline in importance of the *édition de luxe* and its replacement by works more austere in feeling, more modern even to today's eyes. He set what was, in general, to be the standard book style of the world until the appearance of William Morris. (W.E.P.)

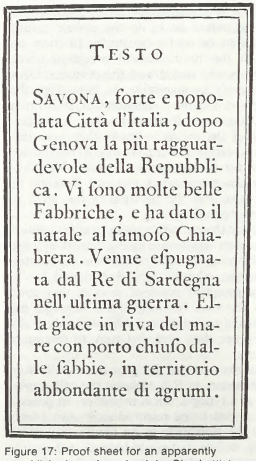
Type and book design since the 19th century. Two late-19th-century developments—one technological, the other aesthetic—profoundly changed the course of book typography and design. The advent of mechanical type composition in the 1880s (the so-called Linotype machine was patented by Ottmar Mergenthaler, a German inventor, in 1884; the Monotype, by an American, Tolbert Lanston, in 1887) had much to do with the look of the 20th century book. The Arts and Crafts Movement, whose leader in typography as in other aspects was William Morris, had an equally great influence on the quality of modern book printing.

The private-press movement. The Industrial Revolution changed the course of printing not only by mechanizing a handicraft but also by greatly increasing the market for its wares. Inventors in the 19th century, in order to produce enough reading matter for a constantly growing and ever more literate population, had to solve a series of problems in paper production, composition, printing, and binding. The solution that most affected the appearance of the book was mechanical composition; the new composing machines imposed new limitations not only on type design but also on the number and kinds of faces available, since the money required to buy a new typeface was enough to inhibit printers from stocking faces of slight utility. As a result, Victorian exuberance of design, which might use a dozen or more typefaces within a single book, was effectively curbed.

It is paradoxical that what became known as the Arts and Crafts Movement, with its roots in the romantic Gothicism propounded by the critic John Ruskin and by Morris, should have had a considerable influence on modern industrial design, including that of the book. An Englishman, William Morris was a fervent Socialist who believed that the Industrial Revolution had killed man's joy in his work and that mechanization, by destroying handicraft, had brought ugliness with it. Morris was above all a decorator; his work in the decorative arts had added great lustre to the fame he had already achieved as a writer when, partly as a result of dissatisfaction with the editions

William Morris and the Arts and Crafts Movement

By courtesy of the Newberry Library, Chicago



TESTO

SAVONA, forte e popolata Città d'Italia, dopo Genova la più ragguardevole della Repubblica. Vi ſono molte belle Fabbriche, e ha dato il natale al famoſo Chiabrera. Venne eſpugnata dal Re di Sardegna nell'ultima guerra. Ella giace in riva del mare con porto chiuſo dalle ſabbie, in territorio abbondante di agrumi.

Figure 17: Proof sheet for an apparently unpublished specimen book by Giambattista Bodoni.

of his own works, he decided to establish a press. In 1888 Morris attended a lecture given by the printer Emery (later Sir Emery) Walker and was entranced by Walker's lantern slides of early types, greatly enlarged. He proposed to Walker that they cut a new font of type that would recapture the strength and beauty of the early letters, based upon medieval calligraphy. The Kelmscott Press, in its brief life (1891–96), printed 52 books that exemplified Morris' standards of perfect workmanship. A firm believer that a return to the past would produce a better society, he commissioned handmade paper like that used in the 15th century, had new, blacker inks made, and used the handpress and hand binding exclusively; a few copies of each title were also printed on vellum. With Walker, he designed three types: a roman, based upon that of Nicolas Jenson, and two Gothics after German models; all were cut and cast by hand. Woodcut initials and borders were engraved to his own design, and wood-block illustrations were cut from drawings by Edward Burne-Jones and other of his friends.

The Kelmscott Press's major book was its *Chaucer,* finished in 1896, a sumptuous folio whose rich decorations and strong black pages are reminiscent of the German incunabula Morris admired. A table book, meant to be looked at rather than read, it is one of the most influential books in the history of printing—a revolutionary book, despite its anachronisms, which caused a whole generation of printers and designers to be dissatisfied with the books they saw about them and to attempt to improve upon the badly made, weakly designed books that were common in the late Victorian age.

Proliferation of private presses

Private presses on the Morris model proliferated in England, on the Continent—especially in Germany and the Scandinavian countries—and in the United States. The best of these, notably the Doves and Ashendene presses in England and the Bremer and Cranach presses in Germany, published books of great style and strength. There were also poorer imitations, as the Roycroft Press in the United States.

The most influential of the private presses was the Doves Press, established in 1900 by T.J. Cobden-Sanderson and Emery Walker. Walker, who was one of the prime movers in fine printing for over half a century, also played an important role in creating type for the Ashendene and Cranach presses. Cobden-Sanderson was one of Morris' circle at Kelmscott House and had become a bookbinder at the suggestion of Mrs. Morris. The bindings executed at his Doves Bindery are notable for their excellent craftsmanship and their clear, simple design, which often used Art Nouveau motifs (see below). The Doves Press books, which were printed in a type based on Nicolas Jenson's 15th-century roman, were austere in their typography, eschewing all decoration and illustration and relying for their effect on the beauty of their type, spacing, and presswork. Occasionally a second colour, a splendid red, was used, and superbly drawn initials adorned many of the 50-odd books. A five-volume Doves Bible, issued between 1903 and 1905, is among the monuments of fine bookmaking, as well as one of the most influential modern books, a result of its virility, purity of design, and perfection in craftsmanship.

The third great English private press, the Ashendene, was conducted by C.H. St. John Hornby, a partner in the English booksellers W.H. Smith and Son. Hornby in 1900 met Emery Walker and Sydney Cockerell (Morris' secretary at the Kelmscott Press), who encouraged and instructed him and helped in devising two types for his own use: Subiaco, based upon Sweynheim's and Pannartz' semiroman of the 1460s, and Ptolemy, based upon a late 15th-century German model. The Ashendene Press books, like those of Morris, were often illustrated with wood engravings, and many had coloured initials.

In Germany Morris' closest counterpart was Rudolf Koch, who gathered around himself at Offenbach, where he taught at the Arts and Crafts School and designed types for the Klingspor foundry, a community of craftsmen who painted, worked in metal, wood, and stone, printed, and wrote. Above all a consummate penman, Koch made the written word the basis of his designs in any medium, whether tapestry or woodcut. A devout Christian, Koch, like the medieval craftsmen he admired, saw the Gothic style as a supreme manifestation of religious spirit; he was no mere imitator but an artist who freely reinterpreted in his types and books the traditional Fraktur type of Germany. Koch also created a number of modern types, among them sans serifs and romans.

Influence of Cobden-Sanderson in Germany

Cobden-Sanderson's influence, however, far exceeded that of Morris in Germany. The most important of the German private presses, the Bremer Presse (1911–39), conducted by Willy Wiegand, like the Doves Press, rejected ornament (except for initials) and relied upon carefully chosen types and painstaking presswork to make its effect. The most cosmopolitan of the German presses was the Cranach, conducted at Weimar by Count Harry Kessler. It produced editions of the classics and of German and English literature illustrated by artists such as Aristide Maillol, Eric Gill, and Gordon Craig and printed with types by Emery Walker and Edward Johnston on paper made by hand in France. Kessler's books did not attempt to imitate medieval or Renaissance models; they sought to create—using the same methods as the early printers—books modern or, rather, timeless, in spirit.

The Netherlands: Krimpen and de Roos

The most notable figures of the private-press movement in The Netherlands were S.H. de Roos and Jan van Krimpen. De Roos, like Morris a utopian Socialist, was an industrial designer who hoped to create a better society by improving the appearance of ordinary utilitarian objects. His first book, *Kunst en Maatschappij* (1903), was, significantly, a collection of Morris' essays in translation. De Roos's decorative style became simple and less florid under the influence of Cobden-Sanderson, whose work he greatly admired, although his ideals remained those of the Arts and Crafts Movement. Unlike Morris and Cobden-Sanderson, de Roos was a book designer, designing books for others, rather than a printer—one of the earliest of the new school of typographers, who provided layouts for the publisher or printer, specifying type, format, and overall design. Increasingly, as technology became more complex

By courtesy of the Newberry Library, Chicago

ENTERED ACCORDING TO ACT OF CONGRESS, IN THE YEAR 1886, BY
THE TRIBUNE ASSOCIATION,
IN THE OFFICE OF THE LIBRARIAN OF CONGRESS, AT WASHINGTON.

ALL RIGHTS RESERVED.

THIS BOOK IS PRINTED WITHOUT TYPE, BEING THE FIRST PRODUCT IN BOOK FORM OF THE MERGENTHALER MACHINE WHICH WHOLLY SUPERSEDES THE USE OF MOVABLE TYPE.

Figure 18: Verso of the title page from *Open-Air Sports*, the first book set entirely by Linotype.

and shops more highly specialized and automated, design became more a profession; the typographer, trained in industrial design or graphic arts, succeeded the printer or the publisher in deciding how a book should look. De Roos, who drew a number of typefaces for the Typefoundry Amsterdam, designed books for the Zilverdistel, the Meidoorn, and other private presses, as well as for trade publishers.

Jan van Krimpen used little decoration in his work, which achieved its effect through a classic clarity of style and impeccable printing. His books, for the Enschedé firm for which he worked, for private presses, or for trade publishers, attempted always to interpret the author's meaning as clearly as possible, to reflect it rather than to enhance it. Krimpen also designed a number of typefaces, all of which show his earlier study of calligraphy. Among them are Lutetia, a modern roman and italic of great distinction; Romulus, a family of text types that includes a sloped roman letter instead of the conventional italic; and Cancellaresca Bastarda, an italic notable for its great number of attractive decorative capitals, ligatures, and other swash (*i.e.*, with strokes ending in flourishes) letters, elegant in appearance.

Marder-steig in Italy

Another typographer working in the classic mode, Giovanni Mardersteig, spent most of his creative life in Italy, though he was born and trained in Germany. His Officina Bodoni utilized Bodoni's types to print the collected works of D'Annunzio. Mardersteig not only used the handpress for limited editions (usually on handmade Italian papers) that rival 15th-century printing in their beauty of spacing and presswork, but also supervised at the Stamperia Valdònega in Verona long-run editions on high-speed presses, which are likewise remarkable for their craftsmanship. In addition, he designed several typefaces, among them Pacioli, Griffo, Zeno, and Dante.

The Art Nouveau movement was an international style, expressed in the consciously archaic types of Grasset in France; in posters and magazine covers by artist Will Bradley in the United States; and in initials and decorations by Henry van de Velde in Belgium and Germany. Van de Velde, the leading spokesman for the movement as well as one of its most skilled practitioners, in his essay

By courtesy of the Newberry Library, Chicago

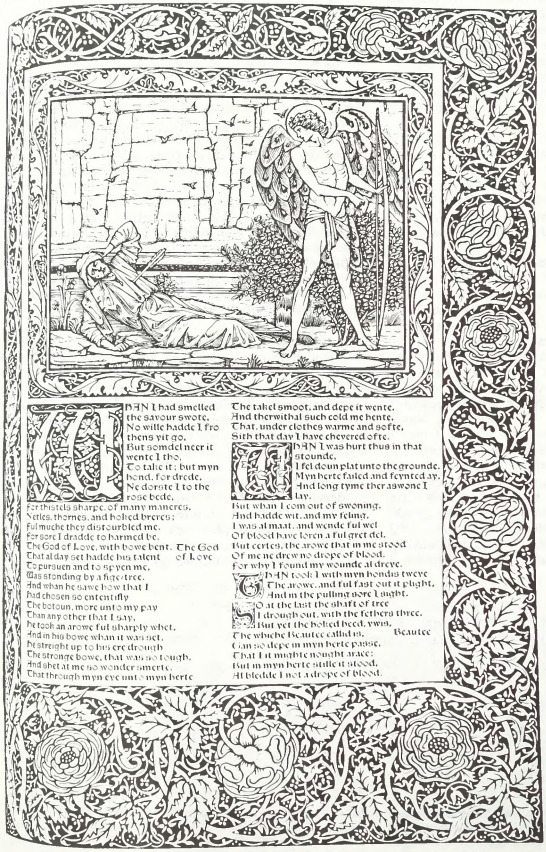


Figure 19: A page from the *Chaucer* printed by the Kelmscott Press, with illustration by Edward Burne-Jones and type and decorations by William Morris.



l'heure dont nous nous souvenons tous, le Bourgeois vivait dans un décor voulu de vertu apparente et solide, d'austérité laide, revêtue de housses blanches. Le revêtement blanc des églises s'était étenduaux maisons. Les meubles n'étaient pas plus provoquants que les lits où s'appendaient des rideaux ingénus et l'on put croire que la vertu s'était installée parmi les hommes.

Mais les juifs aussi s'étaient installés parmi nous qu'une révolution bénévole combla en leur facilitant la brocante, par le fait d'une extension

13

Figure 20: Art Nouveau initial decoration from Henry van de Velde's essay "Déblaiement d'art."
By courtesy of the Newberry Library, Chicago

"Déblaiement d'art" (1892) advocated the development of a new art, one that would be both vital and moral, like the great decorative arts of the past, but that would use contemporary modes. For a reprint of the essay, he designed a series of initials and typographic ornaments that express the characteristics of the style: decoration based upon natural forms; pages whose typography and decoration blend to make overall patterns; and a richness of texture reminiscent of illuminated manuscripts. Van de Velde's most important book was an edition of Nietzsche's *Also sprach Zarathustra,* which he designed for the Insel Verlag and had printed by the Drugulin-Presse of Leipzig and for which he created a series of ornaments printed in gold, as well as endpapers, title page, and binding; a small folio, conceived as an architectonic whole rather than a series of unrelated openings, it is a striking, if dated, volume.

Mechanical composition. The private-press movement did much to raise the standards of the ordinary trade book. Small, independent publishers who wished to make a mark not only through the distinction of their titles but also through the distinctiveness of their house styles acted as a bridge between the deluxe bibliophilic editions and ordinary books. Companies such as those of John Lane and Elkin Mathews, who published Oscar Wilde and the periodical *The Yellow Book;* J.M. Dent, who commissioned Aubrey Beardsley to illustrate Malory and who used Kelmscott-inspired endpapers for his Everyman's Library; Stone and Kimball of Chicago and Thomas Mosher of Maine, who issued small, readable editions of avant-garde writers with Art Nouveau bindings and decorated title pages; the Insel Verlag in Germany, with millions of inexpensive yet well-printed and designed pocket books—these and their many colleagues brought within the reach of the ordinary book buyer mass-produced books whose appearance, if not their method of manufacture, had been profoundly altered and improved by the Arts and Crafts Movement.

Early Linotype and Monotype faces

During the early years of the 20th century, more and more printers installed composing machines (see *Printing* above). The early Linotype and Monotype faces, like the foundry faces they imitated, were weak and poor. The first significant face cut especially for mechanical composition appeared in 1912, when a new face based upon the old-style types of Caslon was produced for *The Imprint,* a short-lived periodical for the printing trade published by Gerard Meynell of the Westminster Press in London. Its contributors included Edward Johnston, who not only wrote for the magazine but designed its calligraphic masthead; and Stanley Morison, who began his career as printing historian and typographer on its staff. Other Monotype faces cut at this time included Plantin, based upon the types of the great Antwerp printer, and Caslon; the latter was made at the instigation of George Bernard Shaw's publishers, since Shaw, who had strong views on typography, would not allow any other face for his books. World War I, however, stopped any further development of types for the composing machines.

In America the generation of designers who had begun as disciples of Morris soon began to develop their own styles. Among the most important were D.B. Updike, Bruce Rogers, F.W. Goudy, and W.A. Dwiggins.

Daniel Berkeley Updike opened the Merrymount Press in Boston in 1893. His books, most of which he designed himself, are noteworthy for the clarity of their organization, their easy readability, and their excellent workmanship, based upon the use of a few carefully selected typefaces and immaculate presswork. Updike stocked only types that met the twin criteria of economy in use and beauty of design. His books, whether a complex folio such as the *Book of Common Prayer* (1930), which is considered by many to be his masterpiece, or the small and amiable *Compleat Angler* (1928), are both functional and pleasing to the eye.

Bruce Rogers was a typographer, trained as an artist, who had the faculty of drawing the best from the printers with whom he worked. His greatest book, a monumental Oxford Lectern Bible of 1935, is the noblest edition of the Bible ever issued in English; his smaller and less ambitious efforts, often decorated with the typographic ornament at which he was a master, possess enormous wit and charm. His one type design, Centaur, which was based upon Jenson, is among the most successful modern adaptations of an early roman, although it is too elegant for frequent use.

Goudy's typefaces

Frederic William Goudy, who was the most prolific American type designer, created more than 100 faces during a long career as a printer, editor, and typographer. In 1908 he began a long association with the Lanston Monotype Corporation, for which he did much of his best work. Among his types were Forum and Trajan, which were based upon the roman capital letters inscribed on Trajan's Column; Goudy Modern, his most successful text face; and a number of black-letter and display faces. Goudy edited two journals, *Typographica* and *Ars Typographica,* in which he expounded his theories of design; he also wrote a number of books, among them *Elements of Lettering* and *The Alphabet.*

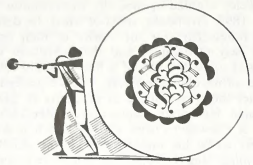
By courtesy of Monsen Typographers, Inc

Education: To prepare us for complete
on which education has to discharge. H
on is properly to draw forth, and implie

Figure 21: Goudy's Old Style typeface.

William Addison Dwiggins, a student of Goudy, was long associated with the publishing firm of Alfred A. Knopf, whose house style he helped to establish. In hundreds of volumes of trade books he designed, typography was taken seriously (each book carried a brief colophon on the history of the type employed); there was an attempt to use contemporary typographic decoration; and the bindings, using designs made up of repeated decorative units like early printers' fleurons, were extremely successful. Dwiggins designed a number of typefaces for the Linotype, two of which, Electra and Caledonia, have had wide use in American bookmaking. In the U.S., unlike England and the Continent, printers have relied far more upon Linotype than Monotype for book composition.

English typography, like that everywhere, marked time during World War I but made remarkable progress soon after. A new generation of typographers, inspired by Morris' ideals of quality but at the same time aware of the need to adapt them to the new mass-production techniques, had begun to make their names. Foremost among these was Stanley Morison, who, after a year's apprenticeship with *The Imprint,* became a typographer on the staff of Burns and Oates, where he worked on a wide variety of books, among them the liturgical texts in which the firm specialized; here he began to develop the rationalistic approach to typographic design that characterizes the English school. Morison demanded that typography be functional: the task of the book and the newspaper designer was to transmit the author's text clearly, and the task of the advertising and display designer was to command attention. In 1922 Morison became typographic adviser to the Monotype Corporation and instituted a program of



ADVERTISEMENT

This pamphlet is the announcement of a new Linotype face to be called "Electra," cut from designs drawn for The Mergenthaler Linotype Company by W. A. Dwiggins. The face—at this stage completed in twelve point, roman and italic—provides a new type texture for book-page composition. In the larger sizes now in preparation it will furnish the printer with a new note in advertising typography.

The face, as may be seen from this specimen, falls within the "modern" family of type styles, but is

Golo Mann, professor and writer of history, was born in 1909 in Munich, Germany, the son of the famous novelist, Thomas Mann. He received his Ph.D. from Heidelberg University in 1932 and, leaving Germany the following year, lectured at French universities until 1937. He then edited the literary review "Mass und Wert" in Zurich, Switzerland, and in the early 1940's moved to the United States. He remained in this country until 1958 and during this time taught at Olivet Col-

Figure 22: *Linotype typefaces designed by W.A. Dwiggins.* (Left) Page from *A Baker's Dozen of Emblems* set in Electra, 1935. (Above) Caledonia italic.
By courtesy of (left) the Newberry Library, Chicago, (above) EB Inc.

cutting for the composing machine a repertory of types culled from the best faces of the past, to which were added a number of contemporary faces designed for modern needs. He had prepared himself for the task by a strenuous course of self-education in paleography and calligraphy, in order to understand the written hands that the early types imitated, and in the history of printing design itself. In 1923 he joined Oliver Simon in publishing *The Fleuron,* a journal of printing history and design in which he published a number of important articles on calligraphy and typography.

Morison's work for Cambridge University Press

In 1925 Morison was made typographic adviser to the Cambridge University Press, whose printer, Walter Lewis, had begun a complete reform of its typographic resources. Cambridge stocked most of the types Morison commissioned for Monotype and demonstrated by their intelligent use that mechanical composition could be used to produce books at once handsome and functional. Among these types were Garamond, based upon a 17th-century French letter (see above); Bembo, after an Aldine roman; Centaur, an adaption of Rogers' foundry face; and Baskerville and Bell, based upon English models. Italics included Arrighi, a version of the letter used by the 16th-century papal writing master and printer (see above). Among the modern faces whose design Morison supervised were Eric Gill's Sans Serif, which enjoyed a wide vogue in advertising and avant-garde book typography; Gill's Perpetua, based upon his stonecut letters; and Times New Roman, designed by Morison himself for *The Times* (London), whose staff he joined in 1930. The last has been called the most successful type design of the 20th century, a result of its economy and legibility when used on high-speed presses.

Francis Meynell was another who demonstrated that mechanical composition and printing, if properly used, could produce aesthetically satisfying books. The books of Meynell's Nonesuch Press, which were usually limited editions of the classics reflecting his own catholic and excellent literary taste, are marked by restrained design, fine papers, and careful presswork. More varied and original than those of the earlier private presses, they were printed not by the proprietor but by large, mechanized shops. Meynell's trade books, published under the same imprint, demonstrated that well-designed and manufactured books need not be costly; the Nonesuch one-volume editions of English classical authors were inexpensive, handsome, and readable.

The most influential modern publisher of English low-priced books, however, was Allen Lane, whose Penguin books, established in 1935 and inspired by such continental publishers as Insel Verlag and Albatross, proved that a well-designed series of inexpensive paperbacks, both worthwhile reprints and new titles, could succeed both commercially and intellectually. They did much to bring about the paperback revolution that swept both the Continent and the United States in the period that followed World War II.

Influence of the Bauhaus

German typography from World War I until the advent of Adolf Hitler was greatly influenced by the Bauhaus, which stressed the graphic arts; its books, which were heavi-

By courtesy of the Newberry Library, Chicago

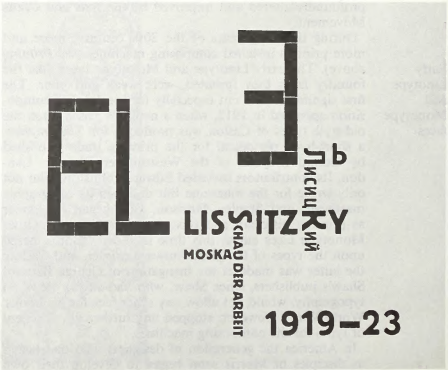


Figure 23: Catalog cover by El Lissitzky, in the Bauhaus Asymmetric style.

ly illustrated, broke away from traditionally symmetrical layouts, in which pictures were inserted into a rigid framework of text, and strove instead for freer arrangements, usually asymmetrical, in which the type supported the illustrations. The attempt was to create graphic patterns on the page and to enhance the reader's consciousness of the illustrations. Many of the Bauhaus faculty were architects and industrial designers, whose principles demanded that the types they used, like the buildings and machines they designed, be sharp and unornamented, symbolic of a machine-dominated society. Their favourite types were sans serifs, such as Gill's Sans Serif and Paul Renner's Futura. When the Nazis dispersed the Bauhaus group, its style became truly international. It has since lost favour among book designers, except for art and architectural books, partly because sans serif types and asymmetrical layout proved less legible than traditional modes and partly because of its rigid limitations.

Other between-war styles, closely linked to literary or artistic movements that affected book design, were Dadaism and Surrealism. The Dadaists' pamphlets, posters, and books employed free, abstract layout, a great mixture of type sizes and faces, and an attempt to create mood through typography. Surrealist writers such as Guillaume Apollinaire and André Breton often collaborated in the design of their own books, attempting to make the typography of their works reflect its mood.

In France, especially, the production of books intended to be works of art in their own right was dominated by painters and sculptors. Publishers such as Ambroise Vollard commissioned members of the School of Paris, among them Braque, Matisse, Bonnard, and Picasso, to illustrate books in which the illustrator worked closely with highly skilled craftsmen to create colourful, original, limited editions, which, while they sometimes may fail as readable books, achieve admirable success as visual decoration.

During the 20th century, styles in book design, as in all the arts, fine or applied, have become increasingly international. Styles born in one country spread throughout the world and die through overuse at a dizzying rate. As a consequence it has become increasingly difficult to distinguish truly individual or national styles—books, magazines, clothes, paintings, music, regardless of country of origin, all resemble one another far more than they differ. (See also WRITING.) (J.M.Ws.)

Photoengraving

Engraving is the broad term describing the procedure used in making printing plates, in which printing and nonprinting areas are distinguished by their height with respect to the general plane of the surface, the artistic decoration created by mechanically incising a design into a surface, and the creation of original works of art by tooling or etching an image into the surface of a metal (or plastic) plate and transferring the resultant image to paper. For detailed information on these last two subjects, see PRINTMAKING. This section is limited to consideration of the procedures whereby a printing surface useful in the production of multiple ink-on-paper images is produced.

The term photoengraving is correctly applied to the procedures discussed here, since the use of light energy, as involved in photographic processes, is essential. A distinction must be made between a relief (or letterpress) printing plate, in which the ink-carrying (or image-bearing) surface coincides with the general level of the plate surface, with nonimage portions cut below the surface, and intaglio (or gravure) printing surfaces, in which the ink-carrying image elements are incised into the plate surface. In letterpress printing, a uniform film of ink is distributed over the surface of the plate and transferred from the individual image elements to the receiving paper surface. In intaglio printing, the plate is flooded with a low-viscosity (thin) ink, then wiped with a blade (doctor blade) to remove any ink adhering to the surface. The doctoring action leaves the incised intaglio image filled with ink; later, as paper is brought into contact with this image and pressure is applied, surface-tension and capillary-action forces cause the ink to transfer from the plate to the paper.

Letterpress and intaglio printing

HISTORY OF PHOTOENGRAVING

The earliest engraved printing units were wood engravings, in which the nonimage areas of an illustration were removed by carving them from the surface of a flat wood block. The oldest known illustration printed from a wooden block was a Buddhist scroll discovered in 1866, in Korea. While the dating of the print is not exact, it is believed to have been prepared about AD 750. The Chinese *Diamond Sūtra*, dated 868, incorporates a woodcut title page and text that includes numerous woodcut images.

From these 8th- and 9th-century dates, it is clear that the use of woodcuts (images cut into a surface parallel to the wood grain) and wood-block engravings (images incised into the end grain of an assembled block) antedates the invention of movable type. The earliest extant example of a European print from a wood engraving to which a reliable date may be attributed is a print entitled "St. Christopher," dated 1423, discovered in the library of the Carthusian monastery in Buxheim, Germany. Another authenticated example of 15th-century wood-block printing is the "Apocalypse of St. John," printed in 1450, after a 14th-century manuscript.

Early etched plates. Plates engraved in wood continued to find use in printing application through the late-medieval and early-modern periods. Plates made of copper, pewter, and other metals were also produced, by a process in which an image in wax or bitumen was drawn on, or transferred to, the surface of the plate and nonimage areas removed by action of appropriate acids.

Preparation of intaglio printing plates by coating a metal plate with an etchant-resistant substance (ground) such as wax, bitumen, or shellac, scratching through this substance (ground) to expose the plate surface, then etching in acids is also a late-medieval European development. This process, however, developed as a medium of artistic expression, rather than a technique for the mass production of printed images.

The first experimental application of light-sensitive materials to the production of printing surfaces was made by Joseph Nicéphore Niepce, of France, an early researcher in lithography who began his experiments in about 1813. He is credited with having produced the first permanent photograph. In 1826 Niepce coated a pewter or copper plate with a photosensitive asphaltum and exposed the surface to bright sunlight through an etching of a portrait, which served as a positive image. Sunlight passing through the background of the etching hardened the asphaltum, while the protected areas, under the inked portion of the etching, were developed in oil of lavender and white petroleum to create an image in exposed metal. This image was then etched into the plate, and from the intaglio image, prints were made on a copperplate press.

Early use of light-sensitive materials

Though this basic discovery was of historical importance, it did not bring about the immediate use of photoengraved images for printing, and many other attempts to produce engravings by exploitation of the photosensitivity of various natural compounds were made by experimenters in Europe and the United States. The origin of the modern photoengraving process rests, however, on the report (1839) by a Scottish scientist and inventor, Mungo Ponton, of the light-sensitive properties of certain chromium compounds. But Ponton, who demonstrated the chemical change that occurs when glue containing a compound of chromium is acted upon by light, was not concerned with preparation of printing plates, and it remained for William Henry Fox Talbot, an English pioneer in photography, to propose the use of chromium-treated colloids such as albumin as an etchant-resistant for preparation of intaglio printing surfaces.

Early 19th-century work on production of chemically etched letterpress printing plates antedated, in many instances, the invention of photography. A researcher in Paris developed a process for the preparation of engravings on zinc. His work involved transfer of an image to the zinc plate by mechanical means, using ink or wax, and the removal of the nonprinting areas in a series of etching operations, each of which involved applying a coating of ink to the sidewalls of the etched lines by means of resilient rollers. The ink served to protect the lines of the

engraving from the action of the etching acid, so that the printing area was not reduced.

Wet-collodion photography. The introduction in 1851 of a so-called wet-collodion process for photography provided a means for producing a photographic negative as the basic element in the preparation of engravings. In this process, a glass plate is coated with an alcohol–ether solution of collodion (cellulose nitrate) containing potassium iodide. While still wet, the plate is immersed in a silver nitrate solution, producing light-sensitive silver iodide in the collodion layer. Without drying the film of collodion, the plate is placed in the camera and exposed, followed by development in ferrous sulfate solution and chemical "intensification" to produce an image of greater opacity. The image consists of deposits of metallic silver and other heavy metals imbedded in the collodion layer.

This photographic process also provided a method of stripping the photographic image from the glass plate, permitting assembly of a number of images for plate making, and also making possible the geometric reversal of the image needed in letterpress plate making to produce a right-reading print on paper. The wet-collodion process was used extensively in engraving until the 1930s, when it was gradually replaced by commercially coated stripping films.

The halftone process. Since the letterpress printing process provides a uniform coating of ink on all printing elements, no provision can be made for reproducing tones intermediate between black and white by varying the thickness of the ink film laid down by the press. The production of shades of gray was then the role of the halftone process, in which the image is broken up into dots, and variations of gray tones are obtained by varying the size of the dots, thus controlling the amount of ink laid down in a given area.

Halftone feasibility proved

The feasibility of the method was demonstrated in about 1850, when a halftone image was produced by photography through a screen of loosely woven fabric. The screen was placed some distance forward of the plane of the receiving photographic surface (film or plate) and had the effect of breaking the gray tones of the subject into dots of varying sizes, through a combination of geometric and diffraction effects involving the spacing of screen from the image surface, the size of the openings in the screen, distance from lens to image plane, and the size of the aperture in the lens. It was obvious that a screen designed for this use could consist of a pattern on a glass or other firm, transparent surface.

A French patent of 1857 described a screen with parallel lines scratched in a single direction in an opaque background. As early as 1869 an image with a crossline halftone was produced in the *Canadian Illustrated News*. Later, in 1882, a crossline halftone was produced using a single-direction screen, by making half the exposure with the screen in one position and half with the screen rotated a quarter turn. Two brothers, Max and Louis Levy, of Philadelphia, in 1890 produced the first commercial halftone screens. The Levy brothers coated selected plates of high-quality optical glass with a lacquer, in which parallel lines were cut. The ruled lines were then etched with hydrofluoric acid and filled with an opaque material. Two such plates were cemented face to face with the lines at 90°, the edges sealed, and the assembly bound in a metal frame.

There has been no significant change in the methods of making halftone screens since those developed by the Levy brothers. Other screen patterns, including triangular dot patterns and the grained (mezzograph) screen, have been proposed, but none has produced consistently satisfactory results. For special effects, screens having straight or wavy line patterns and screens that produce a pattern of circles, concentric about a point that is chosen to be the focal point of the readers' interest, are in use. These screens are generally produced photographically from hand- or machine-drawn patterns and are used in the form of contact screens.

Basis for selection of screen ruling. Halftone screens may be obtained with line frequencies of 50 to 400 lines per inch (one inch equals 25.4 millimetres). The coarser screens are used for reproductions printed on coarse papers, the fine screens for higher quality reproductions on highly finished and coated papers. Screens in the 50–85-line frequency range are used primarily in newspaper illustration, while 100-, 110-, and 120-line halftones are suitable for highly polished papers and for some magazines, where single-colour and some multicolor work is involved. The 120-, 133-, and 150-lines-per-inch screens are generally used for colour illustrations in magazines and books printed on coated papers, when picture detail is important. Screens of 175 and more lines per inch are seldom used in letterpress printing, since the inks tend to fill the screens, causing difficulties in the press run. Such screens, however, do have some use in printing by offset lithography. In general, where paper quality permits, the finer screens are used when reproduction of fine detail is important. But since the letterpress process requires that the diameter of the finest highlight dot should not be less than 0.0015–0.002 inch, the use of very fine screens will lead to loss of image contrast, since some 3 to 5 percent of the picture area, in highlights, will be ink covered.

"Altone Gradar Screen"

An interesting development in glass screens was the "Altone Gradar Screen," manufactured in Germany. These are glass screens, ruled and etched in the usual manner, but with the rulings of the two glass elements filled with a transparent magenta lacquer of two different optical densities. When the screens are assembled, lines in one direction exhibit a density different from that of lines in the perpendicular direction, and the intersections have a density equal to the sum of the densities of the two lacquers. The effect is to provide elongated halftone dots, with improved tonal reproduction in intermediate gray tones on coarse paper such as newspaper stock.

Contact screens. Perhaps the most significant recent advance in the halftone process has been the use of contact screens—films bearing a gray or magenta-dyed image of the light-distribution pattern behind a conventional halftone screen. The screen is placed in contact with the surface of a high-contrast film, in the image plane. The image, as recorded on the film surface, has the characteristic of a halftone exposed through a glass screen, with significant improvements in rendition of detail of the subject. Though first proposed in 1855 and developed by a number of later investigators, this technique was not fully exploited commercially until the 1940s. The contact screen eliminates certain diffraction effects inherent in glass screens, frees the operator from some of the lens-diaphragm restrictions imposed by the glass screen, and eliminates the necessity that lens opening, bellows extension, screen distance from the focal plane, and screen ruling all be in a particular relationship.

Contact screens are made with a silver (gray) screen pattern image and with a magenta dye image. The dyed screen gives some additional control over halftone negative quality through use of colour filters on the camera.

The benday process. An entirely mechanical procedure for production of a halftone image on a metal printing plate is the benday process (1879), named after its inventor, Benjamin Day, a New York newspaper engraver. This process utilizes a series of celluloid screens bearing raised images of dot and line patterns. The screen surface is covered with a waxy ink and the ink transferred, by pressure and rolling, to prepared portions of a metal plate. By selecting different screen patterns for transfer to different parts of the image, a mechanically produced halftone image is rendered. The ink image is reinforced with powdered resins and the plate etched. This process has been supplanted by completely photomechanical techniques.

Special effects. Such techniques as dropping out the highlights from the halftone negative (*i.e.*, eliminating the dots in these areas) in order to achieve increased contrast in illustrations were studied and introduced by several individuals. Such a method was patented in 1893, and in 1925 a camera attachment was introduced, making it possible to impart a slight motion to the image on the film and thus reduce exposure to the point at which small highlight halftone dots were not printed or developed.

The most successful of the highlighting methods were those employing fluorescence phenomena, in which an object produces visible light when exposed to ultraviolet

radiation. In 1938, for example, the fluorographic process, in which fluorescing materials were incorporated in the artist's pigments, was patented. Similar pigments, designed for colour correction in watercolour illustrations, were patented in 1935 and 1938. Another process introduced shortly thereafter utilized a fluorescing paperboard. All of these processes were based on the same procedure: making an exposure under normal lighting for overall reproduction and then making an additional correcting exposure under ultraviolet. The fluorescence produced by the ultraviolet illumination provided additional exposure in the affected areas that gave the necessary correction for highlighting or colour correction, by eliminating the screen pattern from "white" areas, in the case of monochrome, or reducing printing dot sizes, in critical areas of colour work.

Process developments. The discovery of the halftone screen was primarily responsible for the development and growth of photoengraving; further growth was related to other developments in the printing and allied industries.

Effect of colour film on printing developments

The introduction in 1935 of the first practical colour film for amateur and professional use probably did more to accelerate printing developments than any single invention. By making bulky studio-type colour cameras obsolete and permitting the use of readily portable camera equipment for the production of colour images, on-the-spot colour photography became possible, greatly increasing the use of coloured illustrations.

At approximately the same time, the commercial production of coated paper and heat-drying printing inks for letterpress printing began. Many colour developments for films, printing processes, and materials followed.

Chemical etching—traditional and powderless processes. Early methods of etching zinc and copper, methods that have persisted in some areas to the present day, were tedious and inexact and could be learned only through trial-and-error training. The principal difficulty stemmed from the fact that the chemical removal of metal from nonimage areas proceeds in all directions. Thus, etching of the plate surface proceeds not only in the desired direction, to achieve the depth required for satisfactory printing, but also sideways, causing reduction in width of lines and dots of the printing image and also undercutting halftone dots—producing a below-surface dimension smaller than the printing surface. The mechanical weakening of the dot may lead to its collapse under printing pressure.

Some success in overcoming this problem has been achieved by depositing an etchant-resistant material about the sidewalls of etched lines and dots, thus preventing lateral etching. The method of rolling a waxy ink onto sidewalls of lines and dots, called gillotage, has found wide use among European engravers. The "powdering" process, most widely used in the United States, involves brushing a resinous powder (dragons' blood) against the sides of partially etched lines and dots and fusing, with heat, to provide an etchant-resistant coating. Several repetitions of the operation—etching, application of the protective material, and etching again—are needed before sufficient depth is attained. Results of this process are dependent upon the skill of the operator and on such ambient conditions as temperature and relative humidity, since these affect the performance of the powder. A major step toward solving the problem—in fact, the most important development in the field of etching since photoengraving was invented—came with introduction of a process of etching a magnesium plate without the use of powder. Experimenters found that by adding an oily material and a surface-active (wetting) agent to the nitric acid bath and controlling the conditions under which the plate was etched, they could produce characters in relief with adequate etching depth and virtually no printing-area loss during the etching. Later adapted to the etching of zinc, the process was quickly adopted by engravers in all parts of the world.

Powderless etching of copper

With this major hurdle in the etching of zinc and magnesium overcome, attention turned to copper, and in 1954 it was found that a powderless etching process for copper resulted from the addition of an organic compound (thiourea) to the iron chloride etching bath. Further refinements in the process and the introduction of new compounds to add to the etching bath followed.

Electromechanical plate making. While these developments in chemical etching were taking place, other experiments were being conducted to assess the feasibility of replacing traditional methods with the techniques of electronics, optics, and mechanics. The first successful result of these efforts was a device, introduced in 1947, that optically scanned a picture and simultaneously reproduced it as a relief printing plate on a plastic sheet. This device found wide application, particularly in newspaper plants, where the slowness of photoengraving procedures was particularly objectionable. Within a short time, machines were developed that were capable of making etched plates in metals.

Meanwhile, investigators in the United States discovered about 1950 that some methacrylate compounds could be quickly polymerized (converted to products of high molecular weight and low solubility) by exposure to light. Nylon was also found to be photosensitive, and by 1958 both materials were being offered for use in printing plates. Another plate-making system, reportedly based on light-sensitive polyurethane resins, was introduced in 1968.

Colour scanners. Paralleling the development of the electromechanical engraving machine, experimenters in the United States and Europe independently devised a number of electromechanical devices that automatically produce, from a colour-transparency image, corrected film negatives from which the four printing plates used in full-colour reproduction can be prepared.

In one of these, a photographic transparency, wrapped around a glass cylinder, is scanned by a narrow beam of light. After passing through the transparency, the light continues through a colour splitter, and the blue, green, and red components are directed onto the sensitive surfaces of photocells. The electronic signals thus generated are modified and amplified in a computer that functions as an electronic analogue of photographic colour-separation processes. The computer activates a series of lamps, which expose the colour-corrected images onto photographic films mounted on another drum, attached to the same shaft as the transparency holder. This development was based on initial experimentation in commercial laboratories in the late 1930s and early 1940s. Other units, based on similar principles but differing in some details of structure and operating procedures, have been manufactured in the United States, West Germany, and Great Britain.

MODERN PHOTOENGRAVING TECHNIQUES

In terms of cost, engraving methods range in ascending order as follows: line engravings; halftone engravings; combination line-and-halftone engravings; single-colour, two-colour, and duotone engravings; and process colourplates. Each of the types may be produced in any of the customary metals or plastics. Process colourplates are usually made of copper in the United States and United Kingdom and of zinc elsewhere.

Basic production processes. The essential operations for the production of all types of photoengravings are similar. They include photography, photomechanical operations, etching, finishing, routing, blocking, and proofing.

Camera and darkroom equipment. The engravers' camera, called a process camera, is a rigidly built machine designed to allow precise positioning of the lens and copyboard so as to provide control over the enlargement or reduction in size of the copy. It has a colour-corrected lens designed to give the sharpest possible overall image when focussed on a plane surface, without the distortions common (though usually unnoticed) in the average portrait or amateur camera lens. Process cameras are designated as gallery or darkroom types. The gallery camera is free-standing and may be installed in any convenient location, but film must be removed in a light-tight cassette and processed in a separate darkroom. The darkroom camera is installed with its film holder as an integral part of the darkroom wall, giving easy access to the darkroom facilities.

The process camera

The material to be reproduced, called copy, is mounted on a board or glass-covered copyholder, carried on the bed of the camera. Illumination for exposure is provided by arc lamps or high-intensity gas-discharge lamps. The most common camera lamp systems in late years have involved

pulsed xenon lamps, in which a high-voltage alternating current, passing through a glass tube containing the rare gas xenon, causes the emission of a light rich in the ultraviolet wavelengths.

Virtually all photographic work is done on film coated with high-contrast emulsions especially developed for graphic arts work. The introduction of dimensionally stable film bases has nearly eliminated the use of glass plates. Film emulsions used for halftones yield the extremely high contrast needed for halftone or line reproduction. Stripping film, a laminated film with a soft adhesive layer between the base and the emulsion layer, is widely used to permit images to be removed from the base and properly oriented on the glass or film flat through which the metal plate will be exposed.

In the early days of photoengraving, with wet-plate images on a glass support, it was impossible to process photographic images by any means other than immersion in solutions contained in a shallow pan or tray or by dipping into a tank of solution. Such tank and tray processing remains important but is now being supplanted by the use of automatic film-processing machines. Derived from equipment originally designed for processing of motion-picture film or photostat prints, these consist of belt- or roller-driven apparatus that carries the film through developer, fixing, and washing solutions, and, in most cases, through a drier, permitting delivery of a processed, dried film within three to five minutes after insertion into the machine. Such machines, with different processing solutions, may be used for continuous-tone or lith-type films.

Automatic film processors

Plate coating and printing. Photomechanical operations include cleaning the metal plate surfaces, coating with a light-sensitive solution, drying the coating (known as the top or enamel), and making the exposure on this coating through the negative prepared in the photographic step. Throughout these operations care is required to prevent imperfections such as bubbles, dirt, or scratches in the light-sensitive coating. The zinc, magnesium, or copper is prepared by careful cleaning with pumice and water. The light-sensitive coatings are usually poured over the surface, and the plate, held flat, is whirled to ensure uniform coverage by the solution.

Light-sensitive coatings are usually a dichromated colloid material, but light-sensitive resins are also used. "Cold top" enamels are used on zinc and magnesium, which cannot be heated; these are usually slightly alkaline solutions of shellac or polyvinyl alcohol to which a dichromate is added. "Hot top" enamels nearly always contain fish glue as well as some egg albumin, to which is added a dichromate sensitizer. Mixtures of glue and albumin are used when it is necessary to control the etch resistance and the ease with which the edges of the enamel break away during the etching process. Hot top enamels must be set at temperatures of 550°–650° F (285°–345° C) and are used mainly on copper, the crystal structure of which is not altered at these temperatures. Polyvinyl alcohol and shellac resistants are set at temperatures of 350° and 220° F (175° and 105° C) respectively; therefore they are used on zinc and magnesium.

The tops are high-contrast materials that, when exposed to strong ultraviolet light, harden where the light has struck them and lose their solubility in water. Development in water then removes the coating from the unwanted areas of metal, exposing the metal for the etching process. Photosensitive resinous materials find wide application in electronic circuit printing, an operation analogous to photoengraving. They have more limited applications in the making of photoengraved letterpress plates, where they are used especially on zinc and magnesium and where their excellent storage properties permit their application in the metal-finishing plant, obviating the necessity for coating of the resist onto the metal in the photoengraving shop. These resinous materials are developed in organic solvents.

Etching and finishing. Nitric acid is commonly used in etching zinc and magnesium, the strength varying from 6 to 15 percent, depending on the metal. Copper is more readily attacked by ferric chloride (iron chloride), which is commonly used in concentrations of 28–45 percent. The etching may be done in an open tub or tray, though this method does not give the control needed for economical operation and is employed only where control is not critical. Most quality work is carried out in etching machines provided with impellers that break up the etchant into a spray and force it against the plate.

In the conventional etching processes, the acid or iron chloride is used without modification, although great care is needed to prevent overetching. In many cases, especially when making line plates, etchers powder to protect the upper printing areas from attack while continuing to etch in depth. The powderless etching processes, described earlier, have made the powdering technique obsolete and are now almost universally in use. Line plates are usually etched to depths of 0.010 to 0.045 inch. Halftones may be etched to depths of 0.0023 to 0.009 inch, depending on the fineness of the screen. Coarser screens are etched deeper.

Photosensitive plastic plates are not etched in the ordinary sense. Unexposed resins, from nonprinting areas, are washed out with either dilute alkali or alcohol. Overetching is not a problem with this type of plate.

Finishing includes hand operations with engravers' tools, to remove imperfections in the image area of the plate and to improve its appearance. In colourplates, finishing also includes colour correction, a process of further etching or burnishing selected areas to improve the fidelity of reproduction. Finally, unwanted metal in the nonprinting areas of the plate is removed by a mechanical routing machine.

Blocking and proofing. Blocking consists of attaching the plates to cherry wood, plywood, or metal blocks to bring the printing surface to type height, which is 0.918 inch. Until the development of thermoplastic adhesives in the 1940s, blocking was always done by nailing the plates to wooden blocks. This tedious and costly operation has been largely replaced by hot mounting, in which process the plate is placed on a block of wood precoated with adhesive and this sandwich is subjected to heat and pressure. Upon cooling, the plate adheres firmly to the block.

Proofing consists in placing the plates on a precision press and taking sample impressions, or proofs, that show how the plates will print during a regular press run.

Colourplate production. The first printed colour work was produced manually; artists painted in the necessary colours on black-and-white printed sheets. Later, stencils were used to speed this work, and in a further development, colours were printed, either as solids or tints, from hand-engraved plates. All of the work was crude by modern standards, however, and nothing approaching four-colour process printing was possible.

Modern colour printing

Modern colour printing, done with either three or four plates, each using a different colour of ink and overprinting the others, is based on a subtractive system of colours in which intermediate hues are obtained by some combination of two or more of the subtractive, or secondary, colours. The best colour printing is usually done with four process colours: yellow, magenta (blue-red), cyan (blue-green), and black.

The black plate is used to provide added uniformity of colour reproduction, since it will overcome changes in hue of critical neutral tones that could occur with random or cyclic variations in the amount of ink being transferred to the plate from the press inking system. Further, the use of a black plate aids in maintaining sharpness of picture detail.

In theory, black should result whenever the three subtractive colours are superimposed. Thus, it should be possible to produce black wherever all three of the secondary colours are present without affecting reproduction. Further, any colour that is within the range of colours reproducible with inks on paper can theoretically be obtained by using only black plus the proper pair of the secondary colours. But this has not been found practical because of the nature of printing ink pigments and the lack of total precision in the printing operation. Consequently, it is common practice to use the black plate to supplement the colourplates, portions of which are allowed to print in all except pure white areas of an illustration. The colourplates and the black plate must all be printed in register; *i.e.*, they must be superimposed so that identical portions of the image in each plate colour overprint each other.

Colour separation. In manufacture, the production of an individual colourplate involves the same steps used in producing an ordinary black-and-white engraving, once the etchant-resisting image has been printed on the metal. Prior to this, the only differences lie in the use of colour filters on the engraver's camera and in steps to reduce the range of colour contrast of the copy. Negatives representing the images to be printed with each of the coloured inks are obtained by photographing the colour copy through colour filters. These filters, usually used in the form of thin sheets of dyed gelatin inserted into the lens, are complementary in colour to the coloured printing inks used.

Masking is the use of positive or negative images, taken from one or more of the set of colour-separation negatives and used in register with a given negative, to correct for the deficiencies in printing inks and colour of the copy. Common colour errors corrected by masking include the removal of excessive yellow values and magenta values from the blue (yellow printer) and green (magenta printer) negatives.

Colourplates may be made by the use of two general photographic methods—one indirect and one direct. The indirect method produces either continuous-tone negative images, from which halftone negatives are made, or continuous-tone negatives, from which continuous-tone positives are prepared. In the direct method, screen negatives are prepared directly from the copy through the colour-separation filters and a halftone screen onto a high-contrast panchromatic film or plate to produce a negative ready for transfer to the metal plate.

Proofing of colourplates

The proofing of halftone colourplates for wet printing on high-speed presses (when one colour does not have time to dry before the next is laid down) is a critical operation, for the proofing must be carried out under conditions simulating as closely as possible those that will be encountered on the production press. Specially built proof presses make this possible. In appearance they resemble four conventional press units placed end-to-end, and the sheet of paper is passed in turn over the four plates. However, because the production press employs not the original flat plates but curved duplicates made from them, and because ink and paper specifications are highly variable, exact duplication of production results in a proofing operation is difficult.

Elimination of moiré. A serious problem in colour reproduction is the occurrence of an interference pattern, or moiré, caused by the overprinting of the screens in the colourplates (a similar effect can be obtained by superimposing two pieces of window screening or fine net cloth). Because it is impossible to maintain printing register within the degree necessary to avoid such an effect, it is common practice to rotate the halftone screen when making the negatives so that each of the four plates has its screen pattern in a different position.

Electromechanical engraving machines—colour scanners. Reference has been made to devices for the electromechanical production of relief printing plates. The first of these utilized a heated pyramidal stylus, the motion of which was controlled by an electrical signal from a scanning photocell, to penetrate a plastic plate to a distance inversely proportional to the optical density of copy, thus burning out varying areas from the plate surface. In another machine of the same general type, an oscillating gouge cuts a halftone pattern in a flat plastic or metal plate, under control of a signal from a scanning photocell; in yet another a spiral groove, of varying width, is cut into the surface of a plastic plate wrapped on a rotating cylinder.

The colour scanner has been described elsewhere in this article. The first such devices were capable only of producing colour-separation negatives of the same size as the copy that was scanned. In later developments, circuits were provided to produce positive images, and mechanical or electronic devices were developed to allow enlargement or reduction of the size of the final image as compared with size of the original copy. When scanners were first made available, it was believed that their cost would limit their use to a few large plate-making establishments, but their acceptance exceeded expectations.

Production specifications. These include specifications for line plates, halftone specifications, and combination plates.

Line plates. In line illustrations all of the image areas are either black or white, and hence no halftone screen is required to copy them for use in making a printing plate. Suitable copy consists of line drawings, etchings, etc. The negative as it comes from the process camera is suitable to transfer the line image onto the metal.

Plate preparation, coating, burning in, etching, and finishing are essentially the same as for halftone plates. Certain specifications must be met, however. The nonprinting areas must be etched sufficiently deep to prevent the ink rollers from touching them on the press, and to prevent them from rubbing on the surface of the paper during wet colour printing. For presses with accurately adjustable ink rollers, the etch depth may be as little as 0.01 inch. The same depth is permissible in thin, wraparound press plates. For conventional printing presses, the minimum etch depth is about twice this. Plates that are to be duplicated by electrotype or stereotype processes may require slightly greater depths, although normal etching ordinarily is sufficient to produce good duplicates. Plates from which rubber duplicates are to be made will require etch depths as great as 0.045 inch.

Halftone specifications. Etch depths in halftone plates need not be as great as those in line plates, but the contour of the halftone dot and the depth of the etched areas are very important. Etch depth in highlight areas, the most critical portions of halftones, varies from 0.006 inch in a 65-line halftone to 0.002 inch in a 133-line.

Combination line-and-halftone plates. These plates must be prepared by assembling, in the negative form, the halftone and the line portions of the illustration and then, after transferring them onto the metal, etching them in two operations, so as to attain the best results for both portions. The powderless etching processes, however, permit easier etching of coarse-screen combination plates for use in newspapers. Combination line-and-halftone plates may also be produced by making two plates in separate operations and mounting them on a single block in proper position with respect to each other.

Engraving techniques applied to intaglio processes. Procedures similar to those described for production of letterpress printing surfaces are applicable to the production of intaglio printing surfaces. In intaglio, or gravure, printing, the image to be transferred to paper is etched or incised into the surface of the printing plate or cylinder. The entire surface is covered with ink, and by means of a doctoring, or wiping, operation, excess ink is removed from the surface, leaving only that which is retained in the image areas. Paper is brought into contact with the surface, and, under high mechanical pressures, the ink transfers from plate to paper.

Intaglio printing surfaces

Intaglio printing surfaces are of two general types: line intaglio (sometimes referred to as copperplate gravure), in which the ink-retaining image consists of discrete lines that may vary in width and depth; and gravure (also known as rotogravure), in which both continuous-tone and line copy are reproduced as a series of tiny cells etched into the printing surface. These cells commonly vary in depth, and hence in the volume of ink they will retain. Variations in density are produced on paper by the different amounts of ink that the cells transfer to paper.

In the rotogravure printing process, the walls surrounding each cell act as a support for the doctor blade that removes ink from the printing cylinder or plate surface.

Line intaglio. This process is widely used in the production of bank notes, securities, stamps, and engraved documents. The distinctive sharpness of fine lines and readily discernible differences in ink thickness that the process produces make it a preferred technique for production of bank notes and securities. These appearance characteristics cannot be readily counterfeited by photomechanical processes.

The printing surface is created either by mechanically scratching an acid-resistant ground from the plate surface, as described above, or by use of a photographic positive of the desired line pattern to prepare a photoresist image on the metal. The image is etched into the plate, using

the techniques of letterpress line etching, with maximum depth of etch usually less than 0.007 inch. Metals commonly used include steel, brass, and copper.

When a mechanical engraver is used to expose the metal for etching, a pointer or stylus is used to follow a usually enlarged pattern in a metal or plastic master stencil, causing a diamond stylus, which is in contact with the lacquer-covered plate surface, to remove the lacquer in a sharply defined pattern. The intaglio image is then prepared by etching the exposed metal with the appropriate chemicals.

In printing from intaglio forms, the plate is flooded with an ink of medium viscosity and the surface of the plate wiped clean with either a metal doctor blade or a piece of hard-surfaced paper. To minimize wear of the plate from the abrasion of the wiping mechanism, the surface is ordinarily protected by an electroplated chromium layer. Wiped free of excess ink, the plate is brought into contact with the paper surface. A roughly outlined relief image (counter) of the printing pattern is often used to provide high local pressures, forcing the paper into the ink-filled intaglio image. As the paper is pulled from the plate, capillary-attraction and surface-tension forces act to pull the ink from the plate. After drying, the image has a distinctive appearance in which the ink has appreciable thickness, and thin lines have less thickness than wider lines.

Gravure and rotogravure. The gravure printing process is one of the three major processes that are used for catalogs, magazines, newspaper supplements, cartons, floor and wall coverings, textiles, and plastics. The gravure printing process is done with flat plates or, more commonly, with cylindrical surfaces. A screen pattern is superimposed over all image areas; thus the edges of type or lines printed by gravure will have a rough, or sawtoothed, appearance. This does not detract from readability.

Early gravure work

The early work as described above formed the foundation for modern gravure engraving and printing. Karl Klič (also spelled Klietsch) of Bohemia, who was instrumental in making photogravure a practical commercial process, in 1878 exposed a positive transparency over carbon tissue, a film that was made of coloured gelatin sensitized with potassium dichromate and backed by a sheet of paper. The exposed film was pressed down on a copper plate that was coated with an even layer of resin or asphalt powder. The carbon tissue was developed in water, making the gelatin swell in inverse proportion to the exposure it had received. The plate was etched with ferric chloride in successive baths of varying strengths. Penetration of the tissue by the etchant, and hence the resulting depth of etch in the metal surface, was controlled by the degree of swelling of the gelatin. Klič's process produced sure and predictable results and became the preferred method for later workers.

In later developments, the irregular grain pattern, which is produced by use of resinous powders, was replaced by a regular overall pattern of intersecting lines, which is produced by exposing the carbon tissue to a glass screen bearing an overall pattern of clear lines, intersecting at right angles.

In rotogravure, separate negatives of type matter, other line copy, and continuous-tone copy are assembled and positioned according to a prepared layout. After the negatives are retouched, a continuous-tone positive is made and retouched to ensure desired density values.

A sheet of carbon tissue is next exposed under a gravure screen. This screen is a film or sheet of glass on which fine transparent lines, usually 150–175 per inch, cross at right angles to form opaque squares. The lines of the screen are positioned at an angle of 45° to the axis of the printing cylinder. Their purpose is to provide a support for the doctor blade, which wipes ink from nonprinting areas. The lines of the screen allow the light to penetrate to the film and harden the gelatin. The square "islands" remain soft.

Next, the continuous-tone positive is placed in contact with the carbon tissue and is exposed under an arc light. The soft squares are hardened in proportion to the amount of light that penetrates the varying grays of the positive. The carbon tissue is pressed with a rubber roller to a cylinder that has a polished electroplated copper surface. After adhering the exposed carbon tissue to the cylinder surface, it is "developed" with warm water, which has the effect of swelling and removing unhardened gelatin. The result is an image in hardened gelatin, which varies inversely in thickness according to the density of the photographic positive. The cylinder is rotated in a tray of ferric chloride, which produces an etched image in the copper surface. The squares are etched to varying depths, depending on the degree to which they were hardened. The crosslines of the screen, which were entirely light-hardened, are not etched at all. In this way, pits or wells of different depths are etched into the copper. For very long press runs, the cylinder can be strengthened by plating with nickel or chromium.

Rotogravure printing

In rotogravure printing, the cylinder usually is arranged so that during its rotary movement it passes through a trough filled with a thin solution of fast-drying ink. A thin steel doctor blade moves across the cylinder with a slight oscillating action and removes the ink from the surface, but not from the wells beneath. The cylinder then comes in contact with the paper, and the paper draws the ink out of the wells in the plate. After being printed, the paper shows through thin deposits of translucent ink, thus creating pale grays; heavier ink deposits from the more deeply etched wells appear correspondingly opaque. Thus a full range of tonal values can be printed. Since the ink used in rotogravure printing is quite fluid, it penetrates through the pores of the paper surface, obliterating the screen pattern. In reproducing illustrations, gravure comes closest to simulating continuous-tone copy. In colour printing, a separate cylinder is prepared for each colour.

Other methods. In the so-called Dultgen halftone intaglio process, which is widely used in colour work, two positives are made from the continuous-tone copy, one through a halftone screen or a special contact screen and the other without a screen. The carbon tissue is first exposed to the screened positive, which produces an image of dots of varying sizes, then to the continuous-tone positive, which produces differing degrees of hardening of the dot image. When etched, the dots are of differing sizes and of differing depths. This method thus uses two methods for controlling tonal values.

The Henderson process, sometimes referred to as "direct transfer," or "inverse halftone," gravure, has won some acceptance in the printing of packaging materials. Retouched continuous-tone positives are used in preparation of halftone negatives and, by a contact-printing operation, halftone positives. These positives show dot size variations proportional to the desired print density. The cylinder is coated with a cold-top photoresist, as in letterpress engraving. This resist is then exposed to ultraviolet light, through the positives, and the image developed. The cylinder may then be etched either in ferric chloride solution or in a powderless etching bath, similar to that used for letterpress photoengraving. Tonal variation in the resulting print on paper is caused almost entirely by variation in the area of dots. (P.F.B.)

BIBLIOGRAPHY

Printing: The history of printing has been the subject of many volumes. See D.C. MCMURTRIE (ed.), *The Invention of Printing: A Bibliography* (1962). Books retracing the origins of typography include: L.P.V. FEBVRE and H.J. MARTIN (eds.), *L'Apparition du livre* (1958); P. BUTLER, *The Origin of Printing in Europe* (1966); E.G. DUFF, *Early Printed Books* (1893, reprinted 1968); and T.L. DE VINNE, *The Invention of Printing* (1876, reprinted 1969). Biographical works on contributors to the invention of printing are: S. JENNETT, *Pioneers in Printing* (1958); J. GUIGNARD, *Gutenberg et son oeuvre* (1963); G.P. WINSHIP, *Gutenberg to Plantin* (1968); and A. VAN DER LINDE, *De Haarlemsche Costerlegende* (Eng. trans., *The Haarlem Legend of the Invention of Printing, by Lourens Janszoon Coster,* 1871, reprinted 1968). The subject of an entire family devoted to the typographic art is treated by A.J. GEORGE in *The Didot Family and the Progress of Printing* (1961). The various stages of evolution of the techniques of printing are developed in: S.H. STEINBERG, *Five Hundred Years of Printing,* 2nd ed. rev. (1962); M. AUDIN in M. DAUMAS (ed.), *Histoire générale des techniques,* 3 vol. (1962–68; Eng. trans. of vol. 1-2, *A History of Technology and Invention,* 1965–69); I.B. SIMON, *The Story of Printing from Wood Blocks to Electronics* (1965); J. WATSON, *The History of the Art of Printing* (1713, reprinted 1965); C.A.

CLAIR, *Chronology of Printing* (1969); and I. THOMAS, *The History of Printing in America, with a Biography of Printers* (1874; 2nd ed., 2 vol., reprinted 1967). One particular aspect of the history of printing is treated in C.H. BLOY, *A History of Printing Ink: Balls, and Rollers, 1440–1850* (1967). Both history and technology are covered in P. LUCKOMBE, *The History and Art of Printing* (1771, reprinted 1965); G. BAUDRY and R. MARANGE, *Comment on imprime*, 3rd ed. (1966), and also in G.A. STEVENSON, *Graphic Arts Encyclopedia* (1968). Later sources include BENJAMIN FRANKLIN V (ed.), *Boston Printers, Publishers, and Booksellers, 1640–1800* (1980); STANLEY MORISON, *Selected Essays on the History of Letter-Forms in Manuscript and Print*, 2 vol., ed. by DAVID MCKITTERICK (1981); and MIRIAM USHER CHRISMAN, *Lay Culture, Learned Culture* (1982), all three covering different aspects of printing history.

The specialized language of the profession and the technical terms that are appropriate to it are collected in the following dictionaries: *Pocket Encyclopedia of Paper and Graphic Arts Terms* (1960); W.M. PEPPER, *Dictionary of Newspaper and Printing Terms: English–Spanish, Spanish–English* (1959); E.M. ALLEN, *Harper's Dictionary of the Graphic Arts* (1963); W.A. SAVAGE, *Dictionary of the Art of Printing* (1841, reprinted 1966); W.W. PASKO (ed.), *American Dictionary of Printing and Bookmaking* (1894; reprinted with new introduction, 1967); C.T. JACOBI, *The Printer's Vocabulary* (1888, reprinted 1969); and R. HOSTETTLER, *Technical Terms of the Printing Industry*, 5th rev. ed. (1969), which gives the equivalents of the words in five languages. In addition to general technical works on printing, such as F. PATEMAN and L.C. YOUNG, *Printing Science* (1963); and R.R. COUPE, *Science of Printing Technology* (1966); there are numerous treatises for professionals, such as: C.A. HURST and F.R. LAWRENCE, *Letterpress, Composition and Machine Work* (1963); E.A.D. HUTCHINGS, *Printing by Letterpress* (1964); V.S. GANDERTON and H. COPELAND, *Cylinder Presses*, 2nd ed. (1965); L. HEITNER, *Introduction to Offset* (1964); R.R. KARCH and E.J. BUBER, *Graphic Arts Procedures: The Offset Processes* (1967); J.E. COGOLI, *Photo-Offset Fundamentals*, 2nd ed. (1967); A. KINSEY, *Introducing Screen Printing* (1967); A. KOSLOFF, *Photographic Screen Process Printing*, 3rd ed. (1968); A.H. PHILLIPS, *Computer Peripherals and Typesetting* (1968); E.A. APPS, *Printing Ink Technology* (1958) and *Ink Technology for Printers and Students* (1963). On the economics of the printing industry, see *Printing Industry in Britain, U.S.A. and Japan* (1964), ed. by the NATIONAL PRODUCTIVITY COUNCIL, NEW DELHI. Some philosophical and social elements are discussed in MARSHALL MCLUHAN, *The Gutenberg Galaxy: The Making of Typographic Man* (1962); J. CARTER, *Printing and the Mind of Man* (1967); and W.M. IVINS, *Prints and Visual Communication* (1969). See also GEOFFREY A. GLAISTER, *Glaister's Glossary of the Book*, 2nd ed. (1979); JANET N. FIELD (ed.), *Graphic Arts Manual* (1980); PATRICIA B. MINTZ, *Dictionary of Graphic Arts Terms* (1981); and EDWARD BOOTH-CLIBBORN and DANIELE BARONI, *The Language of Graphics*, trans. from Italian (1980).

Typography: There is a vast literature on typography and printing history. *A Bibliography of Printing*, comp. by EDWARD C. BIGMORE and C.W.H. WYMAN, 2nd ed., 2 vol. (1880–86), was still useful enough to merit reprinting in 1945. Several learned and technical journals print annual bibliographies. See especially *Studies in Bibliography* and *Publications of the Modern Language Association*, which emphasize articles dealing with analytic bibliography and printing history. The best brief history in English is SIGFRID H. STEINBERG, *Five Hundred Years of Printing*, rev. ed. (1962). CURT BUHLER, *The Fifteenth-Century Book* (1960), is an excellent survey of early printing and publishing practice. JOSEPH MOXON, *Mechanick Exercises on the Whole Art of Printing* (1683–84), is the earliest comprehensive manual on printing, typography, and type making. The 1962 edition of HERBERT DAVIS and HARRY CARTER, with an excellent introduction and full annotation, gives considerable insight into the typography and printing of the day; there were no significant changes from the invention, *c.* 1450, until the early 19th century. DANIEL B. UPDIKE, *Printing Types*, 3rd ed., 2 vol. (1962), is a thorough and interesting history of the development of type design from the beginnings to about 1930—highly personal, highly dogmatic, but the classic work on the subject. Of the many books and articles by STANLEY MORISON, three are especially noteworthy to the nonprofessional reader: *First Principles of Typography*, 2nd ed., with postscript (1967), is an expanded version of his Britannica article on "Typography," which became the definitive statement of his views on the subject, and has been translated into several languages. *The Typographic Arts* (1950), contains two essays on, *inter alia*, the interrelationship between calligraphy, engraving, and type design. *The Typographic Book*, ed. by KENNETH DAY (1962), is an expanded version of Morison's *Four Centuries of Fine Printing* (1924). It contains good reproductions of specimen titles and text pages spanning 1450–1935. HELLMUT LEHMANN-HAUPT (ed.), *The Book in America*, 2nd ed. (1951), is a historical survey of American printing and publishing from the beginning to the present. KENNETH DAY (ed.), *Book Typography, 1815–1965, in Europe and the United States of America* (1966), contains uneven but generally good articles on its subject. JOHN CARTER and PERCY MUIR (eds.), *Printing and the Mind of Man* (1967), the catalog of two major exhibitions held in London for an International Printing Exhibition, contains much technical information on the technological development of printing and type founding from the invention to today, as well as notes on books important for their intellectual or aesthetic impact. HENRI J. MARTIN and LUCIEN FEBVRE, *L'Apparition du livre* (1958), while heavily French in its emphasis, is a stimulating and original history of the social, economic, cultural, and technical evolution of the book trades from the manuscript period to the 19th century. Of the many modern manuals on typography, mainly reflecting the Bauhaus school, two that are representative and better than average are JAN TSCHICHOLD, *Typographische Gestaltung* (1935; Eng. trans., *Asymmetric Typography*, 1967); and EMIL RUDER, *Typographie* (1967). The latter has text in German, French, and English, and also shows Dadaist and other modern schools. Tschichold became converted to traditional typography, and in *Designing Books* (1951), gives an excellent exposition of his later views. HUGH WILLIAMSON, *Methods of Book Design*, 2nd ed. (1966), is a full and good survey of modern book design and production methods. *The Penrose Annual*, published in London, has technical articles on new developments in design and processes as well as good essays on the history and aesthetics of printing. The *Gutenberg Jahrbuch*, emanating from Mainz, the cradle of printing, emphasizes incunabula but includes articles on later printing, publishing, and binding. Valuable information is found in HERBERT LECHNER, *Geschichte der modernen Typographie: von der Steglitzer Werkstatt zum Kathodenstrahl* (1981); ERIK LINDEGREN, *ABC of Lettering and Printing* (1982); BILL GRAY, *Tips on Type* (1983); and *Words of the World: A Typographic Demonstration of World Alphabets and Languages* (1983).

Photoengraving: J.S. MERTLE and G.L. MONSEN, *Photomechanics and Printing* (1957), a good basic text, emphasizing equipment and techniques of use, now obsolete in many technical details; W.J. SMITH, E.L. TURNER, and C.D. HALLAM, *Photoengraving in Relief*, 3rd ed. (1951), basic information, should be supplemented with more recent texts; R.W.G. HUNT, *Reproduction of Colour*, 2nd ed. (1967), an authoritative presentation of fundamentals of colour printing and platemaking; F.G. WALLIS and R.V. CANNON, *Letterpress Platemaking* (1969), a recent general work on the platemaking process, recommended as both comprehensive and current; H.M. CARTWRIGHT, *Ilford Graphic Arts Manual*, vol. 1, *Photo-engraving* (1962), an excellent basic text, combining historical data and practical information, with bibliographies for further study; *Photogravure*, 2nd ed. (1939), a good presentation of principles and basic technology; and with R. MACKAY, *Rotogravure* (1956), a recent work on the technology of gravure; H. DENISON, *A Treatise on Photogravure* (1894), a classic volume dealing with this process. See also ROBERT M. SWERDLOW, *The Step-by-Step Guide to Photo-Offset Lithography* (1982).

Printmaking

To the modern reader, the word print might suggest mechanically mass-produced commercial products, such as books, newspapers, and textiles. In this article the print refers to the original creation of an artist who, instead of the paintbrush or the chisel, has chosen printmaking tools to express himself.

The fine print is a multiple original. Originality is generally associated with uniqueness, but a print is considered original because the artist from the outset intended to create an etching, woodcut, or other graphic work and thus conceived his image within the possibilities and limitations of that technique. Without doubt, early printmaking was strongly influenced by a desire for multiple prints. Artists quickly discovered, however, that when a drawing is translated into a woodcut or engraving it takes on totally new characteristics. Each technique has its own distinctive style, imposed by the tools, materials, and printing methods. The metamorphosis that takes place between drawing and print became the strongest attraction for the creative artist. It is important to understand that the artist does not select his printing method arbitrarily but chooses the one in which he can best express himself. Thus, any of the proofs printed from an original plate is considered an original work of art, and, although most fine prints are pulled in limited quantities, the number has no bearing on originality, only on commercial value.

What is the difference between a reproduction and an original print? In the very early days of printmaking this was not a serious problem because the print was not looked upon as a precious art object, and prices were low. The question of originality became an issue only in the 18th century, and, in the 19th century, artists started to hand sign their prints. Since then, the signed print has been accepted by most people as the proof of its originality.

With regard to the name with which he signed his works, the Japanese artist followed a bewildering custom: he adopted and discarded names at will. If he admired another artist, he simply adopted his name. Thus, in the art history of Japan, it is common to find several unrelated artists bearing the same name and one artist bearing many names; during his long life, Hokusai, for example, used about 50 different names. In fact, a signature by itself means little or nothing. For instance, Pablo Picasso issued many signed reproductions of his paintings; on the other hand, many of his original etchings have been published in split editions, some signed, some not. These unsigned etchings are original, while the signed reproductions are not. The crucial difference is that Picasso made the plate for the original print, while the signed reproduction was photomechanically produced.

In 1960 the International Congress of Plastic Arts drafted a resolution intended to regulate contemporary prints. The crucial paragraph reads:

> The above principles apply to graphic works which can be considered originals, that is to say to prints for which the artist made the original plate, cut the woodblock, worked on the stone or any other material. Works that do not fulfill these conditions must be considered "reproductions."

Although this is a straightforward statement, later developments have proved it to be highly controversial. Since the rise of the Pop and Op movements, a great number of photographically produced prints have been published and sold as signed originals. Because museum curators, art critics, and artists have not taken a firm stand on the question, any print that the artist declares to be original is now accepted as such, regardless of how it was made. Although the art world is divided on the solution, nearly everybody agrees that something should be done to clarify the situation. The state of New York, for example, has passed a law requiring complete disclosure by the dealer of how, and by whom, the print was made.

Many artists believe that the answer lies in the giving of honest information. In the 17th and 18th centuries in the West, most prints carried all the relevant information on their margins. The name of the individual was followed by a Latin abbreviation indicating his role in the work. Common examples are *del.* (*delineavit*): "he drew it"; *imp.* (*impressit*): "he printed it"; and *sculp.* (*sculpsit*): "he engraved it." This type of information, together with the total edition number, should be furnished by the artist or the dealer to the buyer. Clearly, it is impossible to make completely rigid rules to define originality. Probably the most realistic solution is to establish degrees of originality, based on the degree of the artist's participation in the various steps in the creation of the finished print.

There may also be confusion about edition numbering. In contemporary printmaking, an original print in limited edition should carry information about the size of the total edition and the number of the print. A problem can arise because, in addition to the regular edition, there are "artist's proofs" or the French "H.C." (*hors de commerce*) proofs. These are intended for the artist's personal use and should be no more than 10 percent of the edition; but, unfortunately, this practice is often abused. All of the prints pulled between working stages are called "trial proofs." These can be of great interest because they reveal the artist's working process and of great value because the number of proofs is small.

With prints of old masters in the West, originality is a very complex and difficult issue. These artists did not publish their prints in limited editions but printed as many as they could sell and without signing or numbering their works. There are arguments even between experts about the authenticity of many old prints. Important works of the masters are documented in catalogs and, although these must be revised from time to time, they furnish the only firm information available. After the edition is printed, the modern artist usually either destroys the plate or marks ("strikes") it in a distinctive manner to guarantee that any reprint from the plate is identifiable.

The 19th-century U.S. painter and etcher James McNeill Whistler was one of the first Western artists to hand sign his prints. Signing is now regulated by a convention. Upon completing the edition, the artist signs and numbers each print. Usually the signature is in the lower right corner; the edition number is on the left. Some artists put the title in the centre.

This article is divided into the following sections:

Major techniques of printmaking 113
- Relief processes 113
  - Woodcut
  - Colour woodcut
  - Wood engraving
  - Linoleum cut
  - Metal cut
  - Cardboard (paper) cut
  - Relief etching
  - Rubbing
  - Dotted print (criblé)
- Intaglio processes 115
  - Engraving
  - Drypoint
  - Mezzotint
  - Crayon manner and stipple engraving
  - Etching
  - Metal graphic
  - Printing by intaglio processes
- Surface-printing processes 119
  - Lithography
  - Stencil processes

Special techniques 121
Process prints 122
Contemporary experimentation 122
Mounting and care of prints 122
History of printmaking 123
Printmaking in the 15th century 123
Germany
Italy
Other countries
Printmaking in the 16th century 125
Germany
Other countries
Trends in the late 16th century
Printmaking in the 17th century 126
Portrait engraving
Flemish printmaking
European etching
Japanese Ukiyo-e prints
Printmaking in the 18th century 129
Italy
England
Spain
France
Japan
Printmaking in the 19th century 130
France
Japan
Other countries
Printmaking in the 20th century 132
France
Germany
Other countries
Bibliography 134

Major techniques of printmaking

The techniques of printmaking are divided into three major processes: relief, intaglio, surface. The surface processes are subdivided into two categories: planographic (lithography) and stencil methods. The methods are often combined.

RELIEF PROCESSES

In relief processes, the negative, or nonprinting part of the block or plate, is either cut or etched away, leaving the design standing in relief. Or, instead of cutting away the background, the relief print can be created by building up the printing surface. The relief is the positive image and represents the printing surface. The most familiar relief-printing materials are wood and linoleum, but many other materials can be used, such as aluminum, magnesium, and plastics. Any metal or plastic plate incised or worked in relief can be first inked in the depressions (intaglio inked) and then surface rolled, thus combining relief and intaglio processes.

Relief printing lends itself particularly to a bold conception of design, expressed more in areas than lines. This varies, however, depending on the material used: metal allows more intricate detail than wood, for example.

Woodcut. Woodcut, which appeared in the 8th century in the East and in the early 15th century in the West, is the earliest known relief-printing method. In this method, the design is first either painted directly onto the wood block or pasted on it. Then the surface of the wood is cut away around the design. For fine details and outlines the knife is used; larger areas are removed with gouges. The depth of the relief depends on the design: open areas must be cut deeper than the fine details so that the roller will not deposit ink in these areas. Although woodcuts are generally conceived in bold lines, or large areas, tonal variations can be achieved with textures, a variety of marks made with gouges, chisels, or knives. In contemporary woodcuts many other methods, such as scraping, scratching, and hammering, are also used to create interesting textures.

Originally, woodcut was a facsimile process; *i.e.*, the cutting was a reproduction of a finished design. With most contemporary woodcuts, however, the artist creates his design in the process of cutting.

Wood for blocks

As wood is a natural material, its structure varies enormously and this exercises a strong influence on the cutting. Wood blocks are cut plankwise. The woods most often used are pear, rose, pine, apple, and beech. The old masters preferred fine-grained hardwoods because they allow finer detail work than softwoods, but modern printmakers value the coarse grain of softwoods and often incorporate it into the design.

The printing of woodcuts is a relatively simple process because it does not require great pressure. Although presses are used, even hand rubbing with a wooden spoon can produce a good print. The ink used to print woodcuts must be fairly solid and sticky, so that it lies on the surface without flowing into the hollows. The printing ink can be deposited on the relief either with dabbers or with rollers. Japanese rice or mulberry papers are particularly suitable for woodcuts because they make rich prints without heavy pressure.

Colour woodcut. The standard procedure for making a woodcut with two or more colours is to cut a separate block for each colour. If the colour areas are distinctly separated and the block is large, one block can be used for more than one colour. All blocks must be the same size to assure that in the finished print the colours will appear in their proper relation to one another, that is, properly registered.

The first, the key block, is generally the one that contains most of the structural or descriptive elements of the design, thus serving as a guide for the disposition of the other colours. After the key block is finished and printed, the print is transferred to the second block. This procedure is repeated until all of the blocks are finished.

The registering system depends on the method of printing used. On a press the registering presents no problem:

By courtesy of the Museum of Modern Art, New York, gift of Curt Valentin



"Head of Ludwig Schames," woodcut by Ernst Ludwig Kirchner, 1918. 58.4 cm × 26 cm.



Wood engraving by William Blake, 1820–21, for R.J. Thornton's *Pastorals of Virgil*. 3.5 cm × 7.2 cm.
By courtesy of the trustees of the British Museum; photograph, J.R. Freeman & Co. Ltd.

the wood block is locked into position and the uniformly cut paper is automatically fed into the proper position by the press. For hand rubbing, several registering methods can be used. One method uses a mitred corner nailed to a table or special board. A sheet of paper is attached to one side of this corner, after which the wood block is placed securely in position and the print is made. Once the first colour has been printed, the paper is folded back and the first block is replaced with the second, and so on.

In woodcut colour printing, the artist must consider whether he can print wet on wet or whether the print should dry before it is overprinted. Usually a second colour can be printed immediately but, if the ink deposit is heavy, the print will have to dry before additional colours can be printed. This problem arises mainly with oil colours, which dry more slowly than water-base colours. When using oil paints, the artist has to understand how variations in viscosity affect the overprinting of colours.

Use of small movable blocks

Movable small blocks have also come to be used by a number of printmakers. These involve some planning in order to print them in register with the large blocks. The easiest way is to put a light cardboard that is exactly the size of the main block (the key block) in position. Once the small blocks are registered, their location can be marked on the cardboard. Then the small blocks can be glued down to the cardboard in order to avoid the danger of shifting.

The conception and technique of the Japanese colour woodcut was totally different from that of the European woodcut. Except for chiaroscuro prints, no real colour woodcut existed in Europe before the 19th century. In the West, the woodcut was primarily a reproductive facsimile process: usually, the artist made a completed drawing that was copied by the cutter. The Japanese print, on the other hand, was the result of intricate, perfectly coordinated effort by the designing artist, the cutter, and the printer. Instead of painting a complete picture to be copied, the artist furnished a separate drawing for each colour. The engraver or cutter pasted each drawing on a wood block and cut away the white (negative) part. In this process the drawing was destroyed. Printing started only after all of the blocks had been cut. As the Japanese used water-base colours, often blending tones, printing itself was a very delicate and crucial operation, requiring perfect coordination and speed. Only after the completion of this process could the artist see the total image.

Wood engraving. Wood engraving is a variation of woodcut. The main difference is that, for wood engraving, the block—usually pear, apple, cherry, sycamore, or beech—is cut cross-grained rather than plankwise; on the end-grain block the artist can thus cut freely in any direction, allowing him to do much more intricate work with much finer tools. The image is created by fine white lines and textures. On most wood engravings, the whites appear as the positive image against a dominant black. The blocks are usually cut at the same height as printing type so that they can be printed on a press. Invented in the 18th century, wood engraving was primarily used by illustrators.

Linoleum cut. Since linoleum is easy to cut and does not have a grain, the linoleum cut often is used to introduce children to printmaking. The process was held in low esteem until, in the 1950s, Pablo Picasso made a series of brilliant colour linoleum cuts.

The printing of linoleum cuts is similar to the printing of woodcuts or wood engravings. They can be printed by hand rubbing or, properly mounted, can be printed on a press. The colour printing process follows the woodcut principles.

Metal cut. At times artists have used soft metals, such as lead or zinc, to make prints that are similar to woodcuts or wood engravings. In the 19th century, lead cuts were often used for newspaper illustrations. The distinguished Mexican artist José Guadalupe Posada, for example, used lead frequently for his prints. Lead was used primarily because it was inexpensive and easy to work. Because metal cuts were printed like woodcuts or wood engravings, it is often difficult to tell from the print which material was used.

Cardboard (paper) cut. Elementary school children are often introduced to printmaking by making cardboard cuts, and sophisticated artists use the same material to print complex abstract images. Cardboard and paper are not only inexpensive, readily available, and workable with simple tools but, when properly prepared, have also proved to be remarkably durable. Cardboard cuts can be made either by building up or cutting out. In the first process, cutout pieces are glued to a support. When the plate is finished, it is coated with a plastic varnish to make sure the surface is tough and nonabsorbent. In the cutting-out method a heavy laminated cardboard is used, and the cutout sections are simply peeled off to the desired depth. When finished, the cut is varnished. The printing of cardboard plates follows the same principle as woodcuts or linoleum cuts.

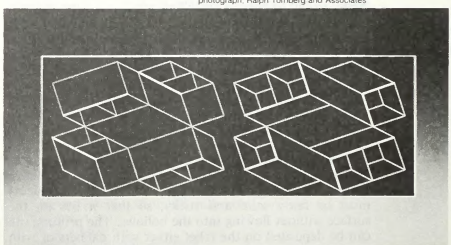
Advantages of cardboard and paper

Relief etching. When large areas of a metal plate are etched out (see below *Etching*), leaving the design in relief to be surface printed, the process is generally called relief etching. Usually the method is used for areas, but it can be also used for lines. The English artist and poet William Blake was the first printmaker to experiment extensively with relief etching. He devised a method of transferring his handwritten poems, together with the illustrations, onto the metal plate to be etched.

In contemporary printmaking, relief etching is used extensively for colour printing. The different levels of the plate can be inked with different colours. Relief etching is also a popular method of making inkless intaglio prints (shallow bas-reliefs on paper).

Rubbing. Simply by placing a fine paper over an incised or carved surface and rubbing the paper with heelball (wax and carbon black) or daubing it with special ink, an artist can use practically any surface for printing—including, as in Japan, the body of a fish. Rubbings were probably the earliest prints made by man. In India rubbings were made of tombstones and temple bas-reliefs, and in China rubbings were used to reproduce calligraphy as early as the 2nd century AD. In addition to fish rubbings, the Japanese made rubbings of metal ornaments.

By courtesy of Multiples, Inc., New York and Los Angeles; photograph, Ralph Tornberg and Associates



"Duo H," inkless intaglio print (relief of design etched on plastic and mounted on Plexiglas) by Josef Albers, 1966. Image 12.7 cm × 34.3 cm.



Scenes from the tomb of the Wu family in Shantung, China, stone rubbing, AD 147 (Han dynasty). 7.5 m × 6.3 m.
By courtesy of the Philadelphia Museum of Art, gift of Horace H.F. Jayne

Today many museums sell rubbings of bas-reliefs in their collections. In the United States rubbings often are made of colonial and early 19th-century gravestones, and in Europe they are applied to brass plaques mounted in stone slabs.

Dotted print (criblé). A traditional technique of the goldsmith long before engraving for printing purposes was developed, criblé was also used to make the earliest metal prints on paper. Criblé was a method of dotting the plate with a hand punch; with punch and hammer; with a serrated, flatheaded tool called a matting punch; with various gouges; or, sometimes, with a hollow, circular-headed ring-punch. Criblé plates were relief printed like woodcuts. On most dotted prints, a black background dominates a fine lacelike design.

INTAGLIO PROCESSES

Intaglio printing is the opposite of relief printing, in that the printing is done from ink that is below the surface of the plate. The design is cut, scratched, or etched into the printing surface or plate, which can be copper, zinc, aluminum, magnesium, plastic, or even coated paper. The printing ink is rubbed into the incisions or grooves, and the surface is wiped clean. Unlike surface printing, intaglio printing—which is actually a process of embossing the paper into the incised lines—requires enormous pressure. The major working methods for intaglio printing are engraving, etching, drypoint, and mezzotint. Intaglio processes are probably the most versatile of the printmaking methods, as various techniques can produce a wide range of effects, from the most delicate to the boldest. The intaglio print also produces the richest printed surface, as it is three-dimensional.

Pressures required for intaglio printing

Engraving. In engraving, the design is cut into metal with a graver or burin. The burin is a steel rod with a square or lozenge-shaped section and a slightly bent shank. The cutting is accomplished by pushing the burin into the metal plate. The deeper it penetrates into the metal, the wider the line; variations in depth create the swelling-tapering character of the engraved line. After the engraving is finished, the slight burr raised by the graver is cleaned off with a scraper. The engraved line is so sharp and clean that it asserts itself even if cut over a densely etched area. In the print, the engraved line is notable for its precision and intensity. In engraving, the hand does not move freely in any direction but pushes the graver forward in a line; a change of direction is achieved by the manipulation of the plate with the other hand. Although copper, zinc, aluminum, and magnesium plates are used—and in the past soft iron and even steel were used—the best all-around metal is copper. It has the most consistent structure and is neither too soft nor too hard.

Drypoint. Next to engraving, the drypoint is the most direct of the intaglio techniques. In printing, however, it represents the opposite end of the spectrum. Engraving is precise; drypoint is rugged, warm, and irregular.

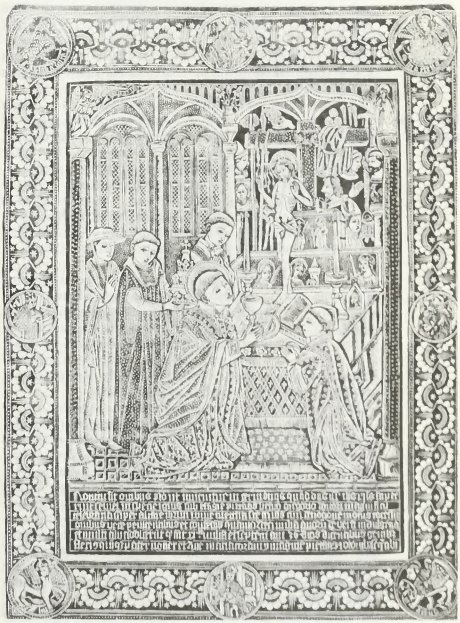
Drypoint is made by scratching lines into metal plates with steel- or diamond-point needles. In this method the penetration into the plate is negligible; it is the metal burr raised by the point that holds the ink. Because the burr is irregular, it prints as a soft, velvety line. The angle of the needle has much more effect on the width of the line than the pressure does. If the needle is perpendicular to the plate, it throws burr on both sides, which then produces a thin double line; for wide lines the optimum angle is 60 degrees. Many artists use an electric graver to make drypoints. The oscillating point of the tool punches little craters into the plate. Because the line consists of thousands of these small craters, it is richer than the conventional scratched line made by the needle and stands up better to printing.

Importance of needle angle

Copper plate is the best for drypoint. The plates are fragile because the burrs are easily flattened down by the printing pressure. Even a too vigorous wiping can damage a plate. Thus, unless the artist is satisfied with a very limited number of proofs (three or four), the plate must be faced with steel, a process in which steel is deposited by electrolytic means on the copper plate. This coating is very thin and, if it is properly done, the burrs are hardened without affecting printing quality. Zinc and aluminum, however, cannot be steel-faced.

Mezzotint. In mezzotint the metal plate is roughened with fine burrs until it prints a rich, velvety black. The

By courtesy of the Metropolitan Museum of Art, New York, Harris Brisbane Dick Fund, 1924



"The Mass of St. Gregory," criblé, or dotted, print by an unknown German artist, 15th century. 35.2 cm × 25.1 cm.



"Self-Portrait with Bowler Hat," drypoint by Max Beckmann, 1921. 31.3 cm × 24.5 cm.
By courtesy of the Museum of Modern Art, New York

plate is then worked back toward the lighter values with scrapers and burnishers. For this reason, mezzotint is also called *manière noire,* or the "black manner."

Mezzotint flourished throughout the 18th and 19th centuries and was primarily used for portraits or to reproduce paintings. None of the important printmakers of the past used the technique. After the invention of photoengraving, the technique of mezzotint was nearly forgotten, but a few printmakers have started to work again with this exotic medium.

Use of the rocker

The first step in preparing a mezzotint plate is to rough up the whole plate surface as evenly as possible. The tool used is the rocker, a blade with a curved serrated edge. The rougher the rocker, the heavier is the burr. The rocker is held with its cutting edge at a right angle to the plate, and the curved edge is rocked systematically over the entire surface. If this is properly done, the entire plate is covered with uniform burrs. Then the work with scrapers and burnishers begins. Where lighter tones are desired, the burr is gradually removed, and in the white areas the plate is burnished back to its original finish.

As with drypoint, mezzotint plates must be steel faced if a large edition is desired. The printing of mezzotints differs slightly from the printing of etchings or engravings. Since the layer of burr on the mezzotint acts as a blotting paper, the ink must be selected with this fact in mind. The inking and wiping must be done gently with soft rags. Printing pressure should be considerably less than that used for engravings or deeply etched plates.

Crayon manner and stipple engraving. Invented in the 18th century, crayon manner was purely a reproduction technique; its aim was the imitation of chalk drawings. The process started with a plate covered with hard ground (see below *Etching*). The design was created using a great variety of etching needles (some of them multiple). After the design was etched in, the ground was removed and the design further developed with various tools. Fine corrections and tonal modifications were made with scrapers and burnishers. Finally, engraving was used for additional strengthening of the design. Pastel manner is essentially the same as the crayon manner except that it is usually used to imitate pastel drawings.

Stipple engraving, also a reproduction method, is closely related to the crayon manner. The exact date of its invention is not known, but it is reasonably certain that it came after the crayon manner. The first step in stipple engraving was to etch in the outlines of the design with fine dots made either with needles or with a roulette, a small wheel with points. The tonal areas were then gradually developed with tiny flick dots made with the curved stipple graver. For very fine tonal gradations, roulettes were also used. The only artist of any importance to use pure stipple engraving was Giulio Campagnola in the 16th century.

Etching. Etching is a process in which lines or textures are bitten (etched) into a metal plate with a variety of mordants (acids). The metal plate is first covered with an acid-resistant coating (ground). The design is then scratched or pressed into the ground, exposing the metal in these areas. Finally, the plate is submerged in an acid solution until the desired depth and width in the exposed areas is reached.

Although the basic principle of etching is very simple, there are many possible variations that have a strong influence on the final result. The materials themselves offer a wide range of possible variations: for example, copper, zinc, aluminum, or magnesium plates can be used; and nitric acid, hydrochloric acid, or ferric chloride can be used for the etching process. Other variations include the strength of the mordants, the biting time, the kinds of grounds and the ways in which they can be worked, and, finally, all the possible methods of printing.

Although all of these matters seem purely technical, every tool or material that is used, every step that is followed, is an integral part of the creative process. The biting action of the acid is just as much part of the drawing as is the incising into the ground. The selection of the paper or the method of wiping the plate can completely change the nature of a print.

Hard-ground etching. Any acid-resistant coating used to make an etching is called a ground. In the past a great variety of different grounds were used, and each master had his own formula. Most of them had wax as a basis, combined with various oils and varnishes. Today, the most commonly used ground consists of two parts Egyptian asphaltum, two parts beeswax, and one part resin.

By courtesy of the Art Institute of Chicago, Mr. and Mrs. Potter Palmer Collection



"St. John on the Isle of Patmos," engraving by Master E.S. (active *c.* 1440–*c.* 1467). 20.6 cm × 14.1 cm.

These ingredients are either dissolved and mixed or fused by heat. Ground comes in either lump or liquid form.

The plate is cleaned before the ground is applied because grease or dirt can affect the ground's adhesion, making it peel or crack. If ground in solid form is used, it is melted on low heat and rolled out evenly. Liquid ground is brushed on the plate, and then the ground is heated to make it more even and to evaporate the solvents. In both cases, after the plate cools, the ground should be solid rather than sticky.

Normally, a good ground is dark enough to offer sufficient contrast with the plate to see the work. If, however, a black ground is desired, it can be achieved by darkening the ground with the smoke of a candle.

Variety of modern etching tools

In etching the ground, any number of tools and instruments may be employed. The old masters were restricted, but the contemporary printmaker uses a whole arsenal, including electrical drills and gravers. The line produced by the etching needle is threadlike and uniform in thickness. The exception is a line made by the tool called *échoppe,* developed by Jacques Callot, which may be used to imitate the engraved line. Other instruments are used to introduce a great variety of marks. The character of the etching is further influenced by the choice of the metal and the type of acid used. For controlled, regular bite, it is common to use Dutch mordant (nine parts of water saturated with potassium chlorate to one part of hydrochloric acid) on copper. For a rugged, irregular bite, nitric acid (one part to nine parts of water) is used on zinc. A plate can be etched in stages by covering some of the already etched areas with stop-out varnish (rosin dissolved in alcohol), which resists the acid, and then etching the rest for a longer period. This procedure can be repeated many times. Most artists develop their plates by repeated bites. After the etching is finished, the ground is removed with solvent (such as kerosene or benzene), and the plate is printed.

The first print is a state, or trial, proof. If further work is desired, the plate is cleaned and covered again with ground, the previous work remaining visible through the new ground. The whole process is repeated as many times as is necessary.

Soft-ground etching. Soft-ground etching is basically the same as hard-ground etching except that the ground contains about one-third grease, which keeps it in a semihard, or tacky, condition.

Initially, in the 19th century, soft ground was used primarily for offset drawings. The artist placed a paper on the grounded plate and made his drawing on the paper with a sharp pencil or other drawing instrument. Under the pressure, the paper picked up the ground and produced a soft granular line. Then the plate was etched normally with a fairly weak acid.

Soft ground has come to be used more often to etch various textures into the plate. Textured materials are placed on the soft ground and the plate run through the press. A thin, even ground picks up the finest textures. The design is controlled by applying a stop-out varnish to areas that should not be etched. The remaining textures are etched into the metal in the same way as in conventional hard-ground etching. This technique lends itself well to collage-type effects on the plate.

Relief etching. To make a relief etching, the areas not to be removed by acid are protected with liquid ground or varnish. The varnish used has to be tough (asphaltum, or ground) because the relief bite takes a long time, and when large areas are bitten, the plate has a tendency to heat up. If various levels are desired, relief etching can be done in stages, as in regular etching.

Aquatint. Aquatint is a process used to etch tonal areas on the plate. The first step is to give the plate a porous ground by dusting it with rosin powder and fusing the powder to the plate by means of heat. When the plate is etched, the acid goes through the pores in the ground and bites tiny cavities in the metal. These cavities hold the ink. A variety of tones and textures can be created, depending on the density, width, and depth of the cavities.

The aquatint method was invented in the 18th century, and, although a great number of pure aquatint plates were done, the technique was mainly used with line etching.



"Minotauromachy," etching by Pablo Picasso, 1935. 49.5 cm × 69.7 cm.
By courtesy of the Museum of Modern Art, New York

Theoretically, there is no limit to the range of tones that can be etched with aquatint.

For the aquatint process, the plate is cleaned, as in hard-ground etching, and then dusted with rosin. Care in this step is crucial, as an incorrectly distributed rosin ground will produce uneven, spotty tones. To achieve even tones, a fine-grain rosin is used. The quantity should cover about 50 percent of the surface, neither too thin nor too thick. The dusting can be done either with a dust box or with dust bags.

Use of the dust box

The dust box is a completely enclosed container with a sliding tray (usually made of steel mesh) that holds the plate in position above the dust tray, which is filled with fine rosin dust. After the plate is placed in the box, the rosin dust is agitated either by a bellows, by an electric fan, or by shaking.

Dusting bags are made of various materials; the finer the material, the finer the dust coming through. The dusting bags have the advantage of allowing the artist to visually control the amount of dust deposited and also to use different textures in different areas.

After dusting, the plate is placed on the heating plate, and the rosin is fused to the metal. When the plate has cooled, the design is applied with a stop-out varnish. To achieve various tones the plate is bitten in stages, much as in hard-ground etching but with one important difference: aquatint is much more delicate, and the time element is more critical. A biting time of a few seconds can produce a fine gray, but a proportionately longer time is needed as the artist proceeds toward the darker tones.

Plastic sprays are also used to make aquatints. These lacquers and enamels are sold in pressurized spray cans and are sufficiently acid resistant to use for moderately long bites. They are easy to control and simpler to use, but they must be used in spray booths or other well-ventilated places.

Lift-ground etching (sugar-lift aquatint). In lift-ground etching, a positive image is etched on an aquatint plate by drawing with a water-soluble ground. In the conventional aquatint technique, the artist controls the image by stopping out negative areas with varnish, thus working around the positive image. But for lift-ground etching, he uses a viscous liquid (such as India ink, gamboge, or ordinary poster paint mixed with sugar syrup) to paint directly on the plate. After the painting is finished and dried, the whole surface is covered with thin, liquid hard ground. When dry, the plate is placed in lukewarm water that dissolves the painted design, lifting the ground and dislodging it from the places that had been painted, thus exposing the metal surface to be etched. Aquatinting can be handled two ways: either the whole plate can be aquatinted before painting with lift ground or it can be aquatinted after the design is lifted. Lift-ground etching is particularly well-suited to free, spontaneous, calligraphic designs.



"The Colossus," aquatint by Francisco de Goya (1746–1828). 29.3 cm × 21.6 cm.
By courtesy of the Metropolitan Museum of Art, New York. Harris Brisbane Dick Fund, 1935

Acid bite

Acids and the etching process. The acid bite of the plate is a critical stage in the making of an etching. The printmakermustbe familiar with the characteristics of the materials that are being used. On a zinc plate nitric acid is used. In the process of biting, this acid develops air bubbles over the bitten area. Under the bubbles the acid action is slower, and, therefore, if the bubbles are not constantly moved around by brushing, the etched line will be uneven. Nitric acid also has a tendency to underbite, that is, to bite not only straight down but also sideways. For this reason, areas of dense texture must be watched very closely.

Nitric acid also can be used on copper, but, except to bite out large areas, Dutch mordant is much better suited for this metal. The action of hydrochloric acid on copper is much more even and controlled than that of nitric acid. Thus, for a bold, rough bite, nitric acid on zinc is fine; but for delicate, controlled etching, Dutch mordant on copper is preferred.

Metal graphic. This method was originated by Rolf Nesch, the German-Norwegian printmaker. In all the intaglio methods previously discussed, the artist's design was created by making incisions in the plate. Nesch's method is the reverse of this process: the design is built up like a montage, by cutting out metal shapes and soldering them on the plate surface. Instead of the etching needle and the graver, the tools are shears, wire cutters, and a soldering iron. These plates are in deep relief and thus produce a heavily embossed print. Often such plates are combined with conventionally etched or engraved sections. In addition to metal shapes, wood and plastics may be used. Because of the extremely high relief, the printing of the plates requires specially prepared presses. A few contemporary artists work in such a high relief that the ordinary etching press cannot print their work and standard printing papers cannot be used. In some cases the high relief is created by compressing paper pulp into molds with hydraulic presses.

The use of embossing is not new. Some Japanese woodcuts have sections that have been decorated with "goufrage" (blind pressing). In contemporary printmaking, embossing has become a major interest, and many artists are exploring the possibilities of the intaglio print by using shallow paper bas-reliefs to exploit the interplay of shadow and light.

The etching press

Printing by intaglio processes. The most important piece of equipment in intaglio printing is the etching press, a simple machine whose basic principle has not changed for centuries. Motorization and the use of pressure gauges are the only major improvements. The press consists of a solid steel plate, called the bed, that is driven between two rollers; a screw mechanism on both sides of the top roller adjusts the pressure. Large modern presses are motor driven.

The print is made by placing the inked plate face up on the bed. Dampened paper is placed carefully on the plate and covered with several layers of pure wool printing felts. The bed is then driven through the rollers. The felts, which are squeezed between the metal rollers and the plate, push the paper into the crevices of the plate, forcing the paper into contact with the ink and thus transferring the image.

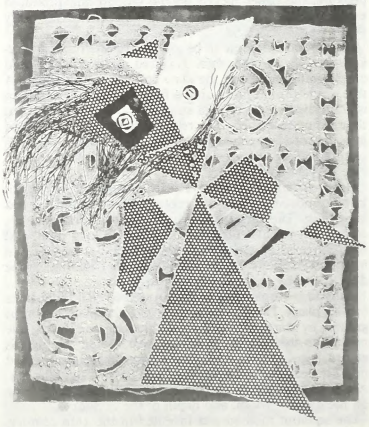
A fairly heavy pure rag paper is normally used. It is soaked until its fibres are softened and then, before printing, it is blotted until no surface water is visible. For inking, the plate is placed on a heater and kept warm throughout the inking and wiping steps. Heat makes the ink looser and thus facilitates both of these processes. Wiping is the operation in which the ink is removed from the surface of the plate, while leaving it in the recesses. Usually a carefully folded starched cheesecloth (tarlatan) is used. When a clean, crisp print is desired, the plate is given a final wiping with the palm of the hand.

Inks for intaglio printing are especially made for this purpose. The consistency of the ink must be such that it comes off the surface of the plate cleanly during the wiping operation, but at the same time it must have enough body to retain its relief on the paper. The printing ink must also have sufficient viscosity to stick to the damp printing paper to produce a clear and rich image.

After the print is pulled, it is dried, either between blotters or taped to a large, stiff board. This choice depends on the size of the print and the type of paper used.

Intaglio colour printing. The intaglio colour print is made with two or more intaglio plates successively overprinted on the same paper. Each plate represents one colour and its possible gradations. In principle, it is possible to take four plates—the three basic colours, yellow, red, and blue, plus black—and make a print that will have the full range of colours. If the colour areas are distinctly separated, more than one colour can be printed from one

By courtesy of the Philadelphia Museum of Art, gift of Lessing J. Rosenwald



"Sort Dame" ("Dark Lady"), "metal graphic" print by Rolf Nesch,1953. 57.1 cm × 50.2 cm.

plate. This method involves an extremely meticulous inking and wiping process.

Problems of colour registration

One of the greatest problems with intaglio colour printing is registering the successive colours in their precise location. If the colours can be printed immediately, wet on wet, then it is relatively simple, but often this is not possible. If the first plate has high relief and is overprinted while wet, the second plate will crush it completely. In this case the first print must be thoroughly dried and then rewetted for the second printing. Because the paper shrinks in the drying process, it is difficult to get it back to the original size when rewetted.

Several methods of registering can be used, depending on the particular problem. For wet-on-wet printing the process is simple. After both plates are inked, the first plate is placed on the press bed and its position is marked. Paper is placed over the plate and secured at one end with masking tape, or, if there is enough margin, the paper is run through so that one end remains caught under the printing roller. The print is then folded back and the first plate is replaced with the second.

Another method uses mats. The paper to be used in the edition is cut to the same size. A cardboard or metal mat is cut, corresponding to the size of the wet paper. The plate position is either cut out or marked on the mat. Registration consists of lining the paper up with the mat.

The most precise registering is with pinholes. Two pinholes are punched in opposite corners of the mat. Corresponding pinholes are punched through all the printing papers. In printing, the paper is picked up with two heavy needles through the punched holes. The needles are then inserted in the corresponding holes on the mat and the paper is released. The holes should be placed close to an edge that will be trimmed after the print is dry.

Stencilled colours with an intaglio plate. Stencilling is one of the simplest ways to use a number of colours combined with an intaglio plate. This method has advantages and also limitations. The main advantage is that it eliminates the registering problems of intaglio colour printing. On the other hand, it is limited to flat, sharply defined colour areas. One method does not replace the other, but each may be used to solve a particular problem.

The procedure itself is very simple. The intaglio plate is inked and wiped normally. The desired colour shape is cut out on a stencil paper. The stencil is placed on the already inked plate and the colour is rolled onto the surface of the plate using a gelatin or soft rubber roller. For surface rolling, regular artist oil colours can be used. The use of stencils allows a great number of colours to be printed with a single run on the press. This is done by surface rolling colours through stencils onto the intaglio inked and wiped plate surface.

For more complex colour combinations, it is possible to combine colours stencilled directly on the paper with colours offset from the intaglio plate. For more sophisticated stencilling, silk screen can be used also in combination with the intaglio plate. When intaglio and stencilling are combined, the process is often designated as mixed or combined technique. This is essentially the same procedure as conventional stencilling except that with silk screen more complex designs and textures can also be stencilled on the plate (see below *Stencil processes*).

Intaglio and surface colour with relief etching. In this technique the main colour structure is defined by the plate surface, which is etched to different levels. The linear or textural elements moving from one level to another bind the whole together.

Inking to different levels

The sequence of printing begins with the intaglio inking and wiping of the plate. Next, the first surface colour is rolled on with a soft gelatin roller that penetrates the lower levels of the relief. The high areas are inked with a hard rubber or composition roller. The sequence of rolling can change, according to the demands of the particular colour problem.

In addition to plate levels and roller variety, control of colour viscosity is an important factor. The thorough description of this method is so complex that the reader is referred to some of the technical books listed in the *Bibliography*.

SURFACE-PRINTING PROCESSES

Surface printing comprises those techniques in which the image is printed from the flat surface of the metal, stone, or other material. The major surface method is lithography, a planographic process. Although many experts place silk screen and stencilling in a separate category, they can be considered surface-printing processes. In lithography, the control of the design is achieved by the chemical treatments of the drawing surface. In stencilling, the design is created by holes in the stencil and the printing ink is either rolled or squeezed through the stencil onto the paper. Silk screen is a special form of stencilling.

Lithography. Lithography is based on the fact that water and grease do not mix. The image is drawn or painted on the stone or metal plate with greasy litho crayon or a greasy black ink (tusche). Once the drawing is finished, it is fixed with an etch to prevent the spreading of the grease.

The etch in lithography

A heavy, syrupy mixture of gum arabic and a small quantity of nitric acid, the etch is used to protect the drawing from water and to further desensitize the undrawn areas to printing ink. The nitric acid opens the pores of the stone, enabling the gum and the grease to enter easily. The gum arabic surrounds the greasy sections, forming an insoluble surface film that sticks to the negative areas and crevices of the grain. This coating around the image repels the water applied during printing and establishes a grease reservoir. It does not smear, and it prevents seepage that would blur the image.

Because of the antipathy of grease and water, the image attracts oily ink but repels water. Thus, when the stone is dampened with a sponge and an ink-charged roller is passed over it, the ink is deposited on the greasy drawing but not on the wet stone.

In lithography, the assumption is that the drawing made on the stone or plate will be closely duplicated on the print. While intaglio processes yield prints unlike any drawing technique, lithography is quite reproductive. Although it is a complex method, if lithography is well done, the effect of the print is deceptively simple and direct, making the technique attractive to artists who wish to avoid the more idiosyncratic printmaking methods.

A highly skilled technician is needed to produce a good lithograph, and most lithography is done in workshops

By courtesy of the National Gallery of Art, Washington, D.C., Rosenwald Collection



"La Mort: Mon ironie dépasse toutes les autres!" (*Tentation de St. Antoine*), lithograph by Odilon Redon (1840–1916). 26.2 cm × 19.7 cm.

where well-trained workers are available. The artist usually works on the stone or plate under the guidance of master printers. When the artist finishes a drawing, the master printers etch the stone and do the printing. In the basic technique, the first step is the preparation of the stone or plate. If a stone has been used before, its surface must be reground. The stone is placed in a sink and thoroughly wetted, and carborundum powder is sprinkled over it. Then, either with a levigator (a heavy steel disk with a handle) or by rubbing two stones together, the surface is thoroughly reground. From time to time the surface should be tested with a steel straightedge to make sure it is level; otherwise it will print unevenly. After the stone has dried, it is ready for work. It is very important to keep the stone clean because any dirt, particularly grease, will show up on the print. Smudges and dirt can be cleaned off with erasers and abrasives.

Metal plates (zinc or aluminum) can also be used, and these, too, may be reground. Although metal plates are satisfactory, stone is far superior, particularly for producing subtle tones and details.

Use of crayons and tusches on stone

With litho crayons and tusches the artist can work on the stone as he would on paper. A whole arsenal of effects is available, including pen, pencil, splashing, sprinkling, spraying, texture transfers, and scraping. After the drawing is finished and before etching, the image must be protected from the etching solution by rubbing rosin and then talcum powder on the stone. The acid-resistant rosin protects the drawing; the talcum absorbs the excess grease, allowing the adhesion of the gum etch to the edges of the drawing.

Next, the whole surface of the stone is coated with undiluted gum arabic, applied with a wide, soft brush. The subsequent etching process is done in stages. The weakest acid solution is usually brushed first on the lightest areas of the drawing. After an appropriate interval, the next strength solution is brushed on, and this continues until the strongest etch has coated the darkest areas.

After the allotted time has elapsed, the excess etch solution is blotted with newsprint paper. The surface is then wiped down and buffed with cheesecloth to a smooth, even layer. When properly handled, the stone should appear dry. It should be allowed to stand for two hours before washing out, the next step.

The washout is done by pouring a small amount of turpentine or Lithotine over the drawn areas. Gently rubbing the drawn areas with a clean dry rag removes the drawing through the gum-etch coating. The image is preserved by the absorbed grease in the porous limestone.

Next, the stone is rubbed with liquid asphaltum or printing ink dissolved in turpentine. This procedure saturates the image and protects it at the same time.

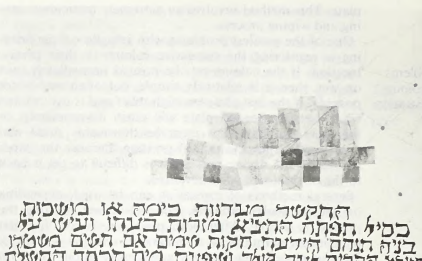
After the stone is dry, it is ready to be inked (rolled up). First, it is dampened with a wet sponge. (In between the rollings, the stone should be redampened.) Ink rolling should be carried out according to a set pattern, gradually building up the image. To facilitate the even distribution of ink it is important to use a roller wider than the image.

The lithographic press prints with scraping pressure. The press itself consists of a metal frame that accommodates a travelling steel plate (the bed), which passes with the stone under a scraping bar (or yoke). The bed can be lowered (to position the stone) and raised (to print). The pressure on the scraping bar can be adjusted.

Lithographs can be printed on either dry or damp paper. The advantage of dampening is that it is possible to use less ink and less pressure, thus minimizing the risk of clogging the image.

To print, the printing paper is first placed on the stone, followed by a newsprint paper, and then a blotter. Last comes the tympan, a sheet of smooth, tough material that can withstand great pressure without stretching. After the bed is raised to printing position, grease is spread evenly in front of the scraping bar on the tympan to allow it to slide easily. Then the print is made.

The prints of the French artist Henri de Toulouse-Lautrec demonstrate that lithography offers endless possibilities in colour printing. Because the effect of lithography is much more painterly than either woodcut or intaglio printing, it is natural that the strong preoccupation with pure colour in contemporary art has created a revival of interest in this medium. The planning and the principle of colour separation are similar to those for the colour woodcut or intaglio colour.



"Pleiades," silk-screen print by Ben Shahn, 1960.
52 cm × 67.3 cm.
By courtesy of the Kennedy Galleries, Inc., New York

Stencil processes. In stencilling, one of the simplest methods of duplication, the design is cut out of paper (or any other suitable thin, strong material) and is then printed by rubbing, rolling, or spraying paint through the cutout areas.

The silk-screen process

Silk screen is a sophisticated stencil process, developed about 1900 and first used mainly for advertising and display work. About 1950, fine artists started to use the process extensively, giving it the name serigraphy.

The silk-screen process got its name from the fine mesh silk that, when tacked to a wooden frame, serves as a support for a cut paper stencil. The stencil is glued to the silk. In the basic process, the open mesh of the silk lets the paint through, while the paper stencil blocks it out. A design can also be blocked out on the screen with glue or other suitable substance.

A common method of stencil preparation is to cut the stencil with a knife. In this method the artist can use commercially produced screen process printing plates or conventional stencil papers. For fine, accurate work, process plates, which consist of a film on a backing, are preferred. Areas to be printed are cut out of the film and peeled off, leaving the rest of the film on the backing paper. After the plate is attached to the screen, the backing paper is removed; thus, the screen is covered with film except in the printing areas. Process plates are available in different colours to make registering easier, and they are attached to the screen either by heat or by the use of a special solvent.

Another method that is quite common is the so-called tusche-and-glue method, which is similar to lift-ground aquatint etching. The design is painted on the screen with tusche and, when dry, the whole screen is covered with glue. When the glue dries, the design is washed out with either kerosene or turpentine. The tusche comes in liquid form for brushing or in solid crayon form. The use of the crayon results in screen prints that deceptively resemble lithographic prints.

Stencil plates can also be made photographically. These plates are made by placing a photographic positive on a photosensitized gelatin stencil plate in a vacuum printing frame. Exposure to light hardens the gelatin under the transparent areas but leaves the gelatin soft under the dark areas. In warm water the soft areas wash out. The stencil is attached to the screen in the same manner as other stencils.

Silk-screen prints

To make a silk-screen print, the wooden frame holding the screen is hinged to a slightly larger wood board. The

printing paper is placed on the board, under the screen. The consistency of the ink is important: it must be liquid enough to pass through the screen but not so liquid that it runs. The ink is pressed through the screen with the squeegee (a rubber blade, usually the same width as the screen, set in a wooden handle). Any number of colours can be used, a separate screen for each colour.

SPECIAL TECHNIQUES

Monoprint (monotype). A monoprint is a unique print. The artist paints on a surface such as metal, plastic, or glass and then transfers the wet design to paper, either by rubbing or with an etching press. The primary reason for making a monoprint is that, when the image is offset from the plate to the paper, the print achieves a separate quality and luminosity totally unlike a painting made directly on paper. In the 19th century, Edgar Degas did considerable experimentation with monoprints and produced a great number of superb ones. He often worked over the proofs with paint or pastel. There has been a strong revival of interest in this method.

Cliché-verre. The method of printing known as *cliché-verre* was used by a few artists in the 19th century during the period when photography was a new and exciting invention. The *cliché-verre* method follows the principle of photography but does not have its tonal variations. The print was made by covering a piece of clear glass with an opaque pigment or emulsion; the design was then scratched through with a sharp etching needle or stylus. When the drawing was finished, the glass plate (negative) was placed on a photosensitized paper, exposed to light, and then developed. The result was a (positive) print with strong black-and-white contrasts. Some of the best *cliché-verre* prints were made by the French landscape painter Camille Corot.

The cellocut. The cellocut method was named by its originator, U.S. printmaker Boris Margo, one of the first to experiment extensively with plastics.

In this method, liquid plastic that has been dissolved in acetone is poured onto a rigid support backing, such as fibreboard or cardboard. The solidified plastic can be textured, raised into relief, and worked with various tools. It can be engraved, scratched, sanded, and filed. The resulting plastic plate can be printed either as a relief or as an intaglio plate, or even both. It can be printed alone or in combination with other techniques. Thin layers of plastics can easily be placed on top of intaglio plates and printed together.

By courtesy of the Museum of Modern Art, New York, gift of Mrs. Sam A. Lewisohn



"Clown with Monkey," monotype by Georges Rouault, 1910. In the Museum of Modern Art, New York City. 57.4 cm × 38.7 cm.



"The Alchemist, No. 2," cellocut by Boris Margo, *c.* 1947. 85.3 cm × 58.3 cm.

By courtesy of the Brooklyn Museum, New York

Collagraphy. Like the metal graphic process, collagraphy is an additive method; the printing surface is built up. It is essentially an intaglio method, but it can be combined with relief printing. The printing surface is created by gluing various materials and textures to a support. Today, with the variety of new material available, the possibilities are limitless.

The support (plate) for collagraphy must be thin and strong. A porous material, such as cardboard, must be treated with a sealer. To build up a tough, durable printing surface, a strong adhesive such as polyvinyl acetate must be used.

Among the materials that can be used for tonal areas are sawdust, sand, carborundum, sandpaper, and ground walnut shells. For specific textures, materials such as tarlatan, laces, and crushed paper can be glued into the adhesive.

After the plate has been constructed, the surface is sealed. The sealer can be either brushed or sprayed on. Plastics are preferred because they are tough and are not dissolved by the solvents generally used to clean the plate.

The printing of collagraphs is essentially the same as for intaglio printing.

Making a plaster print

Plaster print. Good proofs of an intaglio plate can be made by plaster casting, for fine plaster of paris will pick up the most delicate details. This method will produce a particularly attractive proof if the plate has deeply etched or engraved sections.

To make a plaster print, the plate is inked in the same manner as it would be for normal printing. The inked and wiped plate is placed face up on a glass plate, and a precut wood frame is placed around the plate to contain the plaster. After the plaster is poured, it is allowed to cool and set, after which the plate is gently removed.

PROCESS PRINTS

Process-printing methods are primarily used for commercial reproduction. Today, however, many artists use commercial methods to produce fine art. Silk-screen printing itself began as a commercial process, and today it is one of the most popular techniques in printmaking because its character is well suited for hard-edge geometric images. Photomechanical processes are incorporated in the work of many contemporary printmakers.

Linecut. The linecut technique is the simplest and least expensive of all the photoreproductive processes. As it cannot register tone, it is used mostly to reproduce black-and-white line drawings. If tones are needed in a linecut, they are achieved with the use of screens consisting of dots (Ben Day screens). The linecut is similar to the woodcut in that both are used in relief printing.

Linecuts are usually made on zinc plates coated with an emulsion of albumin or gelatin mixed with potassium bichromate. This emulsion hardens on exposure to light. The light passing through the transparent part of the negative hardens the emulsion. The areas of the emulsion that are protected by the black on the negative remain in their soluble state. The plate is then rolled with greasy ink and soaked in water. The unexposed soft emulsion is washed out by the water. The plate is then dried and dusted with powdered rosin, which adheres to the remaining inked emulsion areas. Heating causes the rosin to melt, forming an acid-resistant coating. The plate can then be etched so that the design stands up in high relief.

Halftone cut or plate. Halftone is more sophisticated than linecut, since it is capable of reproducing fine tonal variations. The subject is photographed first through a glass plate that has fine lines printed on it at right angles. The result is an image broken up into tiny dots corresponding to the openings in the screen. When printed, these dots create the optical illusion of continuous tones. There are great variations in screens from coarse (50 lines per inch) to very fine (175 lines per inch). The selection of the screen is dictated by the paper to be used for printing.

After the photonegative of the image is finished, it is printed on a sensitized copper plate. For halftone work, copper is used because of its ability to record fine details. The procedure of washout and etch is similar to that used with linecuts.

Function of the gravure screen

Rotogravure. To make a gravure plate, a screen is used that is the reverse of the halftone screen, in that the lines are transparent and the areas between the lines are opaque squares. When the sensitized plate is exposed to light through the screen, the emulsion on the plate hardens under the lines but leaves the squares soft. Then the plate is exposed again through a diapositive (a positive transparency) of the subject. This time the soft emulsion squares harden in proportion to the range of grays. In etching, the softest squares are affected by the acid first and the hardest ones last. The result, after the etch, is a plate covered with squares of equal size but varying depth. As the deep squares hold more ink than the shallow ones, the tones in the reproduction are controlled in the same manner as in all intaglio printing methods. The rotogravure plate is inked by an ink-carrying cylinder and wiped by a steel blade that removes all the excess ink from its surface.

Although rotogravure is an intaglio printing process, it is printed on dry paper with light pressure and thin ink. Hence, there is hardly any embossing.

Offset lithography. Offset lithography is the application of lithography to commercial mass production. The plate, instead of being a stone, is of specially treated zinc or aluminum, suitable for mounting on a cylinder. The image is photographed on the sensitized lithoplate through a screen. The offset method involves double printing. The image from the plate is printed on another roller covered with a rubber blanket, and from this roller it is transferred to the paper. Because the image is reversed twice, the final print corresponds to the original plate. Since the litho offset ink is thin, to speed up inking and facilitate transfer, the tonal areas lose some of their richness and tend to print gray. Litho offset is often used for colour printing. The colour separation is made photographically.

CONTEMPORARY EXPERIMENTATION

One of the most crucial changes in the 20th century involved the size of the print. All through its history, with few exceptions, the print was considered an intimate art form, enjoyed by the few. The change started with the Lautrec posters: the print started to grow until it became mural size. As the dimensions of the print changed, so did its character. It became increasingly bolder and more colourful. Today, the print often competes with painting, a situation deplored by many people who feel that in the process the print is losing its particular character and beauty. For a time, major print shows tended to exhibit only a limited number of small, delicate prints, but two more recent developments seem to be balancing that trend. One is the reappearance of the intimate, introspective, black-and-white print. The other is the revival of the long-neglected woodcut, due particularly to the interest of the Postmodernist artists in German Expressionism.

Next to the size of the print, the greatest change has been in the technology of colour printing. In this area, techniques have become so varied that practically any effect is possible. This development has contributed to the vitality of printmaking, because it has encouraged the participation of colour-oriented artists. The combining of various media is closely related to the experimentation in colour printing. Each medium has its own capabilities and limitations; combined, the media often complement each other. It is now common to see three or four different techniques combined.

Combining media

Another area of experimentation is in three-dimensional surfaces. The trend started with embossing, and today artists are creating completely three-dimensional printed objects.

The shaped print is a printed paper sculpture, made by cutting and folding the printed paper or by assembling precut printed surfaces. Many of these surfaces are metal or plastic objects with printing, usually done by the silk-screen process.

Instead of using rectangular painting surfaces, many painters now work on shaped canvases. In the same way, printmakers, instead of using rectangular plates, are using many different shapes. Printing with movable plates, which became particularly popular in intaglio colour printing and with colour woodcuts, is the logical extension of this freedom. In this method small cutout plates are placed on top of larger plates and printed together, or they are assembled on a cardboard support and printed. This procedure facilitates the use of many colours and also offers great freedom in composing.

Photography is profoundly affecting printmaking. Photographic methods can be combined with intaglio, lithography, or silk screen to enrich their vocabulary. The possibilities are nearly limitless. Yet photography can be corrupting when it reintroduces reproductive ideas, and, unfortunately, it is often used for this effect.

Kinetic (moving) art, such as the mobile, is a major contemporary preoccupation in painting and sculpture. At present there are few attempts in this direction in printmaking, but there will probably be more. The problem in such works is how to combine the print with motion without destroying its very nature.

MOUNTING AND CARE OF PRINTS

Very few people know how to display prints and how to take care of them properly. It is heartbreaking to see a great master's print glued to a cheap cardboard or the border of a fine print ruined with tape.

Because paper, particularly old paper, is fragile, it should be handled as little as possible and never picked up with one hand since this might put too much stress on the paper and tear it. To protect it, a print should be mounted as soon as possible.

The surface of the print, especially an intaglio print, is delicate, and rubbing might permanently injure it. Prints should not be stacked without protective layers of tissue paper between them. Wood-pulp papers should not be used, as the acid content in these can burn the print. A print should not be exposed to intense sunlight; this is true particularly of colour prints, for very few colours are



Use of photography shown in "Elvis Presley," silk-screen print on canvas by Andy Warhol, 1964. 2.08 m × 1.2 m.
By courtesy of the Leo Castelli Gallery, New York

stable enough to withstand long exposure to direct sun. Light can also affect the paper. Because wood-pulp board contains chemicals that in time can burn or discolour the paper, a permanent mat should be constructed out of pure rag board. A properly constructed mat consists of two parts: the backing board to support the print and the covering frame to display it. The width of the mat frame should be related to the print's dimensions so that the mat does not overpower the image. The window size of the mat should never obscure the printed image itself, or the signature and edition number.

Because temperature changes in a damp climate can cause condensation and the print can develop fungi, prints should be kept from direct contact with the glass. The simplest protection is a deep enough pure rag mat. If this is not sufficient, a filler should be inserted into the frame to increase the space between the mat and the glass.

The backboard of the frame should be also rag board, or at least faced with rag paper, although the latter is not the perfect solution. The back of the frame should be sealed with tape to prevent the penetrating of dust. In damp climates it is advisable to keep the frame away from the wall by placing corks on its four corners. This facilitates the free circulation of air. Air-conditioning and humidity controls are the best protection.

History of printmaking

Engraving is one of the oldest art forms. Engraved designs have been found on prehistoric bones, stones, and cave walls. The technique of duplicating images goes back several thousand years to the Sumerians (*c.* 3000 BC), who engraved designs and cuneiform inscriptions on cylinder seals (usually made of stone), which, when rolled over soft clay tablets, left relief impressions. They conceived not only the idea of multiplication but also the mechanical principle, the roller, which in more sophisticated form became the printing press.

On the basis of stone designs and seals found in China, there is speculation that the Chinese may have produced a primitive form of print—the rubbing—about the 2nd century AD. The first authenticated prints rubbed from wood blocks were Buddhist charms printed in Japan and distributed between AD 764 and 770. It is believed that the first wood-block prints on textiles were made by the Egyptians in the 6th or 7th century; but the earliest printed image with an authenticated date is a scroll of the *Diamond Sūtra* (one of the discourses of the Buddha) printed by Wang Chieh in AD 868, which was found in a cave in eastern Turkestan.

In Europe, stamping (to imprint royal seals and signatures) preceded printing by rubbing or with a press. The earliest documented impressed royal signature is that of Henry VI of England, dated 1436.

Textile printing, however, was known in Europe in the 6th century, the designs consisting largely of repeated decorative patterns. Printing on paper developed from textile printing, following the introduction of paper from the Orient. The first European paper was made in 1151, at Xativa (modern Játiva), Spain. Soon afterward paper manufacturing began in France and then in Germany and Italy, notably by Fabriano, whose enterprise was established in 1276.

Woodcuts of playing cards

The first woodcuts on paper, printed in quantity, were playing cards. The term *Kartenmahler* or *Kartenmacher* ("painter or maker of playing cards," respectively) appears on a German document dated 1402; and documents from both Italy and France from the middle of the 15th century mention wood blocks for the printing of playing cards. The earliest dated woodcut is a "Madonna with Four Virgin Saints in a Garden" from the year 1418.

Many documents from the 15th century indicate that a clear distinction was made between the designer and the cutter of the wood blocks. From the outset, woodcut was primarily a facsimile process; the cutter copied a drawing provided by the designer.

Printing from a metal engraving, introduced a few decades after the woodcut, had an independent development. The art of engraving and etching originated with goldsmiths and armour makers—men who were thoroughly professional craftsmen, practicing an art that had a long, respected tradition. Since the armour makers and goldsmiths were designers themselves, the whole process was controlled by the creative artist.

PRINTMAKING IN THE 15TH CENTURY

Germany. Single prints (in contrast to those printed in a series or as part of an illustrated book) of the early 15th century were not signed or dated, and, because they were religious images carried by pilgrims from one place to another, it is nearly impossible to establish with certainty their place of origin. Their style alone must be relied upon for some indication of origin.

The first phase of woodcut, from about 1402 until about 1425, was dominated by boldly designed single figures against a blank background. Most of the cuts were made to be hand coloured. In the second half of the 15th century the cuts became more complex: architectural and landscape elements came into use, and often the image was framed in an elaborate border.

The first metal prints (criblé, or dotted, print) were made in the second half of the 15th century. The design was created by tiny dots punched into the metal and intermingled with short cuts. Surface printed, the whites are the positive part of the design, which is dominated by the dark background. Tiny holes in the borders indicate that most of these plates were intended as decorations to be mounted rather than as printing plates.

The earliest dated intaglio-printed engraving is from 1446: "The Flagellation," of a Passion series. Around this time, the first distinct personality to have great influence on German engraving appeared. He is known as the Mas-



"Blumen-Dame" ("Cyclamen Queen"), engraving by the Master of the Playing Cards, *c.* 1440. 13.9 cm × 9.2 cm.
Deutsche Fotothek, Dresden

ter of the Playing Cards. His style was simple, nearly monumental; unlike the printwork of goldsmiths, his engravings lack ornamentation. For shading he used slightly diagonal parallel cuts. The Master of the Playing Cards heralds the beginning of a century of great printmakers in Germany. Another significant engraver, the Master of the Banderoles, was named after the ribbon scrolls characteristic of his prints, which are more decorative than those of the Master of the Playing Cards.

Engravings of Master E.S.

In the second half of the 15th century, the outstanding printmaker was Master E.S., who flourished about 1440–67 and was one of the first to use initials as a signature on his plates. Little is known about him, but the personality that emerges from approximately 317 plates is forceful and distinct. Although it is evident from his prints that, like most early engravers, he was first trained as a goldsmith, his work has strong pictorial quality.

Martin Schongauer was the first great engraver who is known to have been a painter rather than a goldsmith. Although Schongauer's style was still Gothic in character, he composed with much greater freedom than his contemporaries, thus representing a transition into the Renaissance. He made about 115 plates, mostly of religious subjects, and was a powerful influence on the young Albrecht Dürer (see below *Printmaking in the 16th century*). During the second half of the 15th century, a group of brilliant engravers known only by their initials emerged in Germany. They are the Masters B.G., B.M., L.G.S., A.G., B.R., and W.H. The controversial figure of Israhel van Meckenem appeared at the end of the 15th century. A superb and extremely prolific engraver, he was a rather eclectic artist, borrowing from other masters and often copying them.

Italy. In the 15th century, Italian printmaking was dominated by the northern cities: Florence, Venice, and Milan. Throughout the century, printmaking was mainly concerned with playing cards and book illustrations, with a few single prints appearing in the second half of the century. While in Germany and the Netherlands the art was completely dominated by devotional, religious subject matter, Italian printmaking covered a relatively broad range. The awakening Renaissance attitude made the artists much more receptive to purely aesthetic, decorative, sensuous experience. In addition to religious subject matter, Italian prints included mythology, pure ornamentation, and some of the finest early portrait engravings.

Giorgio Vasari, the chronicler of the Renaissance, credited the Florentine goldsmith Maso Finiguerra with the invention of printed engraving, but present knowledge indicates that, at the same period in Germany and the Netherlands, printmaking was in a more advanced stage. In spite of the fact that book printing was originally introduced from the northern countries into Italy, engraving remained a national, regional development, free of strong foreign influence until the beginning of the 16th century.

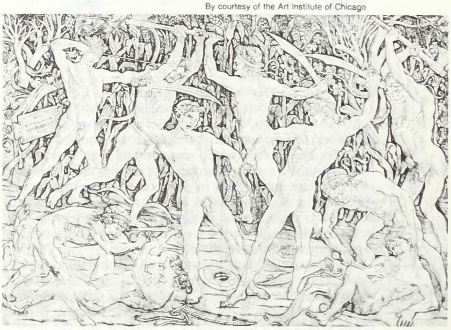
Broad manner and fine manner engraving

Two methods of engraving were practiced in Italy, the broad manner and the fine manner. The fine manner, associated with the Finiguerra school, is characterized by closely cut and extremely fine lines combined with cross-hatching intermingled at times with dots. The broad manner is less dense, and forms are modelled using diagonally cut parallel lines, interlaid at times with short cuts or dots. In shading, the spacing between the lines is wider than in the fine manner and there is no cross-hatching.

Finiguerra himself was not an important artist. His significance lies in his influence on Antonio Pollaiuolo, a Florentine painter, sculptor, and architect whose reputation as one of the most distinguished engravers of the 15th century is based on his one authenticated print, "The Battle of the Nudes" (*c.* 1465)—a powerful image, beautifully engraved in the broad manner.

While Pollaiuolo worked in Florence, Andrea Mantegna, a great painter and certainly the most eminent Italian printmaker, lived and worked in Mantua. Mantegna produced approximately 20 plates (only seven of which are completely authenticated), all line engravings in the broad manner. A superb draftsman and a virtuoso engraver, Mantegna could achieve, in spite of the limitations of his method, an incredible range of colour in his prints, a quality lacking in the work of most of his followers.

By courtesy of the Art Institute of Chicago



"The Battle of the Nudes," engraving by Antonio Pollaiuolo (*c.* 1431/32–98). About 40 cm × 61 cm.

In addition to the masters, talented engravers included Cristofano Robetta, a Florentine who made some rich, intricate engravings in the fine manner; and the Venetian Jacopo de' Barbari, who travelled in Germany and whose refined engravings show the influence of Albrecht Dürer.

Other countries. *The Netherlands and Burgundy.* The first half of the 15th century in the Netherlands and Burgundy was dominated by woodcut book illustrations. Although no single prints of great importance were produced, beautiful books were published. Antwerp and Delft were the main printing centres.

Parallel with, if not even a little earlier than, the emergence of distinguished printmakers in mid-15th-century Germany, a group of great engravers emerged in the Netherlands and neighbouring Burgundy. Superb artists,

they are identified only by the subject of their most characteristic work: the Master of the Death of Mary, Master of the Gardens of Love, and Master of the Mount of Calvary.

Toward the end of the century, the Netherlands produced a brilliant artist—rivalling Master E.S.—known as the Master of the Amsterdam Cabinet or Master of the Hausbuch, who worked between 1450 and 1470. He is known by this name because the finest collection of his extremely rare prints is in the Print Cabinet of the Rijksmuseum. His prints are painterly and almost expressionist in power. His role in the technical development of printmaking is also significant, as he was the first major artist to make drypoints.

Dominance of book illustrations in France

France. In France, book illustrations dominated printmaking throughout the century. Paris, the cultural centre, led in book publishing, although other prosperous cities, such as Lyon, produced many illustrated books. The publications printed by Dupré and Pierre Le Rouge are the glories of French medieval graphic art.

Strangely enough, there was little engraving of importance. Most of the French engravings of this period were either rather crude, provincial illustrations or playing cards.

PRINTMAKING IN THE 16TH CENTURY

Germany. Albrecht Dürer was the master of 16th-century German graphic arts. One of the towering figures in the history of printmaking, he was a complex, truly Renaissance man, interested in philosophy and science as well as art. He was one of the first to break the provincial isolation of Germany by travelling to Italy, where he learned from the Italians and in turn influenced them.

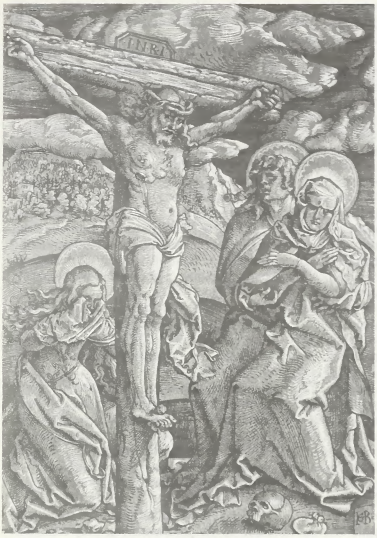
Dürer's subject matter mirrors his thoroughly European intellectual orientation. His prints deal with religion, history, mythology, and folklore. He is also one of the first great portrait engravers.

Dürer was one of the supreme draftsmen of all time and an artist of enormous imagination and sensibility. As a technician he raised the art of engraving to a height it never reached again. As an experimenter, he produced, in addition to his engravings and woodcuts, etchings and drypoints. His best works are metal engravings, which he cut himself. His woodcuts are perfect reproductions of his superb drawings.

By courtesy of the Metropolitan Museum of Art, New York, Harris Brisbane Dick Fund, 1945



"Melencolia I," engraving by Albrecht Dürer (1471–1528). 24.1 cm × 20 cm.



"The Crucifixion," chiaroscuro woodcut by Hans Baldung-Grien (1484–1545). 36.8 cm × 25.8 cm.

By courtesy of the National Gallery of Art, Washington, D.C., Rosenwald Collection

Hans Baldung-Grien, another great German printmaker and one of the most original artists of his time, worked with Dürer. In his images of witchcraft and magic, Baldung expressed the medieval mysticism that lingered in the German Renaissance. Besides his black and white work, he produced fine chiaroscuro woodcuts in which light and shadow are produced by using different wood blocks for different tones of the same colour.

Lucas Cranach the Elder was a typical representative of the German Renaissance—much less affected by the Italian Renaissance influence than Dürer. He had a vivid imagination and an earthy imagery that was full of vitality. He made many woodcuts and relatively few metal engravings.

Landscape etchings of Altdorfer

Albrecht Altdorfer was one of the first great landscape artists and one of the first to make landscape etching and woodcuts for their own sake, rather than as backgrounds for figures. Under his influence, two other artists made fine landscape prints: Augustin Hirschvogel and Hans Sebald Lautensack. The spontaneity and directness of their work foreshadows the lyrical landscapes of the 18th century.

The 16th century also included the artists who were referred to as Little Masters. They were so called because of the size of their prints, not their stature as artists. They were Barthel Beham, Hans Beham, Georg Pencz, and Heinrich Aldegrever.

Other countries. *The Netherlands and Flanders.* The outstanding Dutch printmaker of the period was Lucas van Leyden (1489/94–1533). If the latter birth date is correct, at the age of 14 he was already an accomplished engraver. In maturity, he was a superb engraver, in many respects rivalling Dürer. Besides his metal engravings, which are characterized by a very delicate touch, van Leyden designed many wood blocks and also made a few etchings.

A virtuoso of the burin, the Flemish engraver Hendrik Goltzius (1558–1617) developed an incredible variety of

cuts and textures to imitate the surface qualities of materials. Other printmakers of the period include Allaert Claesz and Cornelis Matsys.

Italy. After the death of Mantegna in 1506, Italian printmaking of the 16th century was dominated by lesser figures. During the 16th century the few etchings produced in Italy have only historical interest.

The most influential engraver of the century was Marcantonio Raimondi. Under the influence of Dürer, Raimondi became a virtuoso engraver; technically Dürer's equal, he lacked his master's originality. Raimondi eventually became the engraver of Raphael, organizing a workshop that was dedicated primarily to making reproductions of the master's work. Thus, Raimondi won the dubious honour of being the first of the many printmakers who ultimately were influential in turning the art of engraving into mere reproduction. He was followed by a whole generation of competent engravers who were devoted solely to reproduction.

One of the exceptions was Giorgio Ghisi of Mantua, who in his isolated regional development escaped the corrupting influence of Rome. His 1550 visit to Antwerp made Ghisi an important link between Italian and northern engraving.

Engravings of Duvet

France. The only major figure in 16th-century French engraving is Jean Duvet, whose predilection for excessive ornamentation indicates that he was trained as a goldsmith. Although Duvet's style was influenced by Mantegna, his imagery was completely original. His greatest work, "The Apocalypse," reveals a feverish, mystical imagination.

Apart from the work of Duvet, ornamental engraving was the most significant achievement of 16th-century French printmaking. Although these elegant engravings cannot be ranked with the work of the great masters, they represent a genuine expression of the French spirit. The outstanding figure of this school was Étienne Delaune. Although his motifs were influenced by those employed by Raphael for his fresco wall paintings in the Vatican, Delaune nonetheless achieved a personal style.

Trends in the late 16th century. By the second half of the 16th century, the quality of printmaking, particularly engraving, had gone into a severe decline. Masters like Dürer and Mantegna were replaced by skilled craftsmen. The trend toward reproduction that had begun with Raimondi gained ground, sapping the vitality of engraving. Yet at the same time, the quantity of production increased. Except for the modern era, this was probably the most prolific period of printmaking. Since it was the beginning of the age of travel, discovery, and religious upheaval, the demand for maps, religious pictures, illustrations, and portraits was enormous.

By courtesy of the Metropolitan Museum of Art, New York, Joseph Pulitzer Bequest, 1917



"Adam and Eve," engraving by Marcantonio Raimondi (*c.* 1480–*c.* 1534). 23.8 cm × 17.6 cm.

One after the other, print publishing houses opened all over Europe. Dutch and Flemish families dominated the new profession: in the Netherlands, the firms Cock, Galle, and Passe; in Augsburg, Dominicus Custos; in Antwerp, Brussels, Prague, and Venice, the Sadeler family. In Italy, Antonio Salamanca cornered the market and flooded it with bad reprints of Raimondi engravings.

The publishers of this period usually bought the original plates outright from the artist and issued prints on demand in unlimited quantities. If the plates wore out, they were reworked in the publishers' own workshops, a practice that was responsible for the destruction of many fine artworks. In many cases, it is no longer possible to identify the creator of the original plate.

Map engravers

In this period, map engravers were particularly important: the maps of Abraham Ortelius were engraved by Franz Hogenberg; Gerardus Mercator engraved his own map designs; and Jodocus Hondius bought the Mercator plates after their use in the edition of 1596 and introduced them in England, along with some of his own work.

PRINTMAKING IN THE 17TH CENTURY

Portrait engraving. *France.* The end of the 16th century and the beginning of the 17th were dominated by ornamental engravers and illustrators, who were working under Flemish influence; by the middle of the 17th century, however, a distinctly French school of portrait engraving had emerged. Although this school did not produce a major master, it represents a significant phase of European printmaking.

Michael Lesne, a French portraitist whose influence was considerable, worked for a time in the Rubens workshop, later returning to France. Claude Mellan, another major influence, was trained in Rome. Technical virtuosity dominated his prints; for example, the modelling of a face with one continuous spiral.

A superb engraver and a fine draftsman, Robert Nanteuil is considered the undisputed master of French portrait engraving. His style is simple, elegant, and free of the mannerism characteristic of his contemporaries. He and his two rivals—Gerard Edelinck, who was born in Antwerp but studied and developed his style in France, and Antoine Masson, who engraved portraits in the grand style—represented the dominant forces in 17th-century French portrait engraving.

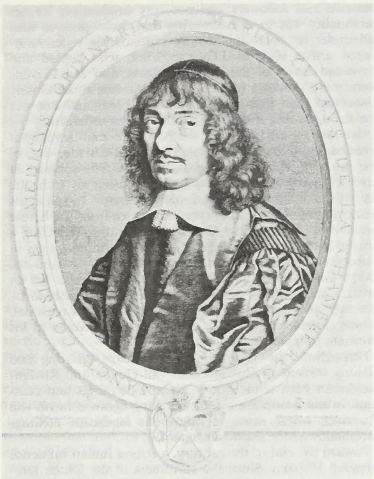
Germany. After the glories of the 15th and 16th centuries, German graphic genius was dormant for nearly three centuries. Historically, Ludwig von Siegen, a minor painter and medalist, is important for his invention of the mezzotint printing method. But the perfecting of this tonal technique increased the reproductive facility of printmaking, thus contributing to the decline of artistic creativity.

Like France, Germany produced a school of portrait engravers; but, although competent technicians, they failed to develop a distinctly national school comparable to the French. Of this group, two are significant: Jeremias Falck, a student of Hondius, and Bartholomäus Kilian, who studied in Paris and later introduced French influence into German printmaking.

The Netherlands. Portrait engraving in Holland was on a higher level than in Germany. Cornelis van Dalen was a fine engraver who emigrated to England and died there. More gifted than his father, Cornelis van Dalen II was an artist of considerable stature, who engraved some of the most powerful portraits of his time.

The development of mezzotint

Abraham Blooteling, a pupil of van Dalen II, was also a fine portrait engraver. His major contribution, however, was in the development of the new technique of mezzotint—specifically, the invention of the rocker, the tool used in the technique. He also introduced the mezzotint into England, where it was adopted with such success that it later became known as the "English Manner."



"Marin Cureau de la Chambre," engraving by Robert Nanteuil (1623–78). 25 cm × 19.4 cm.
By courtesy of the National Gallery of Art, Washington, D.C., Rosenwald Collection

England. In the 17th century, English printmaking produced a portrait engraver of considerable stature, William Faithorne. He studied in France and initially was under the influence of Mellan and Nanteuil; in his late work, however, he developed a style independent of theirs. Faithorne was England's only major native printmaker during this period, when most prints were reproductive engravings. By the end of the century, engraving was in total decline, replaced by the fashionable mezzotint.

Flemish printmaking. One of the dominant figures of European art was Peter Paul Rubens, who was a painter, diplomat, and businessman. Quickly recognizing the commercial potential of printmaking, Rubens organized a graphic workshop where, under his supervision, reproductions of his work were produced. Only one etching, "St. Catherine," is considered as his own. The quality of this one print indicates how great was the loss to the art of printmaking that this great draftsman did not make more original etchings.

Rubens' pupil Anthony Van Dyck was one of the most distinguished portrait painters of his time. At age 27 he undertook a very ambitious project: the etched portraits of the 100 most famous men of his day. For this set of prints, known as the "Iconography," he completed 18 portraits. But only five of these ("Peter Brueghel the Younger," "Snellinx," "Erasmus," "Suttermans," and "Josse de Momper") remained unchanged; another five were retouched by professional engravers, and the rest were completely reworked by them.

European etching. Like the Van Dyck portraits, nearly all of the outstanding prints produced in the 17th century were etchings. Etching emerged as the dominant technique for many reasons. The fact that engraving had become a completely commercialized, reproductive method and that mezzotint had never been anything else alienated many artists. As an unexploited and relatively unexplored medium, etching intrigued the experimentally oriented. Furthermore, the fluid, flexible technique of etching was a lure for the creative painter, whose own medium had become freer and more spontaneous.

Italy. At the beginning of the 17th century, there was more etching in Italy than in any other European country. Strangely enough, probably the three most important etchers—Jacques Callot, Claude Lorrain, and José de Ribera—were foreign-born.

The Bolognese school was formed around Guido Reni, whose delicate etching style of light lines and dots became a standard technique for most Italian etchers of his time. His school, however, did not produce any superior printmakers.

The Spanish painter José de Ribera was the dominant figure of the Neapolitan school. Though he was the first major realist painter in Italy and a strong influence against the idealizing trend, both his paintings and his etchings were outside the mainstream of Italian art.

Next to Ribera, Salvator Rosa, an Italian, was the most notable artist of the Neapolitan school, producing a large number of etchings that are full of charm but of no great importance.

Callot's etching technique

Born in Nancy, France, Jacques Callot ran away from home as a boy to study art in Italy. Of all the artists engaged in 17th-century Italian printmaking, he was historically the most significant; for he was one of the first to use repeated bitings on his plates to achieve tonal variations. His drawing style represented a transition between engraving and etching: using a specially shaped etching needle of his own invention, he imitated the swelling and tapering characteristics of the engraved line. His illustrations record and ironically comment upon the customs, historical events, and morals of his time. Callot's work was often decorative and manneristic; but, at his best, as in the series "The Miseries and Disasters of War" (1633), he transcended mere illustration and achieved powerful images of universal significance.

Claude Lorrain, also French-born, was one of the finest landscape painters in Italy, and he had an intuitive understanding of the etching medium. His spontaneous interpretation of the atmospheric quality of his subject

By courtesy of the trustees of the British Museum; J.R. Freeman & Co. Inc.



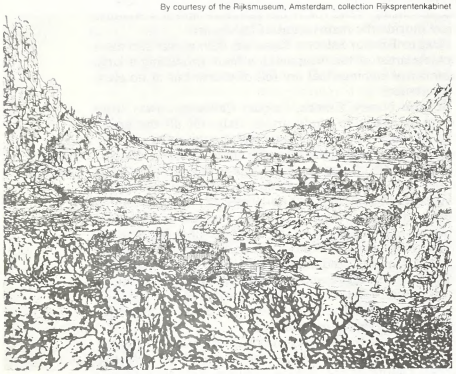
"The Hangman's Tree," etching by Jacques Callot from the series "The Miseries and Disasters of War," 1633. 6.6 cm × 19 cm.

foreshadows the Barbizon school and Impressionism in the 19th century.

The Netherlands. In the beginning of the 17th century, Holland suddenly exploded into a frenzy of creativity in etching. The sensitive, atmospheric etchings of the brothers Esaias and Jan van de Velde can be considered the beginning of the Dutch landscape school. Others were Adriaen van Stalbent, Pieter de Molijn, and Willem Buytewech—all fine printmakers, but all eventually overshadowed by the dramatic personality of Rembrandt. Before him, however, another artist appeared who was so original that no historical precedent could anticipate him. Hercules Seghers is one of the most interesting and mysterious figures in the history of printmaking. He was a lonely, tragic man, an experimenter who was so far ahead of his time that it took centuries for the real significance of his work to become apparent. It is known that Rembrandt appreciated his work. He owned some of Seghers' prints and even reworked one of his plates, "Tobias and the Angel" (*c.* 1633); keeping the landscape, he changed the figures, making it "The Flight into Egypt" (1653).

Seghers was the first real experimenter in intaglio colour printing. His methods were completely unorthodox: he printed on tinted canvas, tried light lines on a dark background, and also mixed printing with hand colouring. Seghers' etching technique was itself very unorthodox. His eroded lines, so well suited to his subject matter, are unlike any etched line made before him, which has led some experts to the conclusion that Seghers invented the lift ground, an aquatint technique.

By courtesy of the Rijksmuseum, Amsterdam, collection Rijksprentenkabinet



"The Rocky River Landscape," 17th-century etching by Hercules Seghers. 17.6 cm × 21.7 cm.

Most of Seghers' etchings represent craggy, arid landscapes. Everything he drew—landscapes, still lifes, even figures—seem to be made of stone. It is a world suspended in the timelessness of death.

The graphic distinction of Rembrandt

Even among the supreme artists of the world, Rembrandt van Rijn occupies a very special place. One of the most eminent painters of all time, he also left a graphic oeuvre of heroic proportions both in quantity and quality. A great innovator, he was the first artist to fully explore the possibilities of the etched line.

Rembrandt made approximately 300 plates. His subject matter represents practically every aspect of human existence: he rendered religious and historical subjects; he explored themes of love and death; and he created profound portraits and sensitive landscapes. Everything that was part of life concerned him, from the highest ideals to the most mundane bodily functions.

While Rembrandt's early prints are pure etchings, his later works frequently combine the techniques of etching and drypoint. Since he often reworked his plates between printings, there are sometimes enormous variations between proofs. Rembrandt's immediate influence on his students and followers was not very productive, for his personality was so overpowering that those close to him fell under his spell and simply imitated his style. Associated artists such as J.G. van Vliet, who copied and reworked many of the master's plates, and Jan Lievens were mere shadows of Rembrandt. Among those closely associated with Rembrandt, probably Ferdinand Bol was the strongest, but even he is dwarfed by comparison.

Also active during Rembrandt's time and somewhat overshadowed by him was Adriaen van Ostade, one of the most gifted of Dutch genre painters. The subjects of both his paintings and prints were taken mainly from the daily lives of simple people, usually peasants. In spirit, his work represented an important departure from the heroic orientation of historic and religious painting, reflecting a crucial social change—the emergence of a middle class in Europe. For the first time, common people replaced the clergy and the nobility as a source of inspiration for an artist. Van Ostade's etching technique was influenced by the early Rembrandt, but his drawing style was personal. It was simple, undramatic, and direct—well suited to his intimate subject matter.

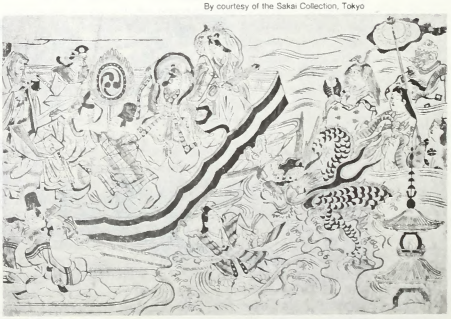
Throughout the 17th century, landscape painting and etching thrived in Holland. Jan van Goyen and Roelant Roghman both made fine landscape paintings and etchings. In this group the most interesting figure is Jacob van Ruisdael, whose sensitive, luminous landscape etchings foreshadowed the Barbizon school.

Toward the end of the century, a strong Italian influence invaded Holland. Since the earthiness of the Dutch temperament did not mix well with the Italian tendency toward idealization, the result was an electicism that drained Dutch art of much of its vitality.

Japanese Ukiyo-e prints. Until the 17th century, Japanese painting was completely dominated by Chinese influence. The Japanese silk paintings and screens of idealized landscapes were hardly distinguishable from their Chinese counterparts. Then, in the early 17th century, an artist of aristocratic origin, Iwasa Matabei, started to paint images related to his environment and personal experience. Although this era of Japanese art history is rather obscure, he is credited with being one of the founders (along with Iwasa Matabei II and Iwasa Matabei of Otsu) of Ukiyo-e, whose woodcuts of the transient world or the world of everyday life represented a drastic break with the classical tradition. Of the three artists Matabei of Otsu was the most original and had the strongest influence on the development of Japanese printmaking. By standards of Western taste, the images the Ukiyo-e school produced are highly stylized and thoroughly refined. Cultured Japanese, however, found them shockingly vulgar. The very fact that ordinary landscapes and the daily life of common people, actors, and courtesans were the inspiration for the Ukiyo-e artist represented a startling departure from tradition. Just as the emerging middle class revolutionized taste in Europe, the prosperous city dwellers of Edo, Kyōto, and Ōsaka de-

Foundation of the Ukiyo-e school

By courtesy of the Sakai Collection, Tokyo



"Tamatori" ("Getting the Pebble"), colour woodcut by Hishikawa Moronobu, 17th century. 30.5 cm × 40 cm.

veloped their own aesthetic subculture. The development of the popular Kabuki theatre, as distinct from the aristocratic Nō drama, parallels the blossoming of Japanese printmaking.

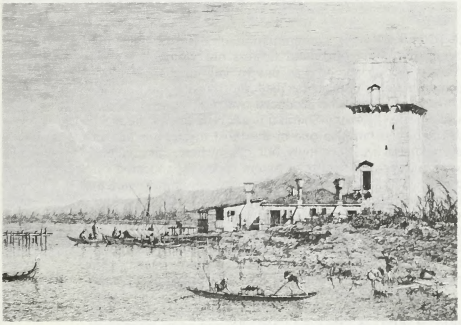
The first great master of Japanese printmaking was Hishikawa Moronobu. A creative innovator, he was the first to use street scenes, peddlers, and crowds as his subject matter and to make his prints available to the common people. As a result, he was looked upon by many as the inventor of printmaking. He illustrated more than 100 books, mirroring the culture and customs of his time. Moronobu's style was a perfect harmony of rhythm, delicacy, and monumental simplicity, leading the way toward the great flowering of Japanese printmaking in the 18th century.

PRINTMAKING IN THE 18TH CENTURY

Italy. In the 18th century, Italy was the most fertile soil in Europe for printmaking. The first outstanding printmaker of the century was the Rococo master Giovanni Battista Tiepolo. His lightly bitten, spontaneous plates reveal superb draftsmanship. With rhythmic, delicate textures, he created a living, luminous space. His 50 plates represent a major contribution to the development of etching—a contribution that was further enhanced by his influence on Goya (see below *Spain*). Giovanni Domenico Tiepolo, the son of Battista, produced a greater quantity of prints than did his father but remained under his influence all his life.

One of the most original printmakers of the period, Canaletto (Giovanni Antonio Canal) created lyrical etchings that were charged with the misty atmosphere of Venice. Inexhaustible in linear and textural invention,

By courtesy of the Yale University Art Gallery, New Haven, Conn., gift of Edward B. Greene



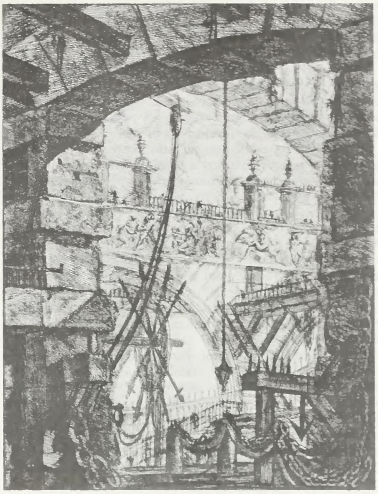
"La Torre di Malghera," etching by Canaletto (1697–1768). 29.8 cm × 43 cm.

they are perfect examples of the simulation of colour and light by purely graphic means. His nephew and pupil, Bernardo Bellotto (Canaletto), who assumed his name, was a prolific printmaker, but, again, he remained under his uncle's influence.

Giambattista Piranesi was the greatest architectural printmaker of his time and probably of all time. Trained as an architect, he was passionately interested in Roman antiquities. Of the approximately 3,000 large etchings completed by Piranesi, all are brilliant, and many rise above documentation. His most important work is the *Carceri d'invenzione* (imaginary prison scenes). The plates, which were made in his youth (published *c.* 1750), are personal, rich, and evocative, far surpassing anything he created after them.

England. Until the 18th century, English printmaking was dominated by foreign influences. William Hogarth, the first major English printmaker, created not only a personal style but a national school. He was a gifted pictorial satirist, belonging in some respects to the tradition of Callot and Goya. He is, however, more earthy than Callot and lacks the savagery of Goya. Hogarth was a printmaker

Pictorial satire of Hogarth



Carceri d'invenzione ("Prisons"), plate 4, etching by Giambattista Piranesi, 1750(?). 55 cm × 41 cm.

By courtesy of the National Gallery of Art, Washington, D.C., Rosenwald Collection

of the people, whose work was so popular that to protect it from imitators he instigated the first engraving copyright act of 1735. Although his drawing was rather pedestrian, Hogarth's prints reveal a sharp observation, projected with robust vitality.

Next to Hogarth, Thomas Rowlandson is the most significant representative of English satire. A brilliant draftsman and a deft caricaturist, he spoofed the moral and social life of England with great humour.

William Blake is by far the most interesting figure in English printmaking. Poet and experimental printmaker, he was a visionary, creating his works totally outside the mainstream of art history. His printed work consisted primarily of book illustrations. Original and inventive in technique, he used a great range of media, from wood and metal engraving to relief etching. In the latter he devised

By courtesy of the Brooklyn Museum, New York, bequest of Samuel E. Haslett



"The Cockpit," engraving by William Hogarth, 1759. 31.7 cm × 38.4 cm.

a transfer method that enabled him to etch the text and the illustration on the same plate. Many of his prints were hand coloured, but he also printed colour by an offset method of his own invention.

Thomas Bewick was a provincial illustrator who made a great number of charming wood engravings, primarily of animals and rural genre scenes. He was a pioneer in the technique of wood engraving, introducing tonal variations by slightly varying the level of his blocks.

To reproduce the fashionable paintings of the day, commercial engravers perfected a whole arsenal of reproduction techniques, such as mezzotint, stipple engraving and etching, and crayon manner.

Spain. Spanish printmaking in the 17th century had been dominated by Flemish and French influences, and no printmaker of importance emerged during the period.

In the 18th-century artist Francisco de Goya, Spain had not only its first truly great printmaker but also the only printmaker whose etchings rival Rembrandt's. Moreover, he is the most eminent satirist printmaking has produced. His visual comments on human folly, war, and religious persecution are devastating.

Goya created four major cycles of prints. The first, "Los caprichos" (1797–98), consists of 80 enigmatic prints commenting on all phases of life. In 1810 he began the 83 plates of "Los desastres de la guerra," a strong visual protest against the brutality of war. After this came "La tauromaquia" (1815–16), a brilliant series on bull fighting. The last important series was "Los disparates," or "Proverbs" (*c.* 1820–24), a biting, though often humorous, interpretation of human folly.

Goya's technique

Technically, the Goya etchings are simple and direct. He usually combined line etching with aquatint; his masterful control of the latter, a relatively new technique, has never been surpassed. Toward the end of his life, he also made a few rich, powerful lithographs.

France. Most 18th-century French etchings were drawings transferred to copper, in which the effects of pencil, pen, or chalk were imitated. Although some distinguished painters, such as Antoine Watteau, made etchings, no prints of importance were produced. Jean-Honoré Fragonard made a few lovely etchings reminiscent of Tiepolo. They have a luminous, transparent quality and express the Rococo spirit but are nevertheless minor works of a major painter. Two artists are notable for technical achievements: Jean-Charles François developed the crayon manner, and Jean-Baptiste Le Prince is credited with the invention of aquatint.

Japan. The first Japanese artist to produce single prints in quantity was Torii Kiyonobu, who specialized in portraits of actors and theatre posters. His school, the Torii, dominated printing for the theatre for two centuries. Another imaginative innovator of the early 18th century was Okumura Masanobu, who experimented with inks, embossing, and gold and silver overlays. He also invented the two-colour print and generally standardized colour printing. His studio greatly influenced the evolution of colour woodcut. Suzuki Harunobu, one of the most charming masters of Japanese woodcut, created prints of infinite delicacy and grace. In this respect he is a forerunner and rival of Utamaro. A highly gifted colourist, he was one of the first to exploit the *nishiki-e,* or full-colour print. He was also the first to colour print backgrounds and to use blind embossing extensively to give his prints three-dimensional textures. Katsukawa Shunshō is notable for his austere portraits of actors, which he designed with much strength and intensity. Some of his portraits are among the finest in Japanese printmaking.

Late 18th-century Ukiyo-e

The period from 1780 to 1790 was dominated by Torii Kiyonaga, whose work represents the Ukiyo-e at its height. He was a great draftsman and designer and could harmonize in his prints the two seemingly contradictory qualities of elegance and power. Kiyonaga was one of the first to experiment with the compositional possibilities of the diptych, triptych, and pentaptych formats. Although he conceived each block as a self-contained unit, they functioned together in harmony. Kitagawa Utamaro can justly be called the supreme poet of Japanese art. Utamaro's prints are the most perfect expression of a tender, loving contemplation of nature, which included not only birds and flowers but women as well. At the age of 50, he was put in jail for an offending print; broken in spirit, he died shortly after his release. During his lifetime he produced over 600 series of books and albums. Toshusai Sharaku is not only one of the most distinguished but also one of the most mysterious figures of Japanese art. Seemingly out of nowhere, his magnificent, powerful portraits of actors suddenly appeared on posters. The boldness of the portraits, verging on caricature; their psychological insight; their richness in colour all represented a daring new attitude. The originality of these prints disturbed the authorities to such an extent that the police prohibited them. In less than two years of working life, Sharaku had produced approximately 145 portraits; then the prodigious flow of work stopped, and he disappeared again.



"Two Lovers Under an Umbrella in Snow," woodcut by Suzuki Harunobu (1724–70). 26.7 cm × 19.7 cm.
By courtesy of the Honolulu Academy of Arts, the James A. Michener Collection

PRINTMAKING IN THE 19TH CENTURY

The 19th century was a turbulent period of art, one aesthetic revolution following the other.

France. French domination of 19th-century art is comparable to northern domination of 15th-century printmaking. Few graphic artists of importance worked outside France. The great French painter Jean-Auguste-Dominique Ingres made only a few etchings, mainly portraits; but, as demonstrated by the lithograph "L'Odalisque" (1825), his draftsmanship was incomparable. Eugène Delacroix left a much more extensive graphic oeuvre: 24 etchings and 131 lithographs. Both in subject matter and style, Delacroix's prints are eloquent expressions of the Romantic spirit. In his tragically short life, Théodore Géricault made a series of powerful lithographs; his horses are considered classics in their genre.

At midcentury, a rebellion against studio painting took place. A group of young landscape painters, most of whom were also printmakers, formed a group that became known as the Barbizon school. The etchings of Charles-François Daubigny, Théodore Rousseau, and Jean-Baptiste-Camille Corot were close to the spirit of the 17th-century Dutch landscapes. Corot made prints whose spontaneity foreshadowed Impressionism; he also experimented with the newly discovered photographic method of *cliché-verre.*

Another member of this group, Jean-François Millet, was concerned particularly with depicting peasant life. His small but simple etchings are reminiscent of the 17th-century Dutch genre painter van Ostade at his best.

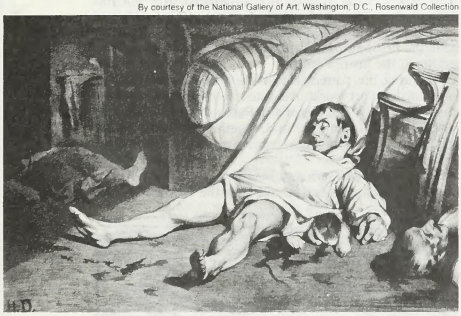
Political satire of Daumier

Honoré Daumier, one of the foremost political satirists of printmaking, was associated with the Barbizon school only through friendships. He produced over 4,000 lithographs (many of them newspaper illustrations), which are visually powerful expressions of his passionate convictions. His best work ranks with that of the greatest masters of printmaking.

A number of French artists were solitary figures working outside of any school; Charles Méryon, Rodolphe Bresdin, and Odilon Redon, for example. Méryon led a short, tragic life, living in poverty and dying insane. His major work is a series of landscapes of Paris—powerfully drawn, moody prints combining an air of mystery with morbid poetry. Bresdin was also a solitary figure, unappreciated and misunderstood most of his life. His etchings and lithographs are characterized by completely personal and fantastically rich imagery. The great symbolist painter Redon initially made prints under the influence of Bresdin. His graphic work—a few etchings but mostly lithographs—consists of about 206 prints, whose strange, often bizarre imagery powerfully influenced the Surrealists of the 20th century.

Although, basically, the Impressionists were concerned with the creation of light through colour, several artists identified with them made major contributions to printmaking. Of these, Édouard Manet and Edgar Degas are the most important. Both were superb draftsmen, and, in spite of their association with an avant-garde movement, their roots were firmly planted in traditional art. Manet made a few fine etchings, but his best and most personal works are lithographs, in which his swift but astonishingly precise drawing found its proper medium. Degas's drawings of horses and ballet dancers are miracles of observation and precision—as are his etchings and lithographs. Degas also made a series of monoprints, including a group of remarkably abstract landscapes. The grand old man of the Impressionists, Camille Pissarro, made 194 prints, both etchings and lithographs. His fine graphic work is representative of forceful Impressionist drawing.

By courtesy of the National Gallery of Art, Washington, D.C., Rosenwald Collection



"Rue Transnonain, April 15, 1834," lithograph by Honoré Daumier. 44.5 cm × 29 cm.

Influence of Japanese woodcuts on European art

The discovery of Japanese colour woodcuts was a revelation that profoundly influenced European art. Until the middle of the 19th century, Japanese printmaking was unknown to the West. As trade relations opened up with Japan, some colour prints came into the hands of young Parisian artists, who responded to the exotic images with great enthusiasm. The simple, abstract handling of colour and design represented a totally new visual experience. Paul Gauguin was one who profited greatly from their influence, which is perhaps more evident in his paintings than in his prints. Following centuries in which the woodcut was used for reproduction, Gauguin's powerful, boldly cut wood blocks were like a breath of fresh air. In the prints of Henri de Toulouse-Lautrec, Japanese influence is more immediate. Although most of his prints were lithographs, the simple bold design, the flat, decorative colour, and the startling disposition of blacks clearly show this influence, which he assimilated and turned into a thoroughly personal expression. A very strong Japanese influence can be seen also in the brilliant color aquatints of the American-born Impressionist Mary Cassatt.



"Le Bon Samaritain" ("The Good Samaritan"), etching by Rodolphe Bresdin (1822–85). 56.3 cm × 44.3 cm.
By courtesy of the Art Institute of Chicago, gift of W.S. Brewster

The giant of Postimpressionism, Paul Cézanne, made three etchings and three lithographs. His immense influence on modern art makes his colour lithograph "The Bathers" (*c.* 1900) an important graphic document. The Dutch artist Johan Barthold Jongkind, who lived in France, created sensitive landscapes and marine etchings that were a transition between the Barbizon school and Impressionism.

By courtesy of the Art Institute of Chicago



"Mahana Atua," woodcut with touches of watercolour by Paul Gauguin, *c.* 1893–95. 18.3 cm × 20.5 cm.

Hokusai's woodcuts

Japan. The most famous Japanese master of woodcut, Hokusai, was born near Edo (Tokyo). From the age of 15, when he became an apprentice, until his death in 1849 at the age of 89, he produced an unending stream of masterpieces—about 35,000 drawings and prints, a staggering figure even considering his long life. He also wrote books and poems. There are few masters in the history of art whose work is comparable to Hokusai's in variety and depth. His interests encompassed history and mythology, popular customs, animal life, and landscape. His output was so enormous and the quality of his work so high that it is difficult to single out individual pieces. The "Thirty-six Views of Mt. Fuji" (*c.* 1826–33) is probably his most popular set of prints. The 15 volumes of the *Hokusai manga* ("Hokusai's Sketches"), published between 1814 and 1878, are fascinating work, for in these rather informal woodcuts the artist gives a comprehensive record of Japanese life and culture. Of all the Japanese masters, the universal genius of Hokusai had the greatest impact on European art.

The last master printmaker of Japan was Andō Hiroshige, whose death in 1858 ends the remarkable dynasty of artists that had begun two centuries earlier. Hiroshige was a great landscape painter and, with Hokusai, the first to capture the European imagination. He was also a versatile artist, famous in Europe as a painter at a time when in Japan he was known mainly as a poet. His greatest period of landscape-painting activity was from 1830 to about 1844. During that time he embarked on a sketching journey (1832), and these sketches formed the basis of "Fifty-three Stages on the Tōkaidō," his most important series of landscapes.

Like the work of the Impressionists in Western art history, Hiroshige's prints are spontaneous both in style and in atmosphere, capturing the essence of the fleeting moments of nature.

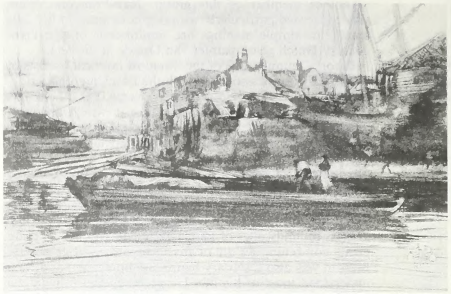
Other countries. In Germany, Max Liebermann made a few etchings of real individuality, but the most important German achievement of the period was the invention of lithography (*c.* 1796) by Aloys Senefelder, who was not an artist. Although the Belgian artist Félicien Rops lived outside France, he was strongly influenced by the school of Paris. His witty, erotic etchings represent a minor but personal expression of the period. In Sweden, the enormously successful Anders Zorn made etchings and drypoints with great virtuosity.

Contributions of Haden and Whistler

English printmaking of the 19th century centred around two great personalities, Sir Francis Seymour Haden and his brother-in-law, James McNeill Whistler. Haden was a Victorian country gentleman, a surgeon who loved and collected etchings. He started to make prints in his leisure time—and ultimately produced over 200 plates. His etchings, sensitively observed documentations of his environment, represent a significant contribution to the English landscape tradition. Whistler was born in America and attended West Point for a period; but he left to study art in Paris, where he met many of the leading artists, including Degas. In 1859 he went to London, where he resided until his death. Whistler was an immensely gifted, complex personality. Simultaneously with his fashionable portraiture, he did a great deal of experimentation; in the nearly abstract paintings and prints that he called "Nocturnes" (begun in 1866), for example, he was far ahead of his time. His graphic oeuvre, 442 etchings and drypoints and 150 lithographs, had great impact on modern printmaking. The freedom and painterliness of Whistler's etchings were particularly significant because they came to act as a strong liberating influence.

Printmaking in 19th-century America was still provincial and did not produce any artist comparable to the European masters. The colour engravings of flora and fauna executed by the naturalist John James Audubon constitute a significant body of work, however.

In Mexico, the popular illustrator José Guadalupe Posada produced thousands of woodcuts and lead cuts for newspapers in a completely original style—a mixture of sophistication and the naïveté of popular art. His work had a substantial influence on the young Mexican revolutionary art movement.



"The Limehouse," lithotint by James McNeill Whistler, *c.* 1887. 17.5 cm × 26.6 cm.

By courtesy of the Philadelphia Museum of Art, bequest of Staunton B. Peck

PRINTMAKING IN THE 20TH CENTURY

Influence of photography on printmaking

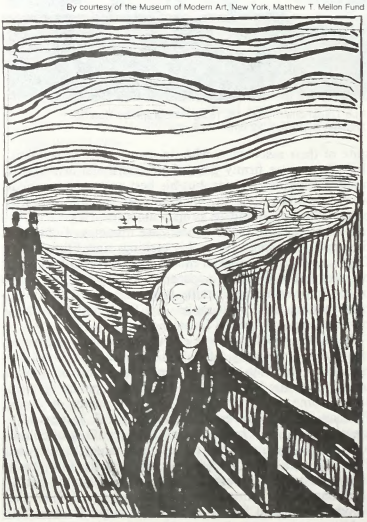
The invention of photography in the early 1800s had a great influence on the development of the visual arts. Its effect was the most immediate on printmaking: photographic reproduction processes made reproductive printmaking obsolete, and printmaking was returned to the creative artist.

The experimental attitude that originated with the Impressionists accelerated in the 20th century. The new styles and new directions that arose with bewildering rapidity made the first half of the century one of the most exciting periods in the history of art.

Continuing the pattern set in the 19th century, France dominated the art world. Attracted by its creative climate, young artists like the Spaniard Pablo Picasso flocked to Paris from other countries and, together with the French, formed the school of Paris, which produced many first-rate artists.

At the same time, Germany became again a vital art centre. German Expressionism and later the Bauhaus school

By courtesy of the Museum of Modern Art, New York, Matthew T. Mellon Fund



"The Cry," lithograph by Edvard Munch, 1895. 35.3 cm × 25.4 cm.

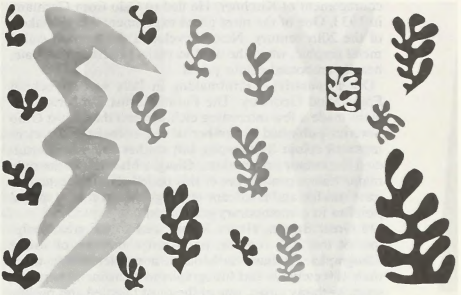
not only produced a number of distinguished artists but eventually exerted international influence.

The following discussion deals only with the "old masters" of contemporary art, those considered to be in historically secure positions. Four transitional figures are singled out as being of particular importance because they represent a bridge between the 19th and 20th centuries. Edvard Munch was an extraordinarily gifted Norwegian painter and printmaker who worked in Paris and in Berlin. His intense imagery, with psychological undertones, relates him to German Expressionism. A versatile artist, he made outstanding etchings, drypoints, colour lithographs, and experimental woodcuts. The Belgian artist James Ensor made superb etchings in a style related to Impressionism, but with fantastic imagery that was close to Surrealism. Close friends, the Frenchmen Pierre Bonnard and Édouard Vuillard produced similar graphic works. Inspired by the Japanese woodcut prints, both made sensitive, beautiful colour lithographs.

France. Pablo Picasso was without doubt the most dramatic and monumental figure of contemporary graphic art. Besides being a superb painter and sculptor, he created a graphic oeuvre so rich and all-encompassing that he stands alone. He made well over 1,000 prints, including etchings, engravings, drypoints, woodcuts, lithographs, and linoleum cuts. Georges Braque, the cofounder with Picasso of Cubism, produced 10 major Cubist etchings. The distinguished French painter Henri Matisse was a remarkable colorist and a highly accomplished draftsman. Although the majority of his more than 500 prints are lithographs, he also made some outstanding line etchings and, late in his career, some cutout prints that are masterpieces of design and colour orchestration. Georges Rouault, the French Expressionist, was a solitary figure in contemporary art. The most important graphic work of this religious painter was the "Miserere," a set of etchings published in 1948. Jacques Villon, a major French printmaker, was recognized late in his life as a great painter. Early in his career he made colour aquatints, after the paintings of his more celebrated contemporaries, that raised the level of intaglio colour printing to new heights. Later he developed a completely personal style within the Cubist tradition. He made more than 600 prints including engravings, etchings, drypoints, and colour lithographs. The poetic, naïve, and, at the same time, sophisticated style of Marc Chagall, a Russian Jewish member of the school of Paris, sets him apart from any art movement. In his significant body of graphic work, the most accomplished prints are illustrations of the Bible, the works of the Russian writer Nikolay Gogol, and the fables of the Frenchman Jean de La Fontaine. Like his compatriot Picasso, the Spanish painter and sculptor Joan Miró was a prolific printmaker. His witty colour etchings and lithographs represent an achievement equal to his paintings. Max Ernst was a founder of Surrealism and one of the most inventive and influential members of the group. In his extensive graphic work, he introduced a number of new techniques; most notable was his imaginative use of the "collage" in printmaking. Stanley William Hayter, an English painter-printmaker who lived in Paris, has an important position in the development of contemporary experimental printmaking. His significance lies not only in his work as an artist but also in his influence as a teacher. In the 1930s his "Atelier 17" printmaking group was the centre of experimental intaglio work in Paris. In the 1940s he came to the United States and, through his teaching in New York, exercised a powerful influence on contemporary American printmaking. Other artists who did noteworthy graphic work in France include Jean (Hans) Arp, Salvador Dalí, André Derain, Jan Dubuffet, André Dunoyer de Segonzac, Alberto Giacometti, Fernand Léger, André Masson, Louis Marcoussis, and Jules Pascin.

Influence of Hayter on printmaking

Germany. Unlike the extremely varied school of Paris, German Expressionism was quite homogeneous and also much less international. The Expressionists were not united by an aesthetic theory but by their human attitudes and spiritual aspirations. Nearly all of them were active in printmaking, and, although they worked in every contemporary graphic medium, the directness of drypoint and woodcut most appealed to their temperaments.



Jazz, Plate XV, serigraph by Henri Matisse, 1947. 41.9 cm × 64.7 cm.
By courtesy of the Fogg Art Museum, Harvard University, Cambridge, Mass.

Lovis Corinth represents a transition from 19th-century naturalism to the Expressionist movement. Although Corinth made etchings, woodcuts, and lithographs, his rich, virile drypoints are his best work. Although not innovative, Käthe Kollwitz's moving, powerful protest prints against war and poverty are significant graphic statements. Ernst Ludwig Kirchner, one of the major figures of German Expressionism, produced a rich graphic oeuvre consisting of etchings, lithographs, and woodcuts. His experimental colour woodcuts represent one of the most distinguished achievements in contemporary graphic art. Emil Nolde produced prints characterized by violent imagery. He worked spontaneously, often making woodcuts without preliminary drawings. Although Nolde came late to graphic work, he left an impressive number of woodcuts, etchings, and lithographs. Max Beckmann was an outstanding draftsman who made many woodcuts and drypoints. In the latter technique he created some of the finest portrait prints of the 20th century. During World War II Beckmann came to the United States, where he exerted considerable influence through his teaching. George Grosz used etchings and lithographs to give savage expression to his social criticism of Germany between the wars. The following Expressionists also left significant graphic work: Ernst Barlach, Erich Heckel, Oskar Kokoschka, and Karl Schmidt-Rottluff.

Artists of the Bauhaus group

Other artists moved in a more formal, abstract direction. Based on their philosophy of "new objectivity," they founded the Bauhaus school in Germany in 1919. The two major artists in this group were the Russian Wassily Kandinsky and the Swiss Paul Klee. Kandinsky was one of the great innovators of contemporary art. In his early, lyrical paintings he was a forerunner of Abstract Expressionism, and in his late mature work he introduced Geometric Abstraction. His graphic work consists of an impressive number of woodcuts and lithographs. The whimsical, lyrical abstractions of Klee also had great influence on the course of modern art. His work—about 120 etchings and lithographs—is full of graphic invention and a rare sense of humour. Lyonel Feininger, born in the United States of German parents, studied in Europe and worked most of his life in Germany. He was associated with the Blaue Reiter group (artists who wished to express through their work the spiritual realities they felt had been ignored by the Impressionists) and then in 1919–33 with the Bauhaus. Feininger concentrated mostly on landscapes, executed in a very personal Cubist style, and was one of the most productive graphic artists at the Bauhaus. In the beginning, he made some etchings and lithographs but from 1918 worked mainly in woodcuts. Josef Albers, also associated with the Bauhaus, was born in Germany and moved to the United States in 1933. He made a considerable number of prints, including colour silk screens. Rolf Nesch was born in Germany, where he started printmaking with the en-

couragement of Kirchner. He fled to Oslo from Germany in 1933. One of the most gifted experimental printmakers of the 20th century, Nesch developed the method called metal graphic, which he used to make extremely intricate, heavily embossed colour prints.

Other countries. Printmaking in Italy was far behind France and Germany. The Futurist artist Umberto Boccioni made a few interesting etchings and the Cubist Gino Severini published a number of rather manneristic etchings and colour lithographs, but neither could be considered important printmakers. Giorgio Morandi is the only major Italian printmaker of this period. His intimate, delicate still-life and landscape etchings occupy a very special position in contemporary graphic art.

In Great Britain, Henry Moore, one of the great sculptors of the 20th century, published a number of strong lithographs. Graham Sutherland, a painter, made more than 100 etchings and lithographs in a distinctly personal style. Anthony Gross, one of the most talented and prolific English printmakers, has published an impressive body of excellent landscape etchings and engravings. Among later artists, the imaginative and personal graphic work of David Hockney should be singled out.

In the United States, after the turn of the century, most of the prominent painters became fairly active printmakers: George Wesley Bellows, in lithography; John Sloan and Reginald Marsh, in etching; Milton Avery, in drypoint and a large number of monoprints; and Stuart Davis, in colour lithography. Among these painter-printmakers, two artists are particularly notable: Edward Hopper, whose few etchings are very personal and of unusually high quality; and Ben Shahn, an extremely prolific printmaker, who left an impressive graphic oeuvre in practically every medium. Of the present generation of established painter-printmakers, only a few are creatively involved in the process, while the rest let the commercial printer take over.

Revival of the Japanese woodcut

A revival of the art of the woodcut began in Japan in the late 1920s as part of the modern art movement. Onchi Kōshirō and Hiratsuka Un-ichi were early exponents who, though working in different styles, did most for the renaissance of this national art, which thrived once again after World War II. Among the notable woodcut artists of the postwar period are Munakata Shikō and Saitō Kiyoshi.

Since the mid-20th century, there has been a spectacular increase in printmaking activity. Artists all over the world are enthusiastically working and experimenting in every conceivable medium. In this period probably more prints were made and more technical innovations introduced than in the previous history of printmaking.

BIBLIOGRAPHY. DONALD SAFF and DELI SACILOTTO, *Printmaking: History and Processes* (1978), with emphasis on the multitude of techniques; RIVA CASTLEMAN, *Prints of the Twentieth Century* (1976), a popularly written survey; FRITZ EICHENBERG, *The Art of the Print* (1976), general history and technique; ARTHUR M. HIND, *A History of Engraving and Etching from the 15th Century to the Year 1914*, 3rd ed. rev. (1923, reprinted 1963), and *An Introduction to a History of Woodcut, with a Detailed Survey of Work Done in the Fifteenth Century*, 2 vol. (1935, reprinted 1963), cover brilliantly the whole history and development of Western printmaking; and JAY A. LEVENSON, KONRAD OBERHUBER, JACQUELYN L. SHEEHAN, *Early Italian Engravings from the National Gallery of Art* (1973), a thorough study of the Italian Renaissance. Other studies particularly recommended are: JEAN LARAN, *L'Estampe*, 2 vol. (1959), excellent documentation coupled with a volume of fine reproductions; WILLY BOLLER, *Masterpieces of the Japanese Color Woodcut* (1957), not a scholarly book but it covers well the high points of Japanese printmaking; CARL ZIGROSSER, *The Book of Fine Prints*, rev. ed. (1956), an easy-to-read introduction into the history of printmaking; ELLEN S. JACOBOWITZ and STEPHANIE L. STEPENAK, *The Prints of Lucas Van Leyden and His Contemporaries* (1983), excellent documentation of the period; DAVID FREEDBERG, *Dutch Landscape Prints of the Seventeenth Century* (1980); A. HYATT MAYOR, *Giovanni Battista Piranesi* (1952), a fine biography of Piranesi with an excellent selection of illustrations; WOLF STUBBE, *Graphic Arts of the Twentieth Century* (1963; originally published in German, 1962), good introduction into the history of contemporary printmaking; JAMES WATROUS, *American Printmaking* (1984), covering 1880 to 1980; UNA E. JOHNSON, *American Prints and Printmakers* (1980), a comprehensive study covering 1900–80; KAREN F. BEALL (comp.), *American Prints in the Library of Congress* (1970); E.S. LUMSDEN, *The Art of Etching* (1929, reprinted 1962), excellent document on the traditional etching techniques; WILLI KURTH (ed.), *The Complete Woodcuts of Albrecht Dürer* (1946), primarily a picture book with historical background; CARL ZIGROSSER and CHRISTA M. GAEHDE, *A Guide to the Collecting and Care of Original Prints* (1965), a wealth of indispensable information for the collector; STANLEY W. HAYTER, *About Prints* (1962), challenging ideas about printmaking by an important artist and teacher; RICHARD T. GODFREY, *Printmaking in Britain: A General History from Its Beginnings to the Present Day* (1978); LEO C. COLLINS, *Hercules Seghers* (1953), an excellent, well-documented book on one of the most important printmakers; LUDWIG MÜNZ, *Rembrandt Etchings*, 2 vol. (1949), interesting because it documents Rembrandt's influence as a teacher; K.G. BOON, *Rembrandt: The Complete Etchings* (1963, reissued 1978), one of the finest books on Rembrandt with excellent rich reproductions; GABOR PETERDI, *Printmaking*, rev. ed. (1971, reissued 1980), a simple but thorough book on both the traditional and experimental intaglio and woodcut methods; *Redon, Moreau, Bresdin*, Museum of Modern Art, New York (1961), primarily interesting for the Bresdin documentation; CARL W. SCHRAUBSTADTER, *Care and Repair of Japanese Prints*, ACL ed. (1978), useful information for the collector; *The Complete Etchings of Goya* (1943), primarily a picture book; ANDRÉ MALRAUX, *Saturne: An Essay on Goya* (1957; originally published in French, 1950), a very subjective response to Goya by a great writer, richly illustrated; MAXIME LALANNE, *A Treatise on Etching* (1880; trans. of 2nd French ed., 1878), primarily interesting as a historical document on technique; FRANK and DOROTHY GETLEIN, *The Bite of the Print: Satire and Irony in Woodcuts, Engravings, Etchings, Lithographs and Serigraphs* (1963), an interesting book highlighting the militant, political aspect of prints; BERNARD S. MYERS, *The German Expressionists* (1957), a good introduction to the history of German Expressionism, well illustrated; JEAN ADHEMAR, *Toulouse-Lautrec: His Complete Lithographs and Drypoints* (1965; originally published in French, 1965), the definitive book on Toulouse-Lautrec as a printmaker; MICHEL MELOT, *Graphic Art of the Pre-Impressionists*, trans. from the French (1981); MICHAEL KNIGIN and MURRAY ZIMILES, *The Technique of Fine Art Lithography*, rev. ed. (1977), an excellent, well-organized book on a complex subject; JOAN LUDMAN and LAURIS MASON (comps.), *Print Reference: A Selected Bibliography of Print-Related Literature* (1982), a classified bibliography of most English-language 20th-century works on the history, production, collecting, and care of fine prints.

(G.F.P.)

Probability Theory

The word probability has several meanings in ordinary conversation. Two of these are particularly important for the development and applications of the mathematical theory of probability. One is the interpretation of probabilities as relative frequencies, for which simple games involving coins, cards, dice, and roulette wheels provide examples. The distinctive feature of games of chance is that the outcome of a given trial cannot be predicted with certainty, although the collective results of a large number of trials display some regularity. For example, the statement that the probability of "heads" in tossing a coin equals one-half, according to the relative frequency interpretation, implies that in a large number of tosses the relative frequency with which "heads" actually occurs will be approximately one-half, although it contains no implication concerning the outcome of any given toss. There are many similar examples involving collections of people, molecules of a gas, genes, and so on. Actuarial statements about the life expectancy for persons of a certain age describe the collective experience of a large number of individuals but do not purport to say what will happen to any particular person. Similarly, predictions about the chance of a genetic disease occurring in a child of parents having a known genetic makeup are statements about relative frequencies of occurrence in a large number of cases but are not predictions about a given individual.

A second interpretation of probability, as a personal measure of uncertainty, is discussed below. For further discussion of the applications of probability theory, see STATISTICS.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 10/22 and 10/23, and the *Index*.

The article is divided into the following sections:

Development of probability theory 135
Experiments, sample space, events, and equally likely probabilities 135
Conditional probability 137
Random variables, distributions, expectation, and variance 138
An alternative interpretation of probability 140
The law of large numbers, the central limit theorem, and the Poisson approximation 140
Infinite sample spaces and axiomatic probability 141
Conditional expectation and least squares prediction 144
The Poisson process and the Brownian motion process 144
Stochastic processes 146
Bibliography 148

DEVELOPMENT OF PROBABILITY THEORY

Although some ideas about probability appear in works from antiquity, the systematic development of probability theory began only in the 16th and 17th centuries when the European mathematicians Gerolamo Cardano, Pierre de Fermat, Blaise Pascal, and Christiaan Huygens began to analyze simple games of chance involving cards and dice. One of the first attempts to use ideas of relative frequency to study human populations arose in the mortality statistics assembled by John Graunt in 17th-century London at the time of the plague.

The most important early contributions to probability theory were Jakob Bernoulli's *Ars Conjectandi* (1713; "The Art of Conjecturing") and Abraham de Moivre's *Doctrine of Chances* (1718, 1738, 1756). Bernoulli was the most distinguished of a distinguished family of Swiss mathematicians. He contributed to many branches of mathematics, and to probability theory he gave the law of large numbers, the first precise mathematical theorem to interpret probabilities in terms of observable relative frequencies. De Moivre was a French Protestant who fled to England to escape persecution. He is known today primarily because of his work in probability theory, notably his contribution to the central limit theorem, which along with the law of large numbers is discussed below.

After a century of comparatively slow progress, the development of probability theory accelerated in the mid-19th century. Pafnuty Lvovich Chebyshev began a tradition of elegant mathematical contributions that continues in Russia to the present day. British interest in heredity, spearheaded in its quantitative development by Sir Francis Galton, and the rapid development at the turn of the century of the physical theories of Brownian motion and statistical mechanics provided scientific sources of new problems.

In the 20th century the mathematical theory of probability has been given an axiomatic foundation, and like other branches of mathematics its development from these axioms depends only on its logical correctness, whether or not its theorems refer to phenomena of the physical world. Nevertheless, its usefulness in describing variable or uncertain phenomena explains the ubiquity of probability theory in modern science and technology.

This article contains a description of the important mathematical concepts of probability theory, illustrated by some of the applications that have stimulated their development. Since applications inevitably involve simplifying assumptions that focus on some features of a problem at the expense of others, it is advantageous to begin, like the 17th-century mathematicians mentioned above, by thinking about simple experiments, such as tossing a coin or rolling dice, and later to see how these apparently frivolous investigations relate to important scientific questions.

EXPERIMENTS, SAMPLE SPACE, EVENTS, AND EQUALLY LIKELY PROBABILITIES

The fundamental ingredient of probability theory is an experiment that can be repeated, at least hypothetically, under essentially identical conditions and that may lead to different outcomes on different trials. The set of all possible outcomes of an experiment is called a "sample space." The sample space

The experiment of tossing a coin once results in a sample space with two possible outcomes, "heads" and "tails." Tossing two dice has a sample space with 36 possible outcomes, each of which can be identified with an ordered pair (i, j), where i and j assume one of the values 1, 2, 3, 4, 5, 6 and denote the faces showing on the individual dice. It is important to think of the dice as identifiable (say by a difference in colour), so that the outcome (1, 2) is different from (2, 1). An "event" is a well-defined subset of the sample space. For example, the event "the sum of the faces showing on the two dice equals six" consists of the five outcomes (1, 5), (2, 4), (3, 3), (4, 2), and (5, 1).

A third example is to draw n balls from an urn containing balls of various colours. A generic outcome to this experiment is an n-tuple, where the ith entry specifies the colour of the ball obtained on the ith draw ($i = 1, 2, \ldots, n$). In spite of the simplicity of this experiment, a thorough understanding gives the theoretical basis for opinion polls and sample surveys. For example, individuals in a population favouring a particular candidate in an election may be identified with balls of a particular colour, those favouring a different candidate may be identified with a different colour, and so on. Probability theory provides the basis for learning about the contents of the urn from the sample of balls drawn from the urn; an application is

to learn about the electoral preferences of a population on the basis of a sample drawn from that population.

Another application of simple urn models is to clinical trials designed to determine whether a new treatment for a disease, a new drug, or a new surgical procedure is better than a standard treatment. In the simple case in which treatment can be regarded as either success or failure, the goal of the clinical trial is to discover whether the new treatment more frequently leads to success than does the standard treatment. Patients with the disease can be identified with balls in an urn. The red balls are those patients who are cured by the new treatment, and the black balls are those not cured. Usually there is a control group, who receive the standard treatment. They are represented by a second urn with a possibly different fraction of red balls. The goal of the experiment of drawing some number of balls from each urn is to discover on the basis of the sample which urn has the larger fraction of red balls. A variation of this idea can be used to test the efficacy of a new vaccine. Perhaps the largest and most famous example was the test of the Salk vaccine for poliomyelitis conducted in 1954. It was organized by the United States Public Health Service and involved almost two million children. Its success has led to the almost complete elimination of polio as a health problem in the industrialized parts of the world. Strictly speaking, these applications are problems of statistics, for which the foundations are provided by probability theory.

Probability theory and statistics

In contrast to the experiments described above, many experiments have infinitely many possible outcomes. For example, one can toss a coin until "heads" appears for the first time. The number of possible tosses is $n = 1, 2, \ldots$. Another example is to twirl a spinner. For an idealized spinner made from a straight line segment having no width and pivoted at its centre, the set of possible outcomes is the set of all angles that the final position of the spinner makes with some fixed direction, equivalently all real numbers in $[0, 2\pi)$. Many measurements in the natural and social sciences, such as volume, voltage, temperature, reaction time, marginal income, and so on, are made on continuous scales and at least in theory involve infinitely many possible values. If the repeated measurements on different subjects or at different times on the same subject can lead to different outcomes, probability theory is a possible tool to study this variability.

Because of their comparative simplicity, experiments with finite sample spaces are discussed first. In the early development of probability theory, mathematicians considered only those experiments for which it seemed reasonable, based on considerations of symmetry, to suppose that all outcomes of the experiment were "equally likely." Then in a large number of trials all outcomes should occur with approximately the same frequency. The probability of an event is defined to be the ratio of the number of cases favourable to the event—*i.e.*, the number of outcomes in the subset of the sample space defining the event—to the total number of cases. Thus, the 36 possible outcomes in the throw of two dice are assumed equally likely, and the probability of obtaining "six" is the number of favourable cases, 5, divided by 36, or 5/36.

Now suppose that a coin is tossed n times, and consider the probability of the event "heads does not occur" in the n tosses. An outcome of the experiment is an n-tuple, the kth entry of which identifies the result of the kth toss. Since there are two possible outcomes for each toss, the number of elements in the sample space is 2^n. Of these, only one outcome corresponds to having no heads, so the required probability is $1/2^n$.

It is only slightly more difficult to determine the probability of "at most one head." In addition to the single case in which no head occurs, there are n cases in which exactly one head occurs, because it can occur on the first, second, . . . , or nth toss. Hence, there are $n+1$ cases favourable to obtaining at most one head, and the desired probability is $(n+1)/2^n$.

This last example illustrates the fundamental principle that, if the event whose probability is sought can be represented as the union of several other events that have no outcomes in common ("at most one head" is the union of "no heads" and "exactly one head"), then the probability of the union is the sum of the probabilities of the individual events making up the union. To describe this situation symbolically, let S denote the sample space. For two events A and B, the intersection of A and B is the set of all experimental outcomes belonging to both A and B and is denoted $A \cap B$; the union of A and B is the set of all experimental outcomes belonging to A or B (or both) and is denoted $A \cup B$. The impossible event—*i.e.*, the event containing no outcomes—is denoted by $\emptyset$. The probability of an event A is written $P(A)$. The principle of addition of probabilities is that, if $A_1, A_2, \ldots, A_n$ are events with $A_i \cap A_j = \emptyset$ for all pairs $i \neq j$, then

The principle of additivity

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n). \tag{1}$$

Equation (1) is consistent with the relative frequency interpretation of probabilities; for, if $A_i \cap A_j = \emptyset$ for all $i \neq j$, the relative frequency with which at least one of the A_i occurs equals the sum of the relative frequencies with which the individual A_i occur.

Equation (1) is fundamental for everything that follows. Indeed, in the modern axiomatic theory of probability, which eschews a definition of probability in terms of "equally likely outcomes" as being hopelessly circular, an extended form of equation (1) plays a basic role (see the discussion of axiomatic probability below).

An elementary, useful consequence of (1) is the following. With each event A is associated the complementary event A^c consisting of those experimental outcomes that do not belong to A. Since $A \cap A^c = \emptyset$, $A \cup A^c = S$, and $P(S) = 1$ (where S denotes the sample space), it follows from (1) that $P(A^c) = 1 - P(A)$. For example, the probability of "at least one head" in n tosses of a coin is one minus the probability of "no head," or $1 - 1/2^n$.

A basic problem first solved by Jakob Bernoulli is to find the probability of obtaining exactly i red balls in the experiment of drawing n times at random with replacement from an urn containing b black and r red balls. To draw at random means that, on a single draw, each of the $r+b$ balls is equally likely to be drawn and, since each ball is replaced before the next draw, there are $(r+b) \times \cdots \times (r+b) = (r+b)^n$ possible outcomes to the experiment. Of these possible outcomes, the number that is favourable to obtaining i red balls and $n-i$ black balls in any one particular order is

$$\overbrace{r \times r \times \cdots \times r}^{i} \times \overbrace{b \times b \times \cdots \times b}^{n-i} = r^i \times b^{n-i}.$$

The number of possible orders in which i red balls and $n-i$ black balls can be drawn from the urn is the binomial coefficient

$$\binom{n}{i} = \frac{n!}{i!(n-i)!}, \tag{2}$$

where $k! = k \times (k-1) \times \cdots \times 2 \times 1$ for positive integers k, and $0! = 1$. Hence, the probability in question, which equals the number of favourable outcomes divided by the number of possible outcomes, is

$$\binom{n}{i} \frac{r^i b^{n-i}}{(r+b)^n} = \binom{n}{i} p^i q^{n-i} \qquad (i = 0, 1, 2, \ldots, n), \tag{3}$$

where $p = r/(r+b)$ and $q = b/(r+b) = 1 - p$.

For example, suppose $r = 2b$ and $n = 4$. According to (3), the probability of "exactly two red balls" is $\binom{4}{2}(\frac{2}{3})^2(\frac{1}{3})^2 = 6 \times \frac{4}{81} = \frac{8}{27}$. In this case the $\binom{4}{2} = 6$ possible outcomes are easily enumerated: (*rrbb*), (*rbrb*), (*brrb*), (*rbbr*), (*brbr*), (*bbrr*).

(For a derivation of (2), observe that in order to draw exactly i red balls in n draws one must either (*a*) draw i red balls in the first $n-1$ draws and a black ball on the nth draw or (*b*) draw $i-1$ red balls in the first $n-1$ draws followed by the ith red ball on the nth draw. Hence, $\binom{n}{i} = \binom{n-1}{i} + \binom{n-1}{i-1}$, from which (2) can be verified by induction on n.)

Two related examples are (i) drawing without replacement from an urn containing r red and b black balls and (ii) drawing with or without replacement from an urn

containing balls of s different colors. If n balls are drawn without replacement from an urn containing r red and b black balls, the number of possible outcomes is $\binom{r+b}{n}$, of which the number favourable to drawing i red and $n-i$ black balls is $\binom{r}{i}\binom{b}{n-i}$. Hence, the probability of drawing exactly i red balls in n draws is the ratio

$$\frac{\binom{r}{i}\binom{b}{n-i}}{\binom{r+b}{n}}.$$

If an urn contains balls of s different colors in the ratios $p_1:p_2:\ldots:p_s$, where $p_1+\cdots+p_s=1$, and if n balls are drawn with replacement, the probability of obtaining i_1 balls of the first colour, i_2 balls of the second colour, and so on is the multinomial probability

$$\frac{n!}{i_1!i_2!\ldots i_s!}p_1^{i_1}p_2^{i_2}\ldots p_s^{i_s}.$$

The evaluation of (3) with pencil and paper grows increasingly difficult with increasing n. It is even more difficult to evaluate related cumulative probabilities—for example the probability of obtaining "at most j red balls" in the n draws, which can be expressed as the sum of (3) for $i=0, 1, \ldots, j$. The problem of approximate computation of probabilities that are known in principle is a recurrent theme throughout the history of probability theory and will be discussed in more detail below.

The birthday problem

An entertaining example is to determine the probability that in a randomly selected group of n people at least two have the same birthday. If one assumes for simplicity that a year contains 365 days and that each day is equally likely to be the birthday of a randomly selected person, then in a group of n people there are 365^n possible combinations of birthdays. Of these possibilities, the number that is unfavourable to the event that at least two have the same birthday is that the first person has any of the 365 days for his birthday, the second any of the other 364 days for his birthday, the third any of the remaining 363 days, ..., and the nth any of the remaining $365-n+1$. Hence, the number of ways that all n people can have different birthdays is $365\times 364\times\cdots\times(365-n+1)$, so that the probability that at least two have the same birthday is

$$P=1-\frac{365\times 364\times\cdots\times(365-n+1)}{365^n}.$$

Numerical evaluation shows, rather surprisingly, that for $n=23$ the probability that at least two people have the same birthday is about 0.5. For $n=42$ the probability is about 0.9.

This example illustrates that applications of probability theory to the physical world are facilitated by assumptions that are not strictly true, although they should be approximately true. Thus, the assumptions that a year has 365 days and that all days are equally likely to be the birthday of a random individual are false, because one year in four has 366 days and because birth dates are not distributed uniformly throughout the year. Moreover, if one attempts to apply this result to an actual group of individuals, it is necessary to ask what it means for these to be "randomly selected." It would naturally be unreasonable to apply it to a group known to contain twins. In spite of the obvious failure of the assumptions to be literally true, as a classroom example it rarely disappoints instructors of classes having more than forty students.

CONDITIONAL PROBABILITY

Suppose two balls are drawn sequentially without replacement from an urn containing r red and b black balls. The probability of getting a red ball on the first draw is $r/(r+b)$. If, however, one is told that a red ball was obtained on the first draw, the conditional probability of getting a red ball on the second draw is $(r-1)/(r+b-1)$, because for the second draw there are $r+b-1$ balls in the urn, of which $r-1$ are red. Similarly, if one is told that the first ball drawn is black, the conditional probability of getting red on the second draw is $r/(r+b-1)$.

In a number of trials the relative frequency with which B occurs among those trials in which A occurs is just the frequency of occurrence of $A\cap B$ divided by the frequency of occurrence of A. This suggests that the conditional probability of B given A (denoted $P(B|A)$ should be defined by

$$P(B|A)=\frac{P(A\cap B)}{P(A)}. \tag{4}$$

If A denotes a red ball on the first draw and B a red ball on the second draw in the experiment of the preceding paragraph, then $P(A)=r/(r+b)$ and

$$P(A\cap B)=\frac{r(r-1)}{[(r+b)(r+b-1)]},$$

so (4) is consistent with the "obvious" answer derived above.

Rewriting (4) as $P(A\cap B)=P(A)P(B|A)$ and adding to this expression the same expression with A replaced by A^c ("not A") leads via (1) to the equality

$$\begin{aligned}P(B)&=P(A\cap B)+P(A^c\cap B)\\&=P(A)P(B|A)+P(A^c)P(B|A^c).\end{aligned}$$

More generally, if $A_1, A_2, \ldots, A_n$ are mutually exclusive events and their union is the entire sample space, so that exactly one of the A_i must occur, essentially the same argument gives a fundamental relation, which is frequently called the law of total probability:

The law of total probability

$$\begin{aligned}P(B)=P(A_1)P(B|A_1)&+P(A_2)P(B|A_2)\\&+\cdots+P(A_n)P(B|A_n).\end{aligned}$$

An application of the law of total probability to a problem originally posed by Huygens is to find the probability of "gambler's ruin." Suppose two players, often called Peter and Paul, initially have x and $m-x$ dollars, respectively. A ball, which is red with probability p and black with probability $q=1-p$, is drawn from an urn. If a red ball is drawn, Paul must pay Peter one dollar, while Peter must pay Paul one dollar if the ball drawn is black. The ball is replaced, and the game continues until one of the players is ruined. It is quite difficult to determine the probability of Peter's ruin by a direct analysis of all possible cases. But let $Q(x)$ denote that probability as a function of Peter's initial fortune x and observe that after one draw the structure of the rest of the game is exactly as it was before the first draw, except that Peter's fortune is now either $x+1$ or $x-1$ according to the results of the first draw. The law of total probability with $A=$ {red ball on first draw} and $A^c=$ {black ball on first draw} shows that

$$Q(x)=pQ(x+1)+qQ(x-1). \tag{5}$$

This equation holds for $x=2, 3, \ldots, m-2$. It also holds for $x=1$ and $m-1$ if one adds the boundary conditions $Q(0)=1$ and $Q(m)=0$, which say that, if Peter has 0 dollars initially, his probability of ruin is 1, while, if he has all m dollars, he is certain to win.

It can be verified by direct substitution that equation (5) together with the indicated boundary conditions are satisfied by

$$\begin{aligned}Q(x)&=\frac{\left(\frac{q}{p}\right)^x-\left(\frac{q}{p}\right)^m}{1-\left(\frac{q}{p}\right)^m} \qquad \left(p\neq\frac{1}{2}\right)\\&=1-\frac{x}{m} \qquad \left(p=\frac{1}{2}\right)\end{aligned} \tag{6}$$

With some additional analysis it is possible to show that these give the only solutions of (5) and hence must be the desired probabilities.

Suppose $m=10x$, so that Paul initially has nine times as much money as Peter. If $p=\frac{1}{2}$, the probability of Peter's ruin is 0.9 regardless of the values of x and m. If $p=0.51$, so that each trial slightly favours Peter, the situation is quite different. For $x=1$ and $m=10$, the probability of Peter's ruin is 0.88, only slightly less than before. However, for $x=100$ and $m=1{,}000$, Peter's slight advantage on each trial becomes so important that the probability of his ultimate ruin is now less than 0.02.

Generalizations of the problem of gambler's ruin play

an important role in statistical sequential analysis, developed by the Hungarian-born American statistician Abraham Wald in response to the demand for more efficient methods of industrial quality control during World War II. They also enter into insurance risk theory, which is discussed below.

The following example shows that, even when it is given that A occurs, it is important in evaluating $P(B|A)$ to recognize that A^c might have occurred, and hence in principle it must be possible also to evaluate $P(B|A^c)$. By lot, two out of three prisoners—Sam, Jean, and Chris—are chosen to be executed. There are $\binom{3}{2} = 3$ possible pairs of prisoners to be selected for execution, of which two contain Sam, so the probability that Sam is slated for execution is $\frac{2}{3}$. Sam asks the guard which of the others is to be executed. Since at least one must be, it appears that the guard would give Sam no information by answering. After hearing that Jean is to be executed, Sam reasons that, since either he or Chris must be the other one, the conditional probability that he will be executed is $\frac{1}{2}$. Thus, it appears that the guard has given Sam some information about his own fate. However, the experiment is incompletely defined, because it is not specified how the guard chooses whether to answer "Jean" or "Chris" in case both of them are to be executed. If the guard answers "Jean" with probability p, the conditional probability of the event "Sam will be executed" given "the guard says Jean will be executed" is

$$\frac{\frac{1}{3}}{\frac{1}{3}+\frac{p}{3}} = \frac{1}{1+p}.$$

Only in the case $p = 1$ is Sam's reasoning correct. If $p = \frac{1}{2}$, the guard in fact gives no information about Sam's fate.

The concept of independence

One of the most important concepts in probability theory is that of "independence." The events A and B are said to be (stochastically) independent if $P(B|A) = P(B)$, or equivalently by (4) if

$$P(A \cap B) = P(A)P(B). \tag{7}$$

The intuitive meaning of the definition in terms of conditional probabilities is that the probability of B is not changed by knowing that A has occurred. Equation (7) shows that the definition is symmetric in A and B.

It is intuitively clear that, in drawing two balls with replacement from an urn containing r red and b black balls, the event "red ball on the first draw" and the event "red ball on the second draw" are independent. (This statement presupposes that the balls are thoroughly mixed before each draw.) An analysis of the $(r+b)^2$ equally likely outcomes of the experiment shows that the formal definition (7) is indeed satisfied.

In terms of the concept of independence, the experiment leading to the binomial distribution (3) can be described as follows. On a single trial a particular event has probability p. An experiment consists of n independent repetitions of this trial. The probability that the particular event occurs exactly i times is given by (3).

Independence plays a central role in the law of large numbers, the central limit theorem, the Poisson process, and the Brownian motion process (see below).

Consider now the defining relation (4) for the conditional probability $P(A_n|B)$, where the A_i are mutually exclusive and their union is the entire sample space. Substitution of $P(A_n)P(B|A_n)$ in the numerator of (4) and substitution of the right-hand side of the law of total probability in the denominator yields a result known as Bayes's theorem (after the 18th-century English clergyman Thomas Bayes) or the law of inverse probability:

$$P(A_n|B) = \frac{P(A_n)P(B|A_n)}{\sum_i P(A_i)P(B|A_i)}.$$

As an example, suppose that two balls are drawn without replacement from an urn containing r red and b black balls. Let A be the event "red on the first draw" and B the event "red on the second draw." From the obvious relations $P(A) = r/(r+b) = 1 - P(A^c)$, $P(B|A) = (r-1)/(r+b-1)$, $P(B|A^c) = r/(r+b-1)$, and Bayes's theorem, it follows that the probability of a red ball on the first draw given that the second one is known to be red equals $(r-1)/(r+b-1)$. A more interesting and important use of Bayes's theorem appears below in the discussion of subjective probabilities.

RANDOM VARIABLES, DISTRIBUTIONS, EXPECTATION, AND VARIANCE

Usually it is more convenient to associate numerical values with the outcomes of an experiment than to work directly with a nonnumerical description such as "red ball on the first draw." For example, an outcome of the experiment of drawing n balls with replacement from an urn containing black and red balls is an n-tuple that tells us whether a red or a black ball was drawn on each of the draws. This n-tuple is conveniently represented by an n-tuple of ones and zeros, where the appearance of a one in the kth position indicates that a red ball was drawn on the kth draw. A quantity of particular interest is the number of red balls drawn, which is just the sum of the entries in this numerical description of the experimental outcome. Mathematically a rule that associates with every element of a given set a unique real number is called a "(real-valued) function." In the history of statistics and probability, real-valued functions defined on a sample space have traditionally been called "random variables." Thus, if a sample space S has the generic element e, the outcome of an experiment, then a random variable is a real-valued function $X = X(e)$. Customarily one omits the argument e in the notation for a random variable. For the experiment of drawing balls from an urn containing black and red balls, R, the number of red balls drawn, is a random variable. A particularly useful random variable is $1[A]$, the indicator variable of the event A, which equals 1 if A occurs and 0 otherwise. If A_k denotes the event "a red ball is drawn on the kth draw," then R has the useful representation: $R = 1[A_1] + \cdots + 1[A_n]$. A "constant" is a trivial random variable that always takes the same value regardless of the outcome of the experiment.

Random variables

Suppose X is a random variable that can assume one of the values $x_1, x_2, \ldots, x_m$ according to the outcome of a random experiment, and consider the event $(X = x_i)$, which is a shorthand notation for the set of all experimental outcomes e such that $X(e) = x_i$. The probability of this event, $P(X = x_i)$, is itself a function of x_i, called the "(probability) distribution" of X. Thus, the distribution of the random variable R defined in the preceding paragraph is the function of $i = 0, 1, \ldots, n$ given in (3). Introducing the notation $f(x_i) = P(X = x_i)$, one sees from the basic properties of probabilities that

$$f(x_i) \geq 0 \text{ for all } i, \quad \sum_i f(x_i) = 1,$$

and

$$P(a < X \leq b) = \sum_{a < x_i \leq b} f(x_i),$$

for any real numbers a and b. If Y is a second random variable defined on the same sample space as X and taking the values $y_1, y_2, \ldots, y_n$, the function of two variables $h(x_i, y_j) = P(X = x_i, Y = y_j)$ is called the "joint distribution" of X and Y. Since $(X = x_i) = \cup_j (X = x_i, Y = y_j)$, and this union consists of disjoint events in the sample space,

$$f(x_i) = \sum_j h(x_i, y_j), \quad \text{for all } i. \tag{8}$$

Often f is called the "marginal distribution" of X to emphasize its relation to the joint distribution of X and Y. Similarly, $g(y_j) = \sum_i h(x_i, y_j)$ is the (marginal) distribution of Y. The random variables X and Y are defined to be independent if the events $(X = x_i)$ and $(Y = y_j)$ are independent for all i and j—*i.e.*, if $h(x_i, y_j) = f(x_i)g(y_j)$ for all i and j. The joint distribution of an arbitrary number of random variables is defined similarly.

Suppose two dice are thrown. Let X denote the sum of the numbers appearing on the two dice, and let Y denote the number of even numbers appearing. The possible values of X are 2, 3, . . . , 12, while the possible values

Joint Distribution of X and Y

j	i											row sum $= g(j)$
	2	3	4	5	6	7	8	9	10	11	12	
0	1/36	0	1/18	0	1/12	0	1/18	0	1/36	0	0	1/4
1	0	1/18	0	1/9	0	1/6	0	1/9	0	1/18	0	1/2
2	0	0	1/36	0	1/18	0	1/12	0	1/18	0	1/36	1/4
column sum $= f(i)$	1/36	1/18	1/12	1/9	5/36	1/6	5/36	1/9	1/12	1/18	1/36	

of Y are 0, 1, 2. Since there are 36 possible outcomes for the two dice, the accompanying table giving the joint distribution $h(i, j)$ $(i = 2, 3, \ldots, 12; j = 0, 1, 2)$ and the marginal distributions $f(i)$ and $g(j)$ is easily computed by direct enumeration.

For more complex experiments, determination of a complete probability distribution usually requires a combination of theoretical analysis and empirical experimentation and is often very difficult. Consequently, it is desirable to describe a distribution insofar as possible by a small number of parameters that are comparatively easy to evaluate and interpret. The most important are the mean and the variance. These are both defined in terms of the "expected value" of a random variable.

The expected value of a random variable

Given a random variable X with distribution f, the expected value of X, denoted $E(X)$, is defined by $E(X) = \sum_i x_i f(x_i)$. In words, the expected value of X is the sum of each of the possible values of X multiplied by the probability of obtaining that value. The expected value of X is also called the "mean" of the distribution f. The basic property of E is that of linearity: if X and Y are random variables and if a and b are constants, then $E(aX + bY) = aE(X) + bE(Y)$. To see why this is true, note that $aX + bY$ is itself a random variable, which assumes the values $ax_i + by_j$ with the probabilities $h(x_i, y_j)$. Hence,

$$E(aX + bY) = \sum_{i,j} (ax_i + by_j)h(x_i, y_j)$$

$$= a \sum_{i,j} x_i h(x_i, y_j) + b \sum_{i,j} y_j h(x_i, y_j).$$

If the first sum on the right-hand side is summed over j while holding i fixed, by (8) the result is $\sum_i x_i f(x_i)$, which by definition is $E(X)$. Similarly, the second sum equals $E(Y)$.

If $1[A]$ denotes the "indicator variable" of A—*i.e.*, a random variable equal to 1 if A occurs and equal to 0 otherwise—then $E\{1[A]\} = 1 \times P(A) + 0 \times P(A^c) = P(A)$. This shows that the concept of expectation includes that of probability as a special case.

As an illustration, consider the number R of red balls in n draws with replacement from an urn containing a proportion p of red balls. From the definition and the distribution of R given in (3), $E(R) = \sum_i i\binom{n}{i} p^i q^{n-i}$, which can be evaluated by algebraic manipulation and found to equal np. It is easier to use the representation $R = 1[A_1] + \cdots + 1[A_n]$, where A_k denotes the event "the kth draw results in a red ball." Since $E\{1[A_k]\} = p$ for all k, by linearity $E(R) = E\{1[A_1]\} + \cdots + E\{1[A_n]\} = np$. This argument illustrates the principle that one can often compute the expected value of a random variable without first computing its distribution. For another example, suppose n balls are dropped at random into n boxes. The number of empty boxes, Y, has the representation $Y = 1[B_1] + \cdots + 1[B_n]$, where B_k is the event that "the kth box is empty." Since the kth box is empty if and only if each of the n balls went into one of the other $n - 1$ boxes, $P(B_k) = [(n-1)/n]^n$ for all k, and consequently $E(Y) = n(1 - 1/n)^n$. The exact distribution of Y is very complicated, especially if n is large.

Many probability distributions have small values of $f(x_i)$ associated with extreme (large or small) values of x_i and larger values of $f(x_i)$ for intermediate x_i. For example, both marginal distributions in the table are symmetrical about a midpoint that has relatively high probability, and the probability of other values decreases as one moves away from the midpoint. Insofar as a distribution $f(x_i)$ follows this kind of pattern, one can interpret the mean of f as a rough measure of location of the bulk of the probability distribution, because in the defining sum the values x_i associated with large values of $f(x_i)$ more or less define the center of the distribution. In the extreme case, the expected value of a constant random variable is just that constant.

It is also of interest to know how closely packed about its mean value a distribution is. The most important measure of concentration is the "variance," denoted by $\mathrm{Var}(X)$ and defined by $\mathrm{Var}(X) = E\{[X - E(X)]^2\}$. By linearity of expectations, one has equivalently $\mathrm{Var}(X) = E(X^2) - (E(X))^2$. The "standard deviation" of X is the square root of its variance. It has a more direct interpretation than the variance because it is in the same units as X. The variance of a constant random variable is 0. Also, if c is a constant, $\mathrm{Var}(cX) = c^2\mathrm{Var}(X)$. Variance

There is no general formula for the expectation of a product of random variables. If the random variables X and Y are independent, $E(XY) = E(X)E(Y)$. This can be used to show that, if $X_1, \ldots, X_n$ are independent random variables, the variance of the sum $X_1 + \cdots + X_n$ is just the sum of the individual variances, $\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)$. If the Xs have the same distribution and are independent, the variance of the average $(X_1 + \cdots + X_n)/n$ is $\mathrm{Var}(X_1)/n$. Equivalently, the standard deviation of $(X_1 + \cdots + X_n)/n$ is the standard deviation of X_1 divided by $n^{1/2}$. This quantifies the intuitive notion that the average of repeated observations is less variable than the individual observations. More precisely, it says that the variability of the average is inversely proportional to the square root of the number of observations. This result is tremendously important in problems of statistical inference. (See the discussion of the law of large numbers and the central limit theorem below.)

Consider again the binomial distribution (3). As in the calculation of the mean value, one can use the definition combined with some algebraic manipulation to show that, if R has the distribution defined by (3), then $\mathrm{Var}(R) = npq$. From the representation $R = 1[A_1] + \cdots + 1[A_n]$ defined above, and the observation that the events A_k are independent and have the same probability, it follows that

$$\mathrm{Var}(R) = \mathrm{Var}\{1[A_1]\} + \cdots + \mathrm{Var}\{1[A_n]\} = n\,\mathrm{Var}\{1[A_1]\}.$$

Moreover,

$$\mathrm{Var}\{1[A_1]\} = E\{1[A_1]^2\} - [E\{1[A_n]\}]^2 = p - p^2 = pq,$$

so $\mathrm{Var}(R) = npq$. Also, $\mathrm{Var}(R/n) = \mathrm{Var}(R)/n^2 = pq/n$.

The conditional distribution of Y given $X = x_i$ is defined by:

$$P\{Y = y_j \mid X = x_i\} = \frac{h(x_i, y_j)}{f(x_i)}$$

(compare (4)), and the conditional expectation of Y given $X = x_i$ is

$$E(Y \mid X = x_i) = \sum_j \frac{y_j h(x_i, y_j)}{f(x_i)}. \tag{9}$$

One can regard $E(Y \mid X)$ as a function of X, since X is a random variable, this function of X must itself be a random variable. The conditional expectation $E(Y \mid X)$ considered as a random variable has its own (unconditional) expectation $E\{E(Y \mid X)\}$, which is calculated by multiplying (9) by $f(x_i)$ and summing over i to obtain the important formula

$$E\{E(Y \mid X)\} = E(Y). \tag{10}$$

Properly interpreted, equation (10) is a generalization of the law of total probability.

For a simple example of the use of (10), recall the problem of gambler's ruin and let $e(x)$ denote the expected duration of the game if Peter's fortune is initially equal to x. The reasoning leading to (5) in conjunction with (10) shows that $e(x)$ satisfies the equations $e(x) = 1 + pe(x+1) + qe(x-1)$ for $x = 1, 2, \ldots, m-1$ with the boundary conditions $e(0) = e(m) = 0$. The solution for $p \neq 1/2$ is rather complicated; for $p = 1/2$, $e(x) = x(m-x)$.

AN ALTERNATIVE INTERPRETATION OF PROBABILITY

In ordinary conversation the word probability is applied not only to variable phenomena but also to propositions of uncertain veracity. The truth of any proposition concerning the outcome of an experiment is uncertain before the experiment is performed. Many other uncertain propositions cannot be defined in terms of repeatable experiments. An individual can be uncertain about the truth of a scientific theory, a religious doctrine, or even about the occurrence of a specific historical event when inadequate or conflicting eyewitness accounts are involved. Using probability as a measure of uncertainty enlarges its domain of application to phenomena that do not meet the requirement of repeatability. The concomitant disadvantage is that probability as a measure of uncertainty is subjective and varies from one person to another.

Probability as a measure of uncertainty

According to one interpretation, to say that someone has subjective probability p that a proposition is true means that for any integers r and b with $r/(r+b) < p$, if that individual is offered an opportunity to bet the same amount on the truth of the proposition or on "red in a single draw" from an urn containing r red and b black balls, he prefers the first bet, while, if $r/(r+b) > p$, he prefers the second bet.

An important stimulus to modern thought about subjective probability has been an attempt to understand decision making in the face of incomplete knowledge. It is assumed that an individual, when faced with the necessity of making a decision that may have different consequences depending on situations about which he has incomplete knowledge, can express his personal preferences and uncertainties in a way consistent with certain axioms of rational behaviour. It can then be deduced that the individual has a "utility function," which measures the value to him of each course of action when each of the uncertain possibilities is the true one, and a "subjective probability distribution," which expresses quantitatively his beliefs about the uncertain situations. The individual's optimal decision is the one that maximizes his expected utility with respect to his subjective probability. The concept of utility goes back at least to Daniel Bernoulli (Jakob Bernoulli's nephew) and has been developed in the twentieth century by John von Neumann and Oskar Morgenstern, Frank P. Ramsey, and Leonard J. Savage, among others. Ramsey and Savage have stressed the importance of subjective probability as a concomitant ingredient of decision making in the face of uncertainty. An alternative approach to subjective probability without the use of utility theory has been developed by Bruno de Finetti.

The mathematical theory of probability is the same regardless of one's interpretation of the concept, although the importance attached to various results can depend very much on the interpretation. In particular, in the theory and applications of subjective probability, Bayes's theorem plays an important role.

For example, suppose that an urn contains N balls, r of which are red and $b = N - r$ of which are black, but r (hence b) is unknown. One is permitted to learn about the value of r by performing the experiment of drawing with replacement n balls from the urn. Suppose also that one has a subjective probability distribution giving the probability $f(r)$ that the number of red balls is in fact r where $f(0) + \cdots + f(N) = 1$. This distribution is called an a priori distribution because it is specified prior to the experiment of drawing balls from the urn. The binomial probability (3) is now a conditional probability, given the value of r. Finally, one can use Bayes's theorem to find the conditional probability that the unknown number of red balls in the urn is r, given that the number of red balls drawn from the urn is i. The result is

A priori and a posteriori distribution

$$\frac{f(r)r^i b^{n-i}}{\sum_{r_0=0}^{n} f(r_0) r_0^i b_0^{n-i}}, \qquad \text{where } b_0 = N - r_0.$$

This distribution, derived by using Bayes's theorem to combine the a priori distribution with the conditional distribution for the outcome of the experiment, is called the a posteriori distribution.

The virtue of this calculation is that it makes possible a probability statement about the composition of the urn, which is not directly observable, in terms of observable data, from the composition of the sample taken from the urn. The weakness, as indicated above, is that different people may choose different subjective probabilities for the composition of the urn a priori and hence reach different conclusions about its composition a posteriori.

To see how this idea might apply in practice, consider a simple urn model of opinion polling to predict which of two candidates will win an election. The red balls in the urn are identified with voters who will vote for candidate A and the black balls with those voting for candidate B. Choosing a sample from the electorate and asking their preferences is a well-defined random experiment, which in theory and in practice is repeatable. The composition of the urn is uncertain and is not the result of a well-defined random experiment. Nevertheless, to the extent that a vote for a candidate is a vote for a political party, other elections provide information about the content of the urn, which if used judiciously should be helpful in supplementing the results of the actual sample to make a prediction. Exactly how to use this information is a difficult problem in which individual judgment plays an important part. One possibility is to incorporate the prior information into an a priori distribution about the electorate, which is then combined via Bayes's theorem with the outcome of the sample and summarized by an a posteriori distribution.

THE LAW OF LARGE NUMBERS, THE CENTRAL LIMIT THEOREM, AND THE POISSON APPROXIMATION

The relative-frequency interpretation of probability is that, if an experiment is repeated a large number of times under identical conditions and independently, then the relative frequency with which an event A actually occurs and the probability of A should be approximately the same. A mathematical expression of this interpretation is the law of large numbers. This theorem says that, if $X_1, X_2, \ldots, X_n$ are independent random variables having a common distribution with mean μ, then for any number $\varepsilon > 0$, no matter how small, as $n \to \infty$,

$$P\{|n^{-1}(X_1 + \cdots + X_n) - \mu| < \varepsilon\} \to 1. \tag{11}$$

The law of large numbers was first proved by Jakob Bernoulli in the special case where X_k is 1 or 0 according as the kth draw (with replacement) from an urn containing r red and b black balls is red or black. Then $E(X_k) = r/(r+b)$, and (11) says that the probability that "the difference between the empirical proportion of red balls in n draws and the probability of red on a single draw is less than ε" converges to 1 as n becomes infinitely large.

Insofar as an event which has probability very close to 1 is practically certain to happen, this result justifies the relative-frequency interpretation of probability. Strictly speaking, however, the justification is circular because the probability in (11), which is very close to but not equal to 1, requires its own relative-frequency interpretation. Perhaps it is better to say that the weak law of large numbers is consistent with the relative-frequency interpretation of probability.

The following simple proof of the law of large numbers is based on Chebyshev's inequality, which illustrates the sense in which the variance of a distribution measures how the distribution is dispersed about its mean. If X is a random variable with distribution f and mean μ, then by definition $\mathrm{Var}(X) = \sum_i (x_i - \mu)^2 f(x_i)$. Since all terms in this sum are positive, the sum can only decrease if some of the terms are omitted. Suppose one omits all terms with $|x_i - \mu| < b$, where b is an arbitrary, given number. Each term remaining in the sum has a factor of the form $(x_i - \mu)^2$,

Chebyshev's inequality

which is greater than or equal to b^2. Hence, $\mathrm{Var}(X) \geq b^2 \sum' f(x_i)$, where the prime on the summation sign indicates that only terms with $|x_i - \mu| \geq b$ are included in the sum. Chebyshev's inequality is this expression rewritten as

$$P(|X - \mu| \geq b) \leq \frac{\mathrm{Var}(X)}{b^2}.$$

This inequality can be applied to the complementary event of that appearing in (11), with $b = \varepsilon$. The Xs are independent and have the same distribution, $E[n^{-1}(X_1 + \cdots + X_n)] = \mu$ and $\mathrm{Var}[(X_1 + \cdots + X_n)/n] = \mathrm{Var}(X_1)/n$, so that

$$P(|\frac{(X_1 + \cdots + X_n)}{n} - \mu| < \varepsilon) \geq 1 - \frac{\mathrm{Var}(X_1)}{n\varepsilon^2}.$$

This not only proves (11), but it also says quantitatively how large n should be in order that the empirical average, $n^{-1}(X_1 + \cdots + X_n)$, approximate its expectation to any required degree of precision.

Suppose, for example, that the proportion p of red balls in an urn is unknown and is to be estimated by the empirical proportion of red balls in a sample of size n drawn from the urn with replacement. Chebyshev's inequality with $X_k = 1\{\text{red ball on the } k\text{th draw}\}$ implies that, in order that the observed proportion be within ε of the true proportion p with probability at least 0.95, it suffices that n be at least $20 \times \mathrm{Var}(X_1)/\varepsilon^2$. Since $\mathrm{Var}(X_1) = p(1-p) \leq \frac{1}{4}$ for all p, for $\varepsilon = 0.03$ it suffices that n be at least 5,555. It is shown below that this value of n is much larger than necessary, because Chebyshev's inequality is not sufficiently precise to be useful in numerical calculations.

Although Jakob Bernoulli did not know Chebyshev's inequality, the inequality he derived was also imprecise, and, perhaps because of his disappointment in not having a quantitatively useful approximation, he did not publish the result during his lifetime. It appeared in 1713, eight years after his death.

The central limit theorem

The desired useful approximation is given by the central limit theorem, which in the special case of the binomial distribution was first discovered by de Moivre about 1730. Let $X_1, \ldots, X_n$ be independent random variables having a common distribution with expectation μ and variance σ^2. The law of large numbers implies that the distribution of the random variable $\overline{X}_n = n^{-1}(X_1 + \cdots + X_n)$ is essentially just the degenerate distribution of the constant μ, because $E(\overline{X}_n) = \mu$ and $\mathrm{Var}(\overline{X}_n) = \sigma^2/n \to 0$ as $n \to \infty$. The standardized random variable $(\overline{X}_n - \mu)/(\sigma/n^{1/2})$ has mean 0 and variance 1. The central limit theorem gives the remarkable result that, for any real numbers a and b, as $n \to \infty$,

$$P(a < \frac{(\overline{X}_n - \mu)}{(\sigma/n^{1/2})} \leq b) \to G(b) - G(a), \qquad (12)$$

where $G(z) = (2\pi)^{-1/2} \int_{-\infty}^{z} \exp(-t^2/2)\, dt$. Thus, if n is large, the standardized average has a distribution that is approximately the same, regardless of the original distribution of the Xs. Equation (12) also illustrates clearly the "square root law": the accuracy of $\overline{X}_n$ as an estimator of μ is inversely proportional to the square root of the sample size n.

Use of (12) to evaluate approximately the probability on the left-hand side of (11), by setting $b = -a = \varepsilon n^{1/2}/\sigma$, yields the approximation $G(\varepsilon n^{1/2}/\sigma) - G(-\varepsilon n^{1/2}/\sigma)$. Since $G(2) - G(-2)$ is approximately 0.95, n must be about $4\sigma^2/\varepsilon^2$ in order that the difference $|\overline{X}_n - \mu|$ will be less than ε with probability 0.95. For the special case of the binomial distribution, one can again use the inequality $\sigma^2 = p(1-p) \leq \frac{1}{4}$ and now conclude that about 1,100 balls must be drawn from the urn in order that the empirical proportion of red balls drawn will be within 0.03 of the true proportion of red balls with probability about 0.95. The frequently appearing statement in newspapers in the United States that a given opinion poll involving a sample of about 1,100 persons has a sampling error of no more than 3 percent is based on this kind of calculation. The qualification that this 3 percent sampling error may be exceeded in about 5 percent of the cases is often omitted. (The actual situation in opinion polls or sample surveys generally is more complicated. The sample is drawn without replacement, so strictly speaking the binomial distribution is not applicable. However, the "urn"—*i.e.*, the population from which the sample is drawn—is extremely large, in many cases infinitely large for practical purposes. Hence, the composition of the urn is effectively the same throughout the sampling process, and the binomial distribution applies as an approximation. Also, the population is usually stratified into relatively homogeneous groups, and the survey is designed to take advantage of this stratification. To pursue the analogy with urn models, one can imagine the balls to be in several urns in varying proportions, and one must decide how to allocate the n draws from the various urns so as to estimate efficiently the overall proportion of red balls.)

Considerable effort has been put into generalizing both the law of large numbers and the central limit theorem, so that it is unnecessary for the variables to be either independent or identically distributed.

The law of large numbers discussed above is often called the "weak law of large numbers," to distinguish it from the "strong law," a conceptually different result discussed below in the section on infinite probability spaces.

The weak law of large numbers and the central limit theorem give information about the distribution of the proportion of successes in a large number of independent trials when the probability of success on each trial is p. In the mathematical formulation of these results, it is assumed that p is an arbitrary, but fixed, number in the interval (0, 1) and $n \to \infty$, so that the expected number of successes in the n trials, np, also increases toward $+\infty$ with n. A rather different kind of approximation is of interest when n is large and the probability p of success on a single trial is inversely proportional to n, so that $np = \mu$ is a fixed number even though $n \to \infty$. An example is the following simple model of radioactive decay of a source consisting of a large number of atoms, which independently of one another decay by spontaneously emitting a particle. The time scale is divided into a large number of very small intervals of equal lengths, and in each interval, independently of what happens in the other intervals, the source emits one or no particle with probability p or $q = 1 - p$ respectively. It is assumed that the intervals are so small that the probability of two or more particles being emitted in a single interval is negligible. One now imagines that the size of the intervals shrinks to 0, so that the number of trials up to any fixed time t becomes infinite. It is reasonable to assume that the probability of emission during a short time interval is proportional to the length of the interval. The result is a different kind of approximation to the binomial distribution, called the Poisson approximation (after the French mathematician Siméon-Denis Poisson), or the law of small numbers.

The Poisson approximation

Assume, then, that a biased coin having probability $p = \mu\delta$ of heads is tossed once in each time interval of length δ, so that by time t the total number of tosses is an integer n approximately equal to t/δ. Introducing these values into (3) and passing to the limit as $\delta \to 0$ gives as the distribution for $N(t)$ the number of radioactive particles emitted in time t:

$$P(N(t) = i) = \frac{(\mu t)^i \exp(-\mu t)}{i!} \qquad (i = 0, 1, \ldots). \qquad (13)$$

The right-hand side of (13) is called the Poisson distribution. Its mean and variance are both equal to μt. Although the Poisson approximation is not comparable to the central limit theorem in importance, it nevertheless provides one of the basic building blocks in the theory of stochastic processes (see below).

INFINITE SAMPLE SPACES AND AXIOMATIC PROBABILITY

The experiments described in the preceding discussion involve finite sample spaces for the most part, although the central limit theorem and the Poisson approximation involve limiting operations and hence lead to integrals and infinite series. In a finite sample space, calculation of the probability of an event A is conceptually straightforward because the principle of additivity tells one to calculate the probability of a complicated event as the sum of the probabilities of the individual experimental outcomes whose union defines the event.

Experiments having a continuum of possible outcomes—for example, that of selecting a number at random from the interval $[r, s]$—involve subtle mathematical difficulties that were not satisfactorily resolved until the 20th century. If one chooses a number at random from $[r, s]$, the probability that the number falls in any interval $[x, y]$ must be proportional to the length of that interval; and, since the probability of the entire sample space $[r, s]$ equals 1, the constant of proportionality equals $1/(s-r)$. Hence, the probability of obtaining a number in the interval $[x, y]$ equals $(y-x)/(s-r)$. From this and the principle of additivity one can determine the probability of any event that can be expressed as a finite union of intervals. There are, however, very complicated sets having no simple relation to the intervals—*e.g.*, the rational numbers—and it is not immediately clear what the probabilities of these sets should be. Also, the probability of selecting exactly the number x must be 0, because the set consisting of x alone is contained in the interval $[x, x+1/n]$ for all n and hence must have no larger probability than $1/[n(s-r)]$, no matter how large n is. Consequently, it makes no sense to try to compute the probability of an event by "adding" the probabilities of the individual outcomes making up the event, because each individual outcome has probability 0.

A closely related experiment, although at first there appears to be no connection, arises as follows. Suppose that a coin is tossed n times, and let $X_k = 1$ or 0 according as the outcome of the kth toss is heads or tails. The weak law of large numbers given above says that a certain sequence of numbers—namely the sequence of probabilities given in (11) and defined in terms of these n Xs—converges to 1 as $n \to \infty$. In order to formulate this result, it is only necessary to imagine that one can toss the coin n times and that this finite number of tosses can be arbitrarily large. In other words, there is a sequence of experiments, but each one involves a finite sample space. It is also natural to ask whether the sequence of random variables $(X_1 + \cdots + X_n)/n$ converges as $n \to \infty$. However, this question cannot even be formulated mathematically unless infinitely many Xs can be defined on the same sample space, which in turn requires that the underlying experiment involve an actual infinity of coin tosses.

For the conceptual experiment of tossing a fair coin infinitely many times, the sequence of zeros and ones, $(X_1, X_2, \ldots)$, can be identified with that real number that has the Xs as the coefficients of its expansion in the base 2, namely $X_1/2^1 + X_2/2^2 + X_3/2^3 + \cdots$. For example, the outcome of getting heads on the first two tosses and tails thereafter corresponds to the real number $1/2 + 1/4 + 0/8 + \cdots = 3/4$. (There are some technical mathematical difficulties that arise from the fact that some numbers have two representations. Obviously $1/2 = 1/2 + 0/4 + \cdots$, and the formula for the sum of an infinite geometric series shows that it also equals $0/2 + 1/4 + 1/8 + \cdots$. It can be shown that these difficulties do not pose a serious problem, and they are ignored in the subsequent discussion.) For any particular specification $i_1, i_2, \ldots, i_n$ of zeros and ones, the event $(X_1 = i_1, X_2 = i_2, \ldots, X_n = i_n)$ must have probability $1/2^n$ in order to be consistent with the experiment of tossing the coin only n times. Moreover, this event corresponds to the interval of real numbers $[i_1/2^1 + i_2/2^2 + \cdots + i_n/2^n,\ i_1/2^1 + i_2/2^2 + \cdots + i_n/2^n + 1/2^n]$ of length $1/2^n$, since any continuation $X_{n+1}, X_{n+2}, \ldots$ corresponds to a number that is at least 0 and at most $1/2^{n+1} + 1/2^{n+2} + \cdots = 1/2^n$ by the formula for an infinite geometric series. It follows that the mathematical model for choosing a number at random from $[0, 1]$ and that of tossing a fair coin infinitely many times assign the same probabilities to all intervals of the form $[k/2^n, l/2^n]$.

The mathematical relation between these two experiments was recognized in 1909 by the French mathematician Émile Borel, who used the then new ideas of measure theory to give a precise mathematical model and to formulate what is now called the "strong law of large numbers" for fair coin tossing. His results can be described as follows. Let e denote a number chosen at random from $[0, 1]$, and let $X_k(e)$ be the kth coordinate in the expansion of e to the base 2. Then $X_1, X_2, \ldots$ are an infinite sequence of independent random variables taking the values 0 or 1 with probability $1/2$ each. Moreover, the subset of $[0, 1]$ consisting of those e for which the sequence $n^{-1}[X_1(e) + \cdots + X_n(e)]$ tends to $1/2$ as $n \to \infty$ has probability 1. Symbolically:

$$P\left\{\lim_{n\to\infty}[n^{-1}(X_1 + \cdots + X_n)] = \frac{1}{2}\right\} = 1. \tag{14}$$

The strong law of large numbers

The weak law of large numbers given in (11) says that for any $\varepsilon > 0$, for each sufficiently large value of n, there is only a small probability of observing a deviation of $\bar{X}_n = n^{-1}(X_1 + \cdots + X_n)$ from $1/2$ which is larger than ε; nevertheless, it leaves open the possibility that sooner or later this rare event will occur if one continues to toss the coin and observe the sequence for a sufficiently long time. The strong law, however, asserts that the occurrence of even one value of $\bar{X}_k$ for $k \geq n$ that differs from $1/2$ by more than ε is an event of arbitrarily small probability provided n is large enough. The proof of (14) and various subsequent generalizations is much more difficult than that of the weak law of large numbers. The adjectives "strong" and "weak" refer to the fact that the truth of a result such as (14) implies the truth of the corresponding version of (11), but not conversely.

During the two decades following 1909, measure theory was used in many concrete problems of probability theory, notably in the American mathematician Norbert Wiener's treatment (1923) of the mathematical theory of Brownian motion (see below), but the notion that all problems of probability theory could be formulated in terms of measure is customarily attributed to the Soviet mathematician Andrey Nikolayevich Kolmogorov in 1933.

The fundamental quantities of the measure-theoretic foundation of probability theory are the sample space S, which as before is just the set of all possible outcomes of an experiment, and a distinguished class M of subsets of S, called events. Unlike the case of finite S, in general not every subset of S is an event. The class M must have certain properties described below. Each event is assigned a probability, which means mathematically that a probability is a function P mapping M into the real numbers that satisfies certain conditions derived from one's physical ideas about probability.

The properties of M are as follows: (i) $S \in M$; (ii) if $A \in M$, then $A^c \in M$; (iii) if $A_1, A_2, \ldots \in M$, then $A_1 \cup A_2 \cup \cdots \in M$. Recalling that M is the domain of definition of the probability P, one can interpret (i) as saying that $P(S)$ is defined, (ii) as saying that, if the probability of A is defined, then the probability of "not A" is also defined, and (iii) as saying that, if one can speak of the probability of each of a sequence of events A_n individually, then one can speak of the probability that at least one of the A_n occurs. A class of subsets of any set that has properties (i)–(iii) is called a "σ-field." From these properties one can prove others. For example, it follows at once from (i) and (ii) that $\emptyset$ (the empty set) belongs to the class M. Since the intersection of any class of sets can be expressed as the complement of the union of the complements of those sets (DeMorgan's law), it follows from (ii) and (iii) that, if $A_1, A_2, \ldots \in M$, then $A_1 \cap A_2 \cap \cdots \in M$.

Given a set S and a σ-field M of subsets of S, a probability measure is a function P that assigns to each set $A \in M$ a nonnegative real number and that has the following two properties: (*a*) $P(S) = 1$ and (*b*) if $A_1, A_2, \ldots \in M$ and $A_i \cap A_j = \emptyset$ for all $i \neq j$, then $P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots$. Property (*b*) is called the "axiom of countable additivity." It is clearly motivated by equation (1), which suffices for finite sample spaces because there are only finitely many events. In infinite sample spaces it implies, but is not implied by, equation (1). There is, however, nothing in one's intuitive notion of probability that requires the acceptance of this property. Indeed, a few mathematicians have developed probability theory with only the weaker axiom of finite additivity, but the absence of interesting models that fail to satisfy the axiom of countable additivity has led to its virtually universal acceptance.

The axiom of countable additivity

To get a better feeling for this distinction, consider the experiment of tossing a biased coin having probability p of heads and $q = 1 - p$ of tails until heads first appears.

To be consistent with the idea that the tosses are independent, the probability that exactly n tosses are required equals $q^{n-1}p$, since the first $n-1$ tosses must be tails, and they must be followed by a head. One can imagine that this experiment never terminates—*i.e.*, that the coin continues to turn up tails forever. By the axiom of countable additivity, however, the probability that heads occurs at some finite value of n equals $p+qp+q^2p+\cdots=p/(1-q)=1$, by the formula for the sum of an infinite geometric series. Hence, the probability that the experiment goes on forever equals 0. Similarly, one can compute the probability that the number of tosses is odd, as $p+q^2p+q^4p+\cdots=p/(1-q^2)=1/(1+q)$. On the other hand, if only finite additivity were required, it would be possible to define the following admittedly bizarre probability. The sample space S is the set of all natural numbers, and the σ-field M is the class of all subsets of S. If an event A contains finitely many elements, $P(A)=0$, and, if the complement of A contains finitely many elements, $P(A)=1$. As a consequence of the deceptively innocuous axiom of choice (which says that, given any collection C of nonempty sets, there exists a rule for selecting a unique point from each set in C), one can show that many finitely additive probabilities consistent with these requirements exist. However, one cannot be certain what the probability of getting an odd number is, because that set is neither finite nor its complement finite, nor can it be expressed as a finite disjoint union of sets whose probability is already defined.

It is a basic problem, and by no means a simple one, to show that the intuitive notion of choosing a number at random from [0, 1], as described above, is consistent with the preceding definitions. Since the probability of an interval is to be its length, the class of events M must contain all intervals; but in order to be a σ-field it must contain other sets, many of which are difficult to describe in an elementary way. One example is the event in (14), which must belong to M in order that one can talk about its probability. Also, although it seems clear that the length of a finite disjoint union of intervals is just the sum of their lengths, a rather subtle argument is required to show that length has the property of countable additivity. A basic theorem says that there is a suitable σ-field containing all the intervals and a unique probability defined on this σ-field for which the probability of an interval is its length.

Lebesgue-measurable sets

The σ-field is called the class of Lebesgue-measurable sets, and the probability is called the Lebesgue measure, after the French mathematician and principal architect of measure theory, Henri-Léon Lebesgue.

In general, a σ-field need not be all subsets of the sample space S. The question of whether all subsets of [0, 1] are Lebesgue-measurable turns out to be a difficult problem that is intimately connected with the foundations of mathematics and in particular with the axiom of choice.

For random variables having a continuum of possible values, the function that plays the same role as the probability distribution of a discrete random variable is called a "probability density function." If the random variable is denoted by X, its probability density function f has the property that

$$P(a<X\le b)=\int_a^b f(x)\,dx$$

for every interval $(a, b]$; *i.e.*, the probability that X falls in $(a, b]$ is the area under the graph of f between a and b (see Figure 1). For example, if X denotes the outcome of selecting a number at random from the interval $[r, s]$, the probability density function of X is given by $f(x)=1/(s-r)$ for $r<x<s$ and $f(x)=0$ for $x<r$ or $x>s$. The function $F(x)$ defined by $F(x)=P(X\le x)$ is called the "distribution function" or "cumulative distribution function" of X. If X has a probability density function $f(x)$, the relation between f and F is $F'(x)=f(x)$ or equivalently $F(x)=\int_{-\infty}^{x} f(t)\,dt$. The distribution function F of a discrete random variable should not be confused with its probability distribution f. In this case the relation between F and f is $F(x)=\sum_{x_i\le x} f(x_i)$.

If a random variable X has a probability density function $f(x)$, its "expectation" can be defined by

$$E(X)=\int_{-\infty}^{\infty} xf(x)\,dx, \tag{15}$$

provided that this integral is convergent. It turns out to be simpler, however, not only to use Lebesgue's theory of measure to define probabilities but also to use his theory of integration to define expectation. Accordingly, for any random variable X, $E(X)$ is defined to be the Lebesgue integral of X with respect to the probability measure P, provided that the integral exists. In this way it is possible to provide a unified theory in which all random variables, both discrete and continuous, can be treated simultaneously. In order to follow this path, it is necessary to restrict the class of those functions X defined on S that are to be called random variables, just as it was necessary to restrict the class of subsets of S that are called events. The appropriate restriction is that a random variable must be a measurable function. The definition is taken over directly from the Lebesgue theory of integration and will not be discussed here. It can be shown that, whenever X has a probability density function, its expectation (provided it exists) is given by (15), which remains a useful formula for calculating $E(X)$.

A unified theory

Some important probability density functions are the following:

(i) Normal: $f(x)=(2\pi\sigma^2)^{-1/2}\exp\left[\dfrac{-(x-\mu)^2}{2\sigma^2}\right]$

$(-\infty<x<+\infty)$; $E(X)=\mu$, $\mathrm{Var}(X)=\sigma^2$.

(ii) Exponential: $f(x)=\mu\exp(-\mu x)$ $(0\le x<+\infty)$,

$f(x)=0$ $(x<0)$; $E(X)=\dfrac{1}{\mu}$.

(iii) Cauchy: $f(x)=\dfrac{1}{[\pi(1+x^2)]}$ $(-\infty<x<+\infty)$;

$E(X)$ does not exist.

The cumulative distribution function of the normal distribution with mean 0 and variance 1 has already appeared as the function G defined following equation (12). The law of large numbers and the central limit theorem continue to hold for random variables on infinite sample spaces. A useful interpretation of the central limit theorem stated formally in (12) is as follows: The probability that the average (or sum) of a large number of independent, identically distributed random variables with finite variance falls in an interval $(c_1, c_2]$ equals approximately the area between c_1 and c_2 underneath the graph of a normal density function chosen to have the same expectation and variance as the given average (or sum). Figure 2 illustrates the normal approximation to the binomial distribution with $n=10$ and $p=1/2$.

The exponential distribution arises naturally in the study of the Poisson distribution introduced in (13). If T_k denotes the time interval between the emission of the k-1st and kth particle, then $T_1, T_2, \ldots$ are independent random variables having an exponential dis-

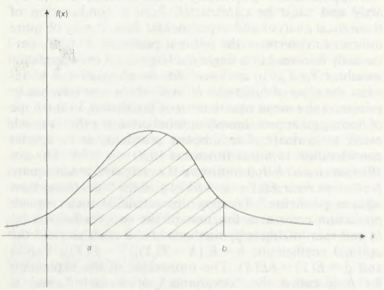


Figure 1: Probability density function.

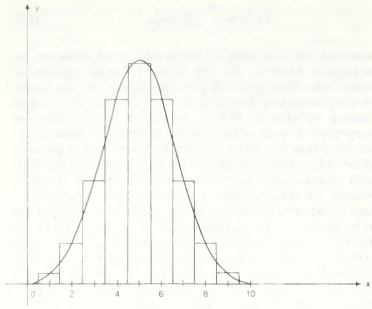


Figure 2: Normal approximation to the binomial distribution.

tribution with parameter μ. This is obvious for T_1 from the observation that $\{T_1 > t\} = \{N(t) = 0\}$. Hence, $P\{T_1 \le t\} = 1 - P\{N(t) = 0\} = 1 - \exp(-\mu t)$, and by differentiation one obtains the exponential density function.

Properties of the Cauchy distribution

The Cauchy distribution does not have a mean value or a variance, because the integral (15) does not converge. As a result it has a number of unusual properties. For example, if $X_1, X_2, \ldots, X_n$ are independent random variables having a Cauchy distribution, the average $(X_1 + \cdots + X_n)/n$ also has a Cauchy distribution. The variability of the average is exactly the same as that of a single observation. Another random variable that does not have an expectation is the waiting time until the number of heads first equals the number of tails in tossing a fair coin.

CONDITIONAL EXPECTATION AND LEAST SQUARES PREDICTION

An important problem of probability theory is to predict the value of a future observation Y given knowledge of a related observation X (or, more generally, given several related observations $X_1, X_2, \ldots$). Examples are to predict the future course of the national economy or the path of a rocket, given its present state.

Prediction is often just one aspect of a "control" problem. For example, in guiding a rocket, measurements of the rocket's location, velocity, and so on are made almost continuously; at each reading, the rocket's future course is predicted, and a control is then used to correct its future course. The same ideas are used to steer automatically large tankers transporting crude oil, for which even slight gains in efficiency result in large financial savings.

Given X, a predictor of Y is just a function $H(X)$. The problem of "least squares prediction" of Y given the observation X is to find that function $H(X)$ that is closest to Y in the sense that the mean square error of prediction, $E\{[Y - H(X)]^2\}$, is minimized. The solution is the conditional expectation $H(X) = E(Y|X)$.

In applications a probability model is rarely known exactly and must be constructed from a combination of theoretical analysis and experimental data. It may be quite difficult to determine the optimal predictor, $E(Y|X)$, particularly if instead of a single X a large number of predictor variables $X_1, X_2, \ldots$ are involved. An alternative is to restrict the class of functions H over which one searches to minimize the mean square error of prediction, in the hope of finding an approximately optimal predictor that is much easier to evaluate. The simplest possibility is to restrict consideration to linear functions $H(X) = a + bX$. The coefficients a and b that minimize the restricted mean square prediction error $E\{(Y - a - bX)^2\}$ give the "best linear least squares predictor." Treating this restricted mean square prediction error as a function of the two coefficients (a, b) and minimizing it by methods of the calculus yield the optimal coefficients: $\hat{b} = E\{[X - E(X)][Y - E(Y)]\}/\mathrm{Var}(X)$ and $\hat{a} = E(Y) - \hat{b}E(X)$. The numerator of the expression for $\hat{b}$ is called the "covariance" of X and Y and is denoted $\mathrm{Cov}(X, Y)$. Let $\hat{Y} = \hat{a} + \hat{b}X$ denote the optimal linear predictor. The mean square error of prediction is $E\{(Y - \hat{Y})^2\} = \mathrm{Var}(Y) - [\mathrm{Cov}(X, Y)]^2/\mathrm{Var}(X)$.

If X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$, the optimal predictor is just $E(Y)$, and the mean square error of prediction is $\mathrm{Var}(Y)$. Hence, $|\mathrm{Cov}(X, Y)|$ is a measure of the value X has in predicting Y. In the extreme case that $[\mathrm{Cov}(X, Y)]^2 = \mathrm{Var}(X)\mathrm{Var}(Y)$, Y is a linear function of X and the optimal linear predictor gives error-free prediction.

There is one important case in which the optimal mean square predictor actually is the same as the optimal linear predictor. If X and Y are jointly normally distributed, the conditional expectation of Y given X is just a linear function of X, and hence the optimal predictor and the optimal linear predictor are the same. The form of the bivariate normal distribution as well as expressions for the coefficients $\hat{a}$ and $\hat{b}$ and for the minimum mean square error of prediction were discovered by the English eugenicist Sir Francis Galton in his studies of the transmission of inheritable characteristics from one generation to the next. They form the foundation of the statistical technique of linear regression.

THE POISSON PROCESS AND THE BROWNIAN MOTION PROCESS

Stochastic processes

The theory of stochastic processes attempts to build probability models for phenomena that evolve over time. A primitive example appearing earlier in this article is the problem of gambler's ruin.

An important stochastic process described implicitly in the discussion of the Poisson approximation to the binomial distribution is the Poisson process. Modeling the emission of radioactive particles by an infinitely large number of tosses of a coin having infinitesimally small probability for heads on each toss led to the conclusion that the number of particles $N(t)$ emitted in the time interval $[0, t]$ has the Poisson distribution given in (13) with expectation μt. The primary concern of the theory of stochastic processes is not this marginal distribution of $N(t)$ at a particular time but rather the evolution of $N(t)$ over time. Two properties of the Poisson process that make it attractive to deal with theoretically are: (i) The times between emission of particles are independent and exponentially distributed with expected value $1/\mu$. (ii) Given that $N(t) = n$, the times at which the n particles are emitted have the same joint distribution as n points distributed independently and uniformly on the interval $[0, t]$.

As a consequence of property (i) a picture of the function $N(t)$ is very easily constructed. Originally $N(0) = 0$. At an exponentially distributed time T_1, the function $N(t)$ jumps from 0 to 1. It remains at 1 another exponentially distributed random time, T_2, which is independent of T_1, and at time $T_1 + T_2$ it jumps from 1 to 2, and so on.

Examples of other phenomena for which the Poisson process often serves as a mathematical model are the number of customers arriving at a counter and requesting service, the number of claims against an insurance company, or the number of malfunctions in a computer system. The importance of the Poisson process consists in (a) its simplicity as a test case for which the mathematical theory, and hence the implications, are more easily understood than for more realistic models and (b) its use as a building block in models of complex systems.

The most important stochastic process is the Brownian motion or Wiener process. It was first discussed by Louis Bachelier (1900), who was interested in modeling fluctuations in prices in financial markets, and by Albert Einstein (1905), who gave a mathematical model for the irregular motion of colloidal particles first observed by the Scottish botanist Robert Brown in 1827. The first mathematically rigorous treatment of this model was given by Wiener (1923). Einstein's results led to an early, dramatic confirmation of the molecular theory of matter in the French physicist Jean Perrin's experiments to determine Avogadro's number, for which Perrin was awarded a Nobel Prize in 1926. Today somewhat different models for physical Brownian motion are deemed more appropriate than Einstein's, but the original mathematical model continues to play a central role in the theory and application of stochastic processes.

Random walk

Let $B(t)$ denote the displacement (in one dimension for simplicity) of a colloidally suspended particle, which is buffeted by the numerous much smaller molecules of the medium in which it is suspended. This displacement will be obtained as a limit of a "random walk" occurring in discrete time as the number of steps becomes infinitely large and the size of each individual step infinitesimally small. Assume that at times $k\delta$, $k = 1, 2, \ldots$, the colloidal particle is displaced a distance hX_k, where $X_1, X_2, \ldots$ are $+1$ or -1 according as the outcomes of tossing a fair coin are heads or tails. By time t the particle has taken m steps, where m is the largest integer $\leq t/\delta$, and its displacement from its original position is $B_m(t) = h(X_1 + \cdots + X_m)$. The expected value of $B_m(t)$ is 0 and its variance is h^2m, or approximately h^2t/δ. Now suppose that $\delta \to 0$, and at the same time $h \to 0$ in such a way that the variance of $B_m(1)$ converges to some positive constant, σ^2. This means that m becomes infinitely large, and h is approximately $\sigma(t/m)^{1/2}$. It follows from the central limit theorem (12) that $\lim P[B_m(t) \leq x) = G(x/\sigma t^{1/2})$, where $G(x)$ is the standard normal cumulative distribution function defined just below equation (12). The Brownian motion process $B(t)$ can be defined to be the limit in a certain technical sense of the $B_m(t)$ as $\delta \to 0$ and $h \to 0$ with $h^2/\delta \to \sigma^2$.

The process $B(t)$ has many other properties, which in principle are all inherited from the approximating random walk $B_m(t)$. For example, if (s_1, t_1) and (s_2, t_2) are disjoint intervals, the increments $B(t_1) - B(s_1)$ and $B(t_2) - B(s_2)$ are independent random variables that are normally distributed with expectation 0 and variances equal to $\sigma^2(t_1 - s_1)$ and $\sigma^2(t_2 - s_2)$, respectively.

Einstein took a different approach and derived various properties of the process $B(t)$ by showing that its probability density function, $g(x, t)$, satisfies the diffusion equation $\partial g/\partial t = D\partial^2 g/\partial x^2$, where $D = \sigma^2/2$. The important implication of Einstein's theory for subsequent experimental research was that he identified the diffusion constant D in terms of certain measurable properties of the particle (its radius) and of the medium (its viscosity and temperature), which allowed one to make predictions and hence to confirm or reject the hypothesized existence of the unseen molecules that were assumed to be the cause of the irregular Brownian motion. Because of the beautiful blend of mathematical and physical reasoning involved, a brief summary of the successor to Einstein's model is given below.

Unlike the Poisson process, it is impossible to "draw" a picture of the path of a particle undergoing mathematical Brownian motion. Wiener (1923) showed that the functions $B(t)$ are continuous, as one expects, but nowhere differentiable. Thus, a particle undergoing mathematical Brownian motion does not have a well-defined velocity, and the curve $y = B(t)$ does not have a well-defined tangent at any value of t. To see why this might be so, recall that the derivative of $B(t)$, if it exists, is the limit as $h \to 0$ of the ratio $[B(t + h) - B(t)]/h$. Since $B(t + h) - B(t)$ is normally distributed with mean 0 and standard deviation $h^{1/2}\sigma$, in very rough terms $B(t + h) - B(t)$ can be expected to equal some multiple (positive or negative) of $h^{1/2}$. But the limit as $h \to 0$ of $h^{1/2}/h = 1/h^{1/2}$ is infinite. A related fact that illustrates the extreme irregularity of $B(t)$ is that in every interval of time, no matter how small, a particle undergoing mathematical Brownian motion travels an infinite distance. Although these properties contradict the commonsense idea of a function—and indeed it is quite difficult to write down explicitly a single example of a continuous, nowhere-differentiable function—they turn out to be typical of a large class of stochastic processes, called "diffusion processes," of which Brownian motion is the most prominent member. Especially notable contributions to the mathematical theory of Brownian motion and diffusion processes were made by Paul Lévy and William Feller during the years 1930–1960.

Diffusion processes

A more sophisticated description of physical Brownian motion can be built on a simple application of Newton's second law: $F = ma$. Let $V(t)$ denote the velocity of a colloidal particle of mass m. It is assumed that

$$mdV(t) = -fV(t)dt + dA(t). \tag{18}$$

The quantity f retarding the movement of the particle is due to friction caused by the surrounding medium. The term $dA(t)$ is the contribution of the very frequent collisions of the particle with unseen molecules of the medium. It is assumed that f can be determined by classical fluid mechanics, in which the molecules making up the surrounding medium are so many and so small that the medium can be considered smooth and homogeneous. Then by Stokes's law, for a spherical particle in a gas, $f = 6\pi a\eta$, where a is the radius of the particle and η the coefficient of viscosity of the medium. Hypotheses concerning $A(t)$ are less specific, because the molecules making up the surrounding medium cannot be observed directly. For example, it is assumed that, for $t \neq s$, the infinitesimal random increments $dA(t) = A(t + dt) - A(t)$ and $A(s + ds) - A(s)$ caused by collisions of the particle with molecules of the surrounding medium are independent random variables having distributions with mean 0 and unknown variances $\sigma^2\,dt$ and $\sigma^2\,ds$ and that $dA(t)$ is independent of $dV(s)$ for $s < t$.

The differential equation (18) has the solution

$$V(t) = V(0)\exp(-\beta t) + m^{-1}\int_0^t \exp[-\beta(t - s)]dA(s), \tag{19}$$

where $\beta = f/m$. From (19) and the assumed properties of $A(t)$, it follows that $E[V^2(t)] \to \sigma^2/(2mf)$ as $t \to \infty$. Now assume that, in accordance with the principle of equipartition of energy, the steady-state average kinetic energy of the particle, $m \lim_{t\to\infty} E[V^2(t)]/2$, equals the average kinetic energy of the molecules of the medium. According to the kinetic theory of an ideal gas, this is $RT/2N$, where R is the ideal gas constant, T is the temperature of the gas in kelvins, and N is Avogadro's number, the number of molecules in one gram molecular weight of the gas. It follows that the unknown value of σ^2 can be determined: $\sigma^2 = 2RTf/N$.

If one also assumes that the functions $V(t)$ are continuous, which is certainly reasonable from physical considerations, it follows by mathematical analysis that $A(t)$ is a Brownian motion process as defined above. This conclusion poses questions about the meaning of the initial equation (18), because for mathematical Brownian motion the term $dA(t)$ does not exist in the usual sense of a derivative. Some additional mathematical analysis shows that the stochastic differential equation (18) and its solution (19) have a precise mathematical interpretation. The process $V(t)$ is called the Ornstein-Uhlenbeck process, after the physicists Leonard Salomon Ornstein and George Eugene Uhlenbeck. The logical outgrowth of these attempts to differentiate and integrate with respect to a Brownian motion process is the Ito (named for the Japanese mathematician Itô Kiyosi) stochastic calculus, which plays an important role in the modern theory of stochastic processes.

The Ito stochastic calculus

The displacement at time t of the particle whose velocity is given by (19) is

$$X(t) - X(0) = \int_0^t V(u)du = \beta^{-1}V(0)[1 - \exp(-\beta t)] + f^{-1}A(t) - f^{-1}\int_0^t \exp[-\beta(t - u)]dA(u).$$

For t large compared to β the first and third terms in this expression are small compared to the second. Hence, $X(t) - X(0)$ is approximately equal to $A(t)/f$, and the mean square displacement, $E[[X(t) - X(0)]^2]$, is approximately $\sigma^2/f^2 = RT/(3\pi a\eta N)$. These final conclusions are consistent with Einstein's model, although here they arise as an approximation to the model obtained from (19). Since it is primarily the conclusions that have observational consequences, there are essentially no new experimental implications. However, the analysis arising directly out of Newton's second law, which yields a process having a well-defined velocity at each point, seems more satisfactory theoretically than Einstein's original model.

STOCHASTIC PROCESSES

A stochastic process is a family of random variables $X(t)$ indexed by a parameter t, which usually takes values in

the discrete set $T = \{0, 1, 2, \dots\}$ or the continuous set $T = [0, +\infty)$. In many cases t represents time, and $X(t)$ is a random variable observed at time t. Examples are the Poisson process, the Brownian motion process, and the Ornstein-Uhlenbeck process described in the preceding section. Considered as a totality, the family of random variables $\{X(t), t \in T\}$ constitutes a "random function."

The mathematical theory of stochastic processes attempts to define classes of processes for which a unified theory can be developed. The most important classes are stationary processes and Markov processes. A stochastic process is called stationary if, for all n, $t_1 < t_2 < \cdots < t_n$, and $h > 0$, the joint distribution of $X(t_1 + h), \dots, X(t_n + h)$ does not depend on h. This means that in effect there is no origin on the time axis; the stochastic behaviour of a stationary process is the same no matter when the process is observed. A sequence of independent identically distributed random variables is an example of a stationary process. A rather different example is defined as follows: $U(0)$ is uniformly distributed on $[0, 1]$; for each $t = 1, 2, \dots$, $U(t) = 2U(t-1)$ if $U(t-1) \le \frac{1}{2}$ and $U(t) = 2U(t-1) - 1$ if $U(t-1) > \frac{1}{2}$. The marginal distributions of $U(t)$, $t = 0, 1, \dots$ are uniformly distributed on $[0, 1]$, but, in contrast to the case of independent identically distributed random variables, the entire sequence can be predicted from knowledge of $U(0)$. A third example of a stationary process is

$$X(t) = \sum_k c_k [Y_k \cos(\theta_k t) + Z_k \sin(\theta_k t)],$$

where the Ys and Zs are independent normally distributed random variables with mean 0 and unit variance, and the cs and θs are constants. Processes of this kind can be useful in modeling seasonal or approximately periodic phenomena.

The ergodic theorem

A remarkable generalization of the strong law of large numbers is the "ergodic theorem": if $X(t)$, $t = 0, 1, \dots$ or $0 \le t < \infty$ is a stationary process such that $E[X(0)]$ is finite, then with probability 1 the average

$$s^{-1} \sum_{t=0}^{t=s-1} X(t), \text{ if } t \text{ is discrete, or } s^{-1} \int_0^s X(t)dt,$$

if t is continuous, converges to a limit as $s \to \infty$. In the special case that t is discrete and the Xs are independent and identically distributed, the strong law of large numbers is also applicable and shows that the limit must equal $E(X(0))$. However, the example that $X(0)$ is an arbitrary random variable and $X(t) \equiv X(0)$ for all $t > 0$ shows that this cannot be true in general. The limit does equal $E(X(0))$ under an additional rather technical assumption to the effect that there is no subset of the state space, having probability strictly between 0 and 1, in which the process can get stuck and never escape. This assumption is not fulfilled by the example $X(t) \equiv X(0)$ for all t, which gets stuck immediately at its initial value. It is satisfied by the sequence $U(t)$ defined above, so by the ergodic theorem the average of these variables converges to $\frac{1}{2}$ with probability 1. The ergodic theorem was first conjectured by the American chemist J. Willard Gibbs in the early 1900s in the context of statistical mechanics and was proved in a corrected, abstract formulation by the American mathematician George David Birkhoff in 1931.

A stochastic process is called Markovian (after the Russian mathematician Andrey Andreyevich Markov) if at any time t the conditional probability of an arbitrary future event given the entire past of the process—*i.e.*, given $X(s)$ for all $s \le t$—equals the conditional probability of that future event given only $X(t)$. Thus, in order to make a probabilistic statement about the future behaviour of a Markov process, it is no more helpful to know the entire history of the process than it is to know only its current state. The conditional distribution of $X(t + h)$ given $X(t)$ is called the "transition probability" of the process. If this conditional distribution does not depend on t, the process is said to have "stationary" transition probabilities. A Markov process with stationary transition probabilities may or may not be a stationary process in the sense of the preceding paragraph. If $Y_1, Y_2, \dots$ are independent random variables and $X(t) = Y_1 + \cdots + Y_t$, the stochastic process $X(t)$ is a Markov process. Given $X(t) = x$, the conditional probability that $X(t + h)$ belongs to an interval (a, b) is just the probability that $Y_{t+1} + \cdots + Y_{t+h}$ belongs to the translated interval $(a - x, b - x)$; and because of independence this conditional probability would be the same if the values of $X(1), \dots, X(t-1)$ were also given. If the Ys are identically distributed as well as independent, this transition probability does not depend on t, and then $X(t)$ is a Markov process with stationary transition probabilities. Sometimes $X(t)$ is called a "random walk," but this terminology is not completely standard. Since both the Poisson process and Brownian motion are created from random walks by simple limiting processes, they too are Markov processes with stationary transition probabilities. The Ornstein-Uhlenbeck process defined as the solution (19) to the stochastic differential equation (18) is also a Markov process with stationary transition probabilities.

The Ornstein-Uhlenbeck process and many other Markov processes with stationary transition probabilities behave like stationary processes as $t \to \infty$. Roughly speaking, the conditional distribution of $X(t)$ given $X(0) = x$ converges as $t \to \infty$ to a distribution, called the "stationary distribution," that does not depend on the starting value $X(0) = x$. Moreover, with probability 1, the proportion of time the process spends in any subset of its state space converges to the stationary probability of that set; and, if $X(0)$ is given the stationary distribution to begin with, the process becomes a stationary process. The Ornstein-Uhlenbeck process defined in (19) is stationary if $V(0)$ has a normal distribution with mean 0 and variance $\sigma^2/(2mf)$.

At another extreme are "absorbing" processes. An example is the Markov process describing Peter's fortune during the game of gambler's ruin. The process is absorbed whenever either Peter or Paul is ruined. Questions of interest involve the probability of being absorbed in one state rather than another and the distribution of the time until absorption occurs. Some additional examples of stochastic processes follow.

The Ehrenfest model of diffusion

The Ehrenfest model of diffusion (named after the Austrian-Dutch physicist Paul Ehrenfest) was proposed in the early 1900s in order to illuminate the statistical interpretation of the second law of thermodynamics, that the entropy of a closed system can only increase. Suppose N molecules of a gas are in a rectangular container divided into two equal parts by a permeable membrane. The state of the system at time t is $X(t)$, the number of molecules on the left-hand side of the membrane. At each time $t = 1, 2, \dots$ a molecule is chosen at random (*i.e.*, each molecule has probability $1/N$ to be chosen) and is moved from its present location to the other side of the membrane. Hence, the system evolves according to the transition probability $p(i, j) = P\{X(t+1) = j \mid X(t) = i\}$, where

$$p(i, i+1) = 1 - \frac{i}{N}, \qquad p(i, i-1) = \frac{i}{N},$$

$$p(i, j) = 0 \quad \text{for } j \ne i+1, i-1.$$

The long run behaviour of the Ehrenfest process can be inferred from general theorems about Markov processes in discrete time with discrete state space and stationary transition probabilities. Let $T(j)$ denote the first time $t \ge 1$ such that $X(t) = j$ and set $T(j) = \infty$ if $X(t) \ne j$ for all t. Assume that for all states i and j it is possible for the process to go from i to j in some number of steps—*i.e.*, $P\{T(j) < \infty \mid X(0) = i\} > 0$. If the equations

$$Q(j) = \sum_i Q(i) p(i, j) \tag{20}$$

have a solution $Q(j)$ that is a probability distribution—*i.e.*, $Q(j) \ge 0$, and $\sum Q(j) = 1$—then that solution is unique and is the stationary distribution of the process. Moreover, $Q(j) = 1/E\{T(j) \mid X(0) = j\}$; and, for any initial state j, the proportion of time t that $X(t) = i$ converges with probability 1 to $Q(i)$.

For the special case of the Ehrenfest process, assume that N is large and $X(0) = 0$. According to the deterministic prediction of the second law of thermodynamics, the entropy of this system can only increase, which means that $X(t)$

will steadily increase until half the molecules are on each side of the membrane. Indeed, according to the stochastic model described above, there is overwhelming probability that $X(t)$ does increase initially. However, because of random fluctuations, the system occasionally moves from configurations having large entropy to those of smaller entropy and eventually even returns to its starting state, in defiance of the second law of thermodynamics.

The accepted resolution of this contradiction is that the length of time such a system must operate in order that an observable decrease of entropy may occur is so enormously long that a decrease could never be verified experimentally. To consider only the most extreme case, let T denote the first time $t \geq 1$ at which $X(t) = 0$—*i.e.*, the time of first return to the starting configuration having all molecules on the right-hand side of the membrane. It can be verified by substitution in (20) that the stationary distribution of the Ehrenfest model is the binomial distribution $Q(j) = \binom{N}{j} 2^{-N}$, and hence $E(T) = 2^N$. For example, if N is only 100 and transitions occur at the rate of 10^6 per second, $E(T)$ is of the order of 10^{15} years. Hence, on the macroscopic scale, on which experimental measurements can be made, the second law of thermodynamics holds.

The symmetric random walk

A Markov process that behaves in quite different and surprising ways is the symmetric random walk. A particle occupies a point with integer coordinates in d-dimensional Euclidean space. At each time $t = 1, 2, \ldots$, it moves from its present location to one of its $2d$ nearest neighbours with equal probabilities $1/(2d)$, independently of its past moves. For $d = 1$ this corresponds to moving a step to the right or left according to the outcome of tossing a fair coin. It may be shown that for $d = 1$ or 2 the particle returns with probability 1 to its initial position and hence to every possible position infinitely many times, if the random walk continues indefinitely. In three or more dimensions, at any time t the number of possible steps that increase the distance of the particle from the origin is much larger than the number decreasing the distance, with the result that the particle eventually moves away from the origin and never returns. Even in one or two dimensions, although the particle eventually returns to its initial position, the expected waiting time until it returns is infinite, there is no stationary distribution, and the proportion of time the particle spends in any state converges to 0!

The simplest service system is a single-server queue, where customers arrive, wait their turn, are served by a single server, and depart. Related stochastic processes are the waiting time of the nth customer and the number of customers in the queue at time t. For example, suppose that customers arrive at times $0 = T_0 < T_1 < T_2 < \cdots$ and wait in a queue until their turn. Let V_n denote the service time required by the nth customer, $n = 0, 1, 2, \ldots$, and set $U_n = T_n - T_{n-1}$. The waiting time, W_n, of the nth customer satisfies the relation $W_0 = 0$, and, for $n \geq 1$, $W_n = \max(0, W_{n-1} + V_{n-1} - U_n)$. To see this, observe that the nth customer must wait for the same length of time as the $n-1$st customer plus the service time of the $n-1$st customer minus the time between the arrival of the $n-1$st and nth customer, during which the $n-1$st customer is already waiting but the nth customer is not. An exception occurs if this quantity is negative, and then the waiting time of the nth customer is 0. Various assumptions can be made about the input and service mechanisms. One possibility is that customers arrive according to a Poisson process and their service times are independent, identically distributed random variables that are also independent of the arrival process. Then, in terms of $Y_n = V_{n-1} - U_n$, which are independent, identically distributed random variables, the recursive relation defining W_n becomes $W_n = \max(0, W_{n-1} + Y_n)$. This process is a Markov process. It is often called a random walk with reflecting barrier at 0, because it behaves like a random walk whenever it is positive and is pushed up to be equal to 0 whenever it tries to become negative. Quantities of interest are the mean and variance of the waiting time of the nth customer and, since these are very difficult to determine exactly, the mean and variance of the stationary distribution. More realistic queuing models try to accommodate systems with several servers and different classes of customers, who are served according to certain priorities. In most cases it is impossible to give a mathematical analysis of the system, which must be simulated on a computer in order to obtain numerical results. The insights gained from theoretical analysis of simple cases can be helpful in performing these simulations. Queuing theory had its origins in attempts to understand traffic in telephone systems. Present-day research is stimulated, among other things, by problems associated with multiple-user computer systems.

Other queuing models

Reflecting barriers arise in other problems as well. For example, if $B(t)$ denotes Brownian motion, then $X(t) = B(t) + ct$ is called Brownian motion with drift c. This model is appropriate for Brownian motion of a particle under the influence of a constant force field such as gravity. One can add a reflecting barrier at 0 to account for reflections of the Brownian particle off the bottom of its container. The result is a model for sedimentation, which for $c < 0$ in the steady state as $t \to \infty$ gives a statistical derivation of the law of pressure as a function of depth in an isothermal atmosphere. Just as ordinary Brownian motion can be obtained as the limit of a rescaled random walk as the number of steps becomes very large and the size of individual steps small, Brownian motion with a reflecting barrier at 0 can be obtained as the limit of a rescaled random walk with reflection at 0. In this way, Brownian motion with a reflecting barrier plays a role in the analysis of queuing systems. In fact, in modern probability theory one of the most important uses of Brownian motion and other diffusion processes is as approximations to more complicated stochastic processes. The exact mathematical description of these approximations gives remarkable generalizations of the central limit theorem from sequences of random variables to sequences of random functions.

The ruin problem of insurance risk theory is closely related to the problem of gambler's ruin described earlier and, rather surprisingly, to the single-server queue as well. Suppose the amount of capital at time t in one portfolio of an insurance company is denoted by $X(t)$. Initially $X(0) = x > 0$. During each unit of time, the portfolio receives an amount $c > 0$ in premiums. At random times claims are made against the insurance company, which must pay the amount $V_n > 0$ to settle the nth claim. If $N(t)$ denotes the number of claims made in time t, then $X(t) = x + ct - \sum_1^{N(t)} V_n$, provided that this quantity has been positive at all earlier times $s < t$. At the first time $X(t)$ becomes negative, however, the portfolio is ruined. A principal problem of insurance risk theory is to find the probability of ultimate ruin. If one imagines that the problem of gambler's ruin is modified so that Peter's opponent has an infinite amount of capital and can never be ruined, then the probability that Peter is ultimately ruined is similar to the ruin probability of insurance risk theory. In fact, with the artificial assumptions that (i) $c = 1$, (ii) time proceeds by discrete units, say $t = 1, 2, \ldots$, (iii) V_n is identically equal to 2 for all n, and (iv) at each time t a claim occurs with probability p or does not occur with probability q independently of what occurs at other times, then the process $X(t)$ is the same stochastic process as Peter's fortune, which is absorbed if it ever reaches the state 0. The probability of Peter's ultimate ruin against an infinitely rich adversary is easily obtained by taking the limit of equation (6) as $m \to \infty$. The answer is $(q/p)^x$ if $p > q$—*i.e.*, the game is favourable to Peter—and 1 if $p \leq q$. More interesting assumptions for the insurance risk problem are that the number of claims $N(t)$ is a Poisson process and the sizes of the claims $V_1, V_2, \ldots$ are independent, identically distributed positive random variables. Rather surprisingly, under these assumptions the probability of ultimate ruin as a function of the initial fortune x is exactly the same as the stationary probability that the waiting time in the single-server queue with Poisson input exceeds x. Unfortunately, neither problem is easy to solve exactly, although there is a very good approximate solution originally derived by the Swedish mathematician Harald Cramér.

As a final example, it seems appropriate to mention one of the dominant ideas of modern probability theory, which at the same time springs directly from the relation of probability to games of chance. Suppose that

Martingales

$X_1, X_2, \ldots$ is any stochastic process and, for each $n = 0, 1, \ldots, f_n = f_n(X_1, \ldots, X_n)$ is a (Borel-measurable) function of the indicated observations. The new stochastic process f_n is called a "martingale" if $E(f_n|X_1, \ldots, X_{n-1}) = f_{n-1}$ for every value of $n > 0$ and all values of $X_1, \ldots, X_{n-1}$. If the sequence of Xs are outcomes in successive trials of a game of chance and f_n is the fortune of a gambler after the nth trial, then the martingale condition says that the game is absolutely fair in the sense that, no matter what the past history of the game, the gambler's conditional expected fortune after one more trial is exactly equal to his present fortune. For example, let $X_0 = x$, and for $n \geq 1$ let X_n equal 1 or -1 according as a coin having probability p of heads and $q = 1 - p$ of tails turns up heads or tails on the nth toss. Let $S_n = X_0 + \cdots + X_n$. Then $f_n = S_n - n(p - q)$ and $f_n = (q/p)^{S_n}$ are martingales. One of the basic results of martingale theory is that, if the gambler is free to quit the game at any time using any strategy whatever, provided only that this strategy does not foresee the future, then the game remains fair. This means that, if N denotes the stopping time at which the gambler's strategy tells him to quit the game, so that his final fortune is f_N, then

$$E(f_N|f_0) = f_0. \tag{21}$$

Strictly speaking, this result is not true without some additional conditions that must be verified for any particular application. To see how efficiently it works, consider once again the problem of gambler's ruin and let N be the first value of n such that $S_n = 0$ or m; *i.e.*, N denotes the random time at which ruin first occurs and the game ends. In the case $p = 1/2$, application of (21) to the martingale $f_n = S_n$, together with the observation that $f_N =$ either 0 or m, yields the equalities $x = f_0 = E(f_N|f_0 = x) = m[1 - Q(x)]$, which can be immediately solved to give the answer in (6). For $p \neq 1/2$, one uses the martingale $f_n = (q/p)^{S_n}$ and similar reasoning to obtain

$$\left(\frac{q}{p}\right)^x = E(f_N|f_0 = x) = 1Q(x) + \left(\frac{q}{p}\right)^m [1 - Q(x)],$$

from which the first equation in (6) easily follows. The expected duration of the game is obtained by a similar argument.

A particularly beautiful and important result is the martingale convergence theorem, which implies that a nonnegative martingale converges with probability 1 as $n \to \infty$. This means that, if a gambler's successive fortunes form a (nonnegative) martingale, they cannot continue to fluctuate indefinitely but must approach some limiting value.

Basic martingale theory and many of its applications were developed by the American mathematician Joseph Leo Doob during the 1940s and '50s following some earlier results due to Paul Lévy. Subsequently it has become one of the most powerful tools available to study stochastic processes.

BIBLIOGRAPHY. F.N. DAVID, *Games, Gods, and Gambling: The Origins and History of Probability and Statistical Ideas from the Earliest Times to the Newtonian Era* (1962), covers the early history of probability theory. STEPHEN M. STIGLER, *The History of Statistics: The Measurement of Uncertainty Before 1900* (1986), describes the attempts of early statisticians to use probability theory and to understand its significance in scientific problems. W. FELLER, *An Introduction to Probability Theory and Its Applications*, vol. 1, 3rd ed. (1967), and vol. 2, 2nd ed. (1971), contains a masterly exposition of discrete probability theory in vol. 1, while vol. 2 requires a more sophisticated mathematical background. A.N. KOLMOGOROV, *Foundations of the Theory of Probability*, 2nd ed. (1956; originally published in German, 1933), is eminently readable, although it requires knowledge of measure theory. JOSEPH L. DOOB, *Stochastic Processes* (1953, reissued 1964), is a comprehensive treatment of stochastic processes, including much of Doob's original development of martingale theory. NELSON WAX, *Selected Papers on Noise and Stochastic Processes* (1954), collects six classical papers on probability theory, especially in its relation to the physical sciences. M. LOÈVE, *Probability Theory*, 4th ed., 2 vol. (1977–78), is an encyclopaedic reference book covering discrete probability theory and developing measure theory, the laws of large numbers, the central limit theorem, and stochastic processes. See also WALTER LEDERMANN (ed.), *Handbook of Applicable Mathematics*, vol. 6, *Probability*, ed. by EMLYN LLOYD (1980), a practical text written for the educated lay reader.

(D.O.S.)

Procedural Law

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

Because procedural law is only a means for enforcing substantive rules, there are different kinds of procedural law, corresponding to the various kinds of substantive law. Criminal law, for example, is the branch of substantive law dealing with punishment for offenses against the public and has as its corollary criminal procedure, which indicates how the sanctions of criminal law must be applied. Substantive private law, which deals with the relations between private (that is, nongovernmental) persons, whether individuals or corporate bodies, has as its corollary the rules of civil procedure. Because the object of judicial proceedings is to arrive at the truth using the best available evidence, there must be procedural laws of evidence to govern the presentation of witnesses, documentation, and physical proof. The law of conflict of laws, in both its civil and criminal applications, provides methods for resolving problems that arise from the diversity of legal systems in the world.

This article deals with procedural laws as they apply to the Anglo-American common law and the continental European civil law. For discussion of other legal systems, such as the laws of Communist countries, see the *Macropædia* article LEGAL SYSTEMS, THE EVOLUTION OF MODERN WESTERN. Substantive laws are covered in such articles as CRIMINAL LAW; BUSINESS LAW; and CONSTITUTIONAL LAW. For treatment of administrative procedural law, see PUBLIC ADMINISTRATION. (P.E.H./Ma.E.O.)

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 551, 552, and 553.

The article is divided into the following sections:

Civil procedure 149
- Historical development 149
  - Roman law
  - Medieval European law
  - Common law
  - Civil-law codifications
  - Constitutional bases of civil procedure
- Civil-law procedure and common-law procedure 151
  - The impact of the jury
  - Convergence of civil- and common-law procedure
- Preliminaries to proceedings 152
  - Jurisdiction, competence, and venue
  - Parties
  - Provisional remedies
  - Commencement of action
  - The preparatory stage
- The trial or main hearing 155
  - The Anglo-American jury trial
  - The civil-law main hearing
  - Judgment and execution
- Appeals and other methods of review 158
  - Common-law appellate procedure
  - Civil-law appellate procedure

Criminal procedure 159
- Procedure before trial 159
  - The investigatory phase
  - The decision to prosecute
- Trial procedure 160
  - Criminal courts
  - Pretrial matters
  - Publicity of the trial
  - Presentation of evidence
  - Finding the verdict
  - Sentencing
- Post-conviction procedure 161
  - Common law
  - Civil law

The law of evidence 161
- The early law of evidence 161
  - Nonrational sources of evidence
  - Semirational sources of evidence
  - The influence of Roman-canonical law
- Comparative survey of modern principles 162
  - Oral proceedings
  - The burden of proof
  - Relevance and admissibility
  - The free evaluation of evidence
- Sources of proof 163
  - Witnesses
  - Party testimony
  - Expert evidence
  - Documentary evidence
  - Real evidence

Conflict of laws 166
- Diversity of laws 166
  - Diversity within countries
  - Diversity between countries
  - Rules on the conflict of laws
- Private international law 166
  - Jurisdiction
  - Proceedings
  - Foreign judgments
- International criminal law 169
  - Jurisdiction
  - Proceedings
  - Foreign judgments

Bibliography 170

Civil procedure

HISTORICAL DEVELOPMENT

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

The formulary system

The formulary system with its separation of fact-finding

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

Medieval European law. The Germanic tribes that conquered the Roman Empire in the 5th century carried their own procedure with them into the conquered territories.

Adaptation of Roman to Germanic procedure

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

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

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

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

Development of English royal courts

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

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

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

Movement toward reform

The procedure of the common-law courts and the existence of a separate procedure for equity matters were both taken over in the United States. In the 19th century there developed in both England and the United States movements to simplify procedural complexities. These involved several related approaches: (1) a reform in court organization, doing away with separate courts of equity and, to the extent they existed, with coordinated common-law courts of general jurisdiction and establishing a more rational system of appeals courts; (2) a reform of pleading, abandoning largely the need to plead a specific cause of action based on writs, and giving judges power to promulgate rules of procedure. In the United States these principles were first embodied in the widely followed New York Code of Civil Procedure of 1848. In the 20th century, however, the notion gained ground that legislation was too slow a means for the adoption of new procedural rules. This led to the Rules Enabling Act of 1934, which authorized the U.S. Supreme Court to adopt (subject to congressional veto) Rules of Civil Procedure for the federal district courts, though some matters, such as subject-matter jurisdiction, remained governed by acts of Congress. The federal Rules of Civil Procedure were later followed by Rules of Appellate Procedure and Rules

of Evidence. There were similar developments in many of the states and also in England and Wales.

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

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

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

During the 19th century, codifications of procedural law were enacted in other countries (Italy in 1865 and Germany in 1877). They usually retained large elements of the Roman-canonical or French procedure and were often cumbersome and slow. Austria departed from the Roman-canonical model in 1895 with the adoption of a new Code of Civil Procedure. The new code adopted comprehensively the principle of oral presentation: only matters presented orally in open court were important for a decision of the case; writings could have only a preparatory role; witnesses were no longer heard before a delegated judge who prepared a written record but by the court or judge that actually decided the case; finally, the parties were obligated to present their cases fully and truthfully, and the judge was directed to make certain that all relevant facts were stated. These notions were widely followed by other countries when they amended their codes of civil procedure.

Revisions in the 20th century

Changes made in French civil procedure beginning in 1958 were to some extent inspired by the Austrian model. Originally adopted in a series of individual decrees, they were consolidated in the new Code of Civil Procedure of 1975. Following earlier amendments to the 1877 German code that had strengthened the role of the judge, a statute adopted in 1976 in West Germany, called the simplification amendment, was designed to expedite proceedings further. A somewhat contrary trend occurred in Italy, where later amendments to the more progressive 1942 Code of Civil Procedure to some extent reemphasized written presentations. A step contrary to some modern European thinking was also taken by the Belgian Judicial Code of 1967 (effective 1969). It reduced the role of the judge and correspondingly increased that of the parties and their counsel. Even more atypical were developments in Japan. In 1890 that country adopted a Code of Civil Procedure, very largely modeled on the German Code of 1877. In 1926 the code was amended in order to expedite procedures. Austrian ideas about the role of the judge were heavily relied on. But after the defeat of Japan at the end of World War II, an attempt was made to introduce some of the features of the American civil trial, with its heavy reliance on the presentation of facts by the parties' attorneys and the correspondingly less significant role of the judge. For a variety of reasons, the attempt was not entirely successful. Present Japanese law blends a procedure largely based on the German model with some features of Anglo-American origin.

Constitutional bases of civil procedure. The U.S. Supreme Court holds that all procedural rules, whether found in statutes, rules of court, or case law, must be consistent with the mandates of the Constitution—in particular with the due process clauses of the Fifth and Fourteenth amendments. Thus, a defendant from one state or a foreign country cannot be required to defend a suit in another state unless the defendant has had enough contacts with that state not to offend "traditional notions of fairness and substantial justice." Likewise, "due process" implies that a party may not be deprived of substantial rights without having had an opportunity to present his side of the case. As a result of the adoption in many other countries of written constitutions with legally binding fundamental rights—and of the creation, after World War II, of special constitutional courts—constitutional rules granting a right to be heard, and, more generally, access to justice (often including access to legal aid) were created. These developments were reinforced by certain international agreements, in particular Article 6 of the 1950 European Convention for the Protection of Human Rights and Fundamental Freedoms.

CIVIL-LAW PROCEDURE AND COMMON-LAW PROCEDURE

It is sometimes said that the Anglo-American common-law procedure is adversarial, while the continental European civil-law procedure is inquisitorial. This means that, in the common law, a lawsuit is essentially the concern of the adversaries, that is, the parties and their lawyers. It is the lawyers who present the evidence, and, unless a procedural problem arises, the judge simply listens to the presentation. By contrast, in the civil law there is a greater emphasis on the judge as a guarantor of a just outcome of the case, regardless of the lawyers' abilities. To this end he often functions as an inquisitor, questioning the parties as to the factual matters of the case. In some countries, such as Germany, the judge is required to guide the proceedings—for instance, by suggesting to the parties that they direct their attention to a particular point of fact or law. These differences in procedure create problems when, for a lawsuit pending in a country of one system, it is necessary to obtain evidence located in a country of the other system. In such cases, "judicial assistance" must be given to the courts in one country by those in the other (see below *Conflict of laws*).

The roles of lawyers and judges

The impact of the jury. The differences between various aspects of civil- and common-law procedure have their origin in several factors, but the most important is the institution of the jury trial. A lay jury will be able to determine factual issues only if they are presented in a focused manner; hence the need, in common-law countries, for initial pleadings that serve to establish beforehand the factual matters in dispute. This is less necessary in civil-law procedure, where the case is handled over a series of hearings by professional judges. Furthermore, because a jury of 12 laypersons cannot be kept together for an indefinite period of time, proceedings must be conducted in a concentrated fashion. This gives the Anglo-American trial its peculiar dramatic character. Where the determination of factual issues is entrusted to a professional judge (who, presumably, will be available for a considerable time), the process can be extended over several shorter hearings. Because a disbanded jury cannot easily be reassembled, the evidence presented by parties must be available at the beginning of the trial. Consequently, there must be procedures (called "discovery") that enable the parties to obtain information they will need at the trial (see below *Discovery procedures*). Such procedures are much less important in civil-law procedure, where evidence that has come to light during the proceedings can be submitted at a subsequent hearing. Also, when factual matters are to be decided by a body of laypersons, the law must ensure that the jury will not be misled by evidence that is plausible on the surface but may be misleading. There is less need to guard against that danger whenever factual determinations are made by professional judges.

Finally, because the jury decides questions of fact while the judge decides only questions of law, in common-law procedure a clear distinction must be drawn from the beginning between factual and legal issues. In civil-law procedure, on the other hand, where the judges decide both questions of fact and questions of law, there is normally no need to make a sharp distinction between these two issues until a case reaches the highest level of civil courts, where only questions of law are open for review.

Convergence of civil- and common-law procedure. In spite of the distinctions between civil and common law, some trends toward a convergence of procedure exist. In

private-law matters, courts in civil-law countries do not initiate proceedings on their own motion; rather, they decide only claims brought forward by the parties and normally only on the basis of evidence proposed by them. Nor do judges in common-law countries always play merely the role of an impartial arbiter. In some cases, such as those involving the welfare of children, they often take a more active role in seeking out the facts.

Because a series of separate hearings makes a proceeding unduly long, there are procedural reforms in some civil-law countries that favour (but do not mandate) a single, well-prepared, main hearing at which the decision is reached. On the other hand, in England, where the civil jury trial originated, the jury has fallen into almost complete disuse in civil cases, except in suits of defamation.

Control over jury-trial proceedings

In the United States, trial by jury is a constitutional right, but steps have been taken toward more judicial control over proceedings. Thus, Rule 11 of the Federal Rules of Civil Procedure authorizes judges to penalize frivolous or harassing procedural tactics by lawyers; there is also a trend toward greater judicial supervision of discovery proceedings.

PRELIMINARIES TO PROCEEDINGS

Jurisdiction, competence, and venue. The words jurisdiction and competence refer generally to the power of an official body (legislative, judicial, or administrative) to deal with a specific matter. This section is concerned with judicial jurisdiction, the power of a court to act. That power may depend on the relationship of the court to the subject matter of the action; in such an instance one speaks generally of subject matter jurisdiction. The jurisdiction of a court may also depend on the relationship between the court and the defendant in the action. As to that relationship, important conceptual differences exist between the countries of the common-law orbit, which usually refer to this problem as the question of "jurisdiction over the defendant" and countries within the civil-law tradition, which are likely to subdivide the problem into questions of "international jurisdiction" (*i.e.,* which country may take the case) and questions of "territorial jurisdiction" (*i.e.,* courts in which part of the country may take the case). As noted above, in the United States, the due process clause of the Constitution imposes limits on the states' power to confer jurisdiction on their courts. It has been suggested that the word jurisdiction should be used only when discussing the power of the courts in a state generally to act in a given situation without violating the due process clause, whereas the word competence should be used to refer to the power of a particular court in a state to act pursuant to the laws of that state, but frequently the terms jurisdiction and competence are used interchangeably. (For a more detailed discussion, especially in relation to matters containing foreign elements, see below *Conflict of laws.*)

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

Venue

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

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

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

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

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

Class suits

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

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

Amicus curiae

ask the court for leave to appear as amicus curiae (friend of the court) so that he may present arguments in favour of the person he supports. In certain cases, furthermore, defendants are authorized to bring third parties into an action when, for instance, these third parties are or may be liable to the defendants on account of the claim asserted against the defendants. This is known as impleader.

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

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

Attachment

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

Temporary injunctions and restraining orders

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

In countries with a common-law tradition a person disobeying an injunction issued by a court is guilty of "contempt of court" and can be punished quite severely. In civil-law countries, punishment for contempt is largely unknown, and since broad orders to defendants may therefore be difficult to enforce, such orders are sometimes limited to specific narrow situations. For constitutional reasons, the U.S. Supreme Court has limited ex parte remedies.

Commencement of action. In Anglo-American procedure, a lawsuit is generally divided into two stages, the first, or pleading, stage and the trial stage. At the pleading stage the parties notify each other of their claims and defenses; at the trial stage, they or their counsel prove their factual contentions before the jury primarily through the oral examination of witnesses produced by them. The verdict of the jury and the judgment based on it follow immediately thereafter.

In civil-law countries, the procedure consists essentially of a series of hearings at which counsel argue their clients' position and submit documentary evidence; any other form of evidence can be utilized only with a special court order definitely describing the type of evidence and the matter to be proved by it.

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

Delivering the summons

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

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

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

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

In the United States during the 19th century, numerous procedural reforms were instituted. The parties were no longer required to plead on the basis of legal theories but instead were to allege a statement of facts constituting the cause of action or defense; the court could then apply any legal theory that was applicable under the facts alleged and later proved. The insistence upon fact pleading had substantial drawbacks, however, especially since the courts demanded a high degree of specificity, made technical

distinctions between fact and evidence (forbidding the insertion of the latter in the pleading), and bound the parties to prove the facts alleged or lose the lawsuit. This last rule was particularly harsh since it forced the party to allege detailed facts early in the proceedings when he frequently was not yet certain precisely what facts had occurred.

Modern reforms have gone a long way toward elimination of the injustices of the former system. U.S. federal rules require only "a short and plain statement of the claim showing that the pleader is entitled to relief"; the defendant "shall state in short and plain terms his defenses." There is no requirement that legal theory be stated in the pleading or that only facts be alleged. Other rules specifically permit the parties to plead alternative or contradictory claims or defenses and provide that in the usual case, only two pleadings, the complaint and the answer, shall be permitted. The effect of these changes has been to substantially downgrade the importance of the pleading stage of the lawsuit. The primary function of the pleadings is now only to give a general notice of the subject matter of the suit to the opposing party.

Civil-law pleadings

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

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

The time limits for the appearance vary greatly. European countries frequently provide a great many different time periods varying with the distance between defendant and the court where the action is pending. In some countries the time to appear is fixed by the court. Less attention is usually paid to geography in the United States. In New York, for instance, the defendant must appear in 20 days if the summons was served personally, and 30 days if some other form of service was employed.

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

The preparatory stage. As noted above, in countries whose procedure is based on English common law, the concentrated trial, traditionally before a jury, serves as a climax to earlier procedures. At this time, the parties attempt to prove the facts at issue, primarily through the presentation of oral evidence. The climax of a European proceeding, however, is the hearing before the full bench of judges—a hearing that is essentially devoted to argument of counsel and the presentation of documentary evidence. In both legal systems there are procedures to prepare for the trial or hearing.

Attempts to end a suit before trial

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

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

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

The European system is in some ways similar to the Anglo-American. Frequently, such questions as jurisdiction can be decided in the preliminary phase, without waiting for the full hearing. The preliminary phase may also serve to narrow issues and produce a settlement. Furthermore, proof proceedings sometimes occur during the preliminary phases rather than at the main hearing; though in Austria the full court holds hearings devoted to all aspects of the case, without distinguishing between matters considered preliminary and those more pertinent to the main hearing.

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

Summary judgment

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

In those civil-law countries that have a preparatory phase before a single judge and a final hearing before a three-judge bench, procedural defenses similar to pretrial motions are ordinarily raised before the single judge. Sometimes, however, in cases of lack of jurisdiction or lack of competence a hearing is held before the full court. Where the issue is one of territorial competence the result may be the transfer of the case to the proper court. General summary proceedings have lost considerable importance in France and have been abandoned completely in Italy, but in actions involving claims based on negotiable or other written instruments, for instance, special procedures have

been developed that permit a judgment to be obtained with great dispatch, particularly if the defendant has no effective defense on the merits.

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

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

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

European procedures to secure information

With the exception of procedures to secure, in advance of lawsuit, evidence that is in danger of being lost (for instance, because a witness may die), there are few procedures in civil-law countries to enable a party to secure information to use later. There are several reasons for this. The absence of a concentrated trial makes it less important to have all information available at once, and the greater role given the judge in bringing out factual matters further reduces the need to obtain information in anticipation of the hearing.

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

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

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

Settling out of court

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

In civil-law countries, procedures somewhat analogous in purpose to pretrial conferences are fairly prevalent. Since such preliminary hearings are ordinarily held before a single judge, rather than a formal three-judge court, a considerable saving of judicial time may result. Under the French code of civil procedure, each case is assigned to a special "prehearing" judge, who sets time limits for the exchange of pleadings, decides how many pleadings after the original summons and complaint shall be used and when they shall be submitted, and may penalize dilatory parties by delivering a default judgment or, if both sides are dilatory, by striking the case off the calendar. In addition, he may call in the parties' counsel for a conference and must make sure that all documents that the parties intend to use at the main hearing have been filed. He may also call in the parties themselves for a conference concerning a possible settlement. He must, in short, either settle the case or put it in shape for the formal hearing. Under the 1976 reforms to the West German Code of Civil Procedure, the parties may be directed, through a preliminary written or oral procedure, to prepare the main hearing in such a manner that it can lead to an immediate decision of the case.

THE TRIAL OR MAIN HEARING

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

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

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

Questioning and cross-examination

jury will retire to deliberate in private until it reaches a verdict, which will then be announced in open court.

Rules of evidence. Although the parties, and not the judge, are charged with the primary obligation to call and question the witnesses, the judge must act as arbiter in all disputes between the parties concerning the admissibility of evidence. When one party objects to the introduction of testimony, the judge will decide whether or not, in accordance with established rules of admissibility, the evidence sought to be introduced is to be heard by the jury. In keeping with the adversary system, the judge is not entitled to rule that evidence is inadmissible unless a party objects to its introduction. The party objecting to the evidence must state the grounds for his objection and the judge must permit the jury to hear the evidence unless the specified grounds given by the attorney are applicable. Even within this narrow framework, the judge's role is limited, for the rules of evidence leave little room for discretion on the part of the judge. (See below *The law of evidence.*)

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

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

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

Comments on evidence

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

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

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

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

Active role of the judge

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

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

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

The examination of an expert is obtained in the same manner as that of a witness. Although the parties may suggest an expert to the court, those chosen are ordinarily taken from a list of experts approved by the court. The expert is considered an impartial auxiliary of the court; his use is ordinarily limited to cases involving some technical or scientific problem. The court or judge issuing the proof

order may authorize him to make certain scientific investigations (*e.g.*, in an automobile accident case, to examine the car involved) and to report thereon.

Parties not considered witnesses

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

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

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

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

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

Dissenting opinion

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

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

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

Res judicata in civil-law countries

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

Enforcement of judgment. All countries have procedures intended to overcome the resistance of a party who fails to comply with the judgment of a court. This is usually known as the enforcement or execution of a judgment. Rules vary greatly, and they are usually highly technical and thus can be dealt with only generally. In the United States a party who obtains a judgment for a sum of money is entitled normally to avail himself at once of the procedural devices designed to enforce the judgment. The fact that the period for appeal has not yet passed or that an appeal is filed does not, of itself, affect the right to enforce the judgment; the losing party, seeking to postpone enforcement of the judgment pending appeal, must request such relief either from the trial court or the court to which the appeal is taken. Frequently, such a request will be granted if the losing party posts a bond or other security to ensure that the delay in enforcement will not adversely affect the rights of the winning party should the appellate court affirm the judgment of the trial court.

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

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

The remedy of garnishing the earnings of the defendant, although generally permitted, is accompanied by certain safeguards to prevent oppression. Thus, only if the debtor

fails to make payments voluntarily, can his wages be seized, and even then only a limited percentage of the wages.

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

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

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

Costs and disbursements. Generally, the prevailing party recovers not only the amount of the judgment but also the costs and expenses of the suit. These include filing fees, government taxes, witness fees, and the like, but not funds spent in the preparation of the case. In countries like Austria and Germany that regulate the fees of attorneys by an official schedule, such fees are ordinarily recoverable. In the United States, where such fees normally are not regulated by schedule, they usually must be borne by the party that has incurred them.

APPEALS AND OTHER METHODS OF REVIEW

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

Differences between common law and civil law

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

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

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

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

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

Civil-law appellate procedure. Appeals to intermediate appellate courts from courts of first instance are available quite broadly in Europe, frequently for all judgments exceeding a certain amount and at times for certain types of judgments, regardless of amount. This encourages appeals to intermediate appellate courts and explains their frequently very heavy case load.

Since the appeal involves a new hearing of the case, the procedure is essentially similar to that in use in courts of first instance, though entirely new claims may not be presented. In the case of a review of a nonfinal judgment, the appellate court frequently limits its review to an examination of the legal correctness of that judgment and then remands the case, so that proceedings in the court below may be completed. Occasionally, appellate courts are authorized to use the occasion of an appeal of a nonfinal judgment in order to decide the entire case themselves.

Appeals to supreme courts

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

quired. Consequently, case loads are extremely heavy, and to handle them the full court does not usually sit together but instead is divided into panels. In important matters two or more panels may sit together. (P.E.H./Ma.E.O.)

Criminal procedure

The law of criminal procedure regulates the modes of apprehending, charging, and trying suspected offenders; the imposition of penalties on convicted offenders; and the methods of challenging the legality of conviction after judgment is entered. Litigation in this area frequently deals with conflicts of fundamental importance for the allocation of power between the state and its citizens.

PROCEDURE BEFORE TRIAL

The investigatory phase. When a criminal offense has been reported, the competent authority (the police, the public prosecutor, or the investigating magistrate) commences the criminal process by investigating the circumstances. In this phase, relevant evidence is collected and preserved for a possible trial. The suspect also has the right to collect evidence in his favour. In the civil-law countries of continental Europe, he can typically request the investigating authority to assist him in this endeavour; in common-law countries, the suspect is expected to take the initiative in preparing the case for his defense.

Gathering evidence

The role of the police. The police play a primary role in the investigation. They are responsible for interrogating suspects and witnesses, and they carry out arrests, searches, and seizures. In Anglo-American legal systems the police perform investigations on their own authority, whereas on the Continent they act under the formal supervision of public prosecutors or investigating magistrates.

The role of the magistrate. In some countries, such as France and Italy, a magistrate conducts the investigation in cases of serious criminal offense, personally hearing witnesses and directing police to perform such relevant acts as the seizure of evidence.

In many other jurisdictions, as in the United States and Germany, magistrates do not organize or conduct the investigation. Their role is limited to authorizing particular acts of investigation involving serious invasions of civil rights—most important, instances of arrest, pretrial detention, search, seizure, and surveillance of mail and telecommunication. Generally, such acts are lawful only upon prior written judicial authorization (the warrant). Under U.S. law, warrants are issued only upon probable cause—that is, when there is evidence leading to a reasonable belief that the person to be arrested has committed a crime or that an object connected with criminal activity can be found at the place to be searched. Other legal systems employ less-stringent standards of suspicion.

When it is necessary for police to act on the spot—for example, because the suspect is about to escape or because he will destroy the contraband sought—they can take the proper measures without prior judicial authorization. In most cases, such provisional measures can or must be submitted later to judicial control.

The role of the prosecutor. Public prosecutors are lawyers appointed by the government as its representatives in criminal matters. In the United States, most state or county prosecutors are elected.

In some legal systems, as in Germany, the prosecutor is formally responsible for conducting criminal investigations. In practice, however, his role is generally limited to advising and supervising police. Only in very serious or politically sensitive matters does he personally conduct the investigation.

The role of the suspect. Since the 19th century, the law has gradually recognized the suspect's autonomous position as a subject of the criminal process. His right to remain silent in order to avoid incriminating himself has, in principle, been acknowledged universally. However, few legal systems go so far as the United States, where, under the *Miranda* v. *Arizona* ruling of 1966, the defendant's statements will be excluded from evidence if he is not specifically warned of his right to remain silent before interrogation while in police custody. In most countries, evidence of a confession is admissible in court unless the confession is shown to have been "involuntary"—for example, acquired by torture or threats.

Defendants' rights

On the other hand, the defendant has a universally recognized right to present to the court his view of the facts. In many jurisdictions, this right can be exercised even before the court decides whether there is sufficient evidence to hold a trial.

The role of defense counsel. The defense lawyer has a double function in the investigation phase of the criminal process: to assist the suspect in gathering exonerating evidence and to protect him from violations of his rights at the hands of law-enforcement personnel. All legal systems grant the suspect the right to the assistance of an attorney, and in many countries the suspect must be informed of this right before police interrogate him. If the suspect does not have the means to hire a lawyer, often the state will pay the attorney's fee or provide the suspect with state-employed counsel.

However, the law also restricts defense counsel's ability to carry out his functions. In some jurisdictions, as in France, the attorney has no right to be present when the suspect or a witness is interrogated by the police; only a few countries, such as the United States, grant the defense the right to compel witnesses on its behalf to appear in court. Moreover, in most jurisdictions the defense has no or only limited access to information gathered by the prosecution before the case reaches the court.

Pretrial detention. Incarceration of the suspect before trial most seriously impairs the preparation of an effective defense. Nevertheless, all legal systems permit pretrial detention, though under differing conditions.

In Anglo-American jurisdictions the rule is that suspects arrested and not released immediately for want of cause are held in custody. However, the suspect generally has a right to be released on a financial surety, or bail, the amount of which is set by the magistrate according to the individual circumstances of the case. The purpose of bail is to assure appearance of the suspect at the trial; hence, it will be forfeited if the suspect absconds. In appropriate cases the suspect can be released on his own recognizance (*i.e.*, without providing bail). Only under special circumstances—for example, when it is thought that the suspect might commit further offenses if released—can bail be denied altogether.

Bail

In continental Europe the law treats pretrial detention as the exception rather than the rule. The magistrate can remand the suspect to custody before trial only if this is necessary to prevent him from escaping, tampering with evidence, or committing further serious offenses. Even on the Continent, the law authorizes the court to release a suspect from custody if sufficient surety is posted.

The decision to prosecute. A formal accusation is universally regarded as an indispensable prerequisite for a criminal trial. It is typically the public prosecutor who, on the basis of the results of the investigation, determines whether to file a complaint and for which offense to bring charges.

Private prosecution. Private citizens, such as the victim of the offense, are not generally permitted to institute a criminal action, though the law on this point differs among jurisdictions. In the United States private criminal complaints are practically impossible. In England anyone can institute criminal proceedings for most offenses, but the director of public prosecutions can take over and discontinue prosecution at any time. In Germany citizens can prosecute only for certain minor offenses such as libel and assault. In France victims of crime can combine criminal prosecution with civil claims for damages.

In many countries victims can prevent prosecution for certain offenses—*e.g.*, assault, libel, and some sexual offenses—by not filing a special request for public prosecution.

Grand jury. In the federal system of the United States, and in about half of the state systems, charges are brought not by the public prosecutor but by the grand jury, a group of 12 to 23 citizens selected by lot. The grand jury also has investigative authority, and it is to serve as a protective shield against unwarranted prosecution. In

practice, however, grand juries are usually dominated by the public prosecutors, who are responsible for presenting the evidence to them.

Prosecutorial discretion. In all legal systems the prosecutor should bring an accusation only if he thinks that the available evidence, discounted by probable defense evidence, is so strong that the defendant is likely to be convicted after trial. In some countries, such as Italy, the prosecutor is required by law to bring charges whenever there is sufficient evidence for conviction. In other jurisdictions—for example, in the United States, France, and Japan—the public prosecutor has discretion whether or not to file a formal accusation; in effect, this means that he can informally grant clemency to an offender who would certainly be convicted in court. In still other countries, such as Germany, prosecutorial discretion applies only to minor offenses, whereas prosecution of serious crimes is mandatory. To the extent the prosecutor has discretion, he can make the decision not to prosecute dependent upon certain conditions—e.g., that the offender pay restitution to the victim.

Legal obligation to prosecute

Plea bargaining. In some countries, such as the United States and Spain, the prosecutor's discretion extends to determining the crime with which the defendant is to be charged. Hence, in a case of armed robbery, the U.S. prosecutor may charge the suspect with armed robbery, simple robbery, assault, simple theft, or any combination of these offenses. All of these offenses carry quite different penalties, and normally the prosecutor charges the most serious offense that can be sustained by the evidence. However, since the cooperation of the defendant, especially in offering a plea of guilty, drastically shortens or simplifies the trial, prosecutors in some countries reduce charges on the condition that the defendant not contest the accusation in court. Especially in the United States, this creates a system of "plea bargaining," in which defense attorneys negotiate with prosecutors the charges (and resulting penalties) most acceptable to their clients. Similar transactions, though sometimes performed discreetly because of their dubious legality, occur in many other jurisdictions.

TRIAL PROCEDURE

Criminal courts. In most countries, two or three types of courts have jurisdiction in criminal matters. Petty offenses are usually dealt with by one professional judge; in England, however, two or more lay justices may sit in Magistrates' Court. Matters of greater importance are, in many countries, tried by panels of two or more judges. Often such panels consist of lawyers and lay judges, as in Germany, where two laypersons sit with one to three jurists. The French *cour d'assises,* which hears serious criminal matters, is composed of three professional judges and nine lay assessors. Such "mixed courts" of professionals and ordinary citizens deliberate together and decide by majority vote, with lawyers and laypersons having one vote each.

By contrast, the jury system, a distinctive feature of the Anglo-American criminal process, involves a division of functions between the presiding judge and the laypersons sitting as jurors. The judge presides over the trial, determines the admissibility of evidence, and instructs the jury on the applicable law, but he does not participate in the deliberations of the jury. The jurors usually remain silent during trial but are autonomous in finding the verdict of guilty or not guilty.

The U.S. Constitution guarantees every defendant in a non-petty case the right to be tried before a jury; the defendant can also waive this right and have a professional judge sitting alone decide on the verdict. To ensure the impartiality of the jury, prosecution and defense can reject (in legal parlance, challenge) jurors whom they establish to be biased. Moreover, the defense (and in the United States the prosecution as well) has the right of peremptory challenge, in which it can challenge a number of jurors without having to give a reason.

Formation of the jury

Pretrial matters. In many legal systems, the court checks the accuracy of the accusation before admitting the case for trial. In France a special panel called the *chambre d'accusation* determines whether there is enough evidence for the case to proceed; in England the Magistrate's Court makes the decision on "binding over" the defendant for trial; and in Germany the trial court itself (sitting without lay assessors) decides whether there is sufficient evidence. In the Anglo-American system, the court holds a hearing to determine "probable cause" for trial; under continental law, courts usually make that determination on the basis of the documents assembled in the course of the investigation.

A characteristic feature of the Anglo-American criminal process is the opportunity for defendants to plead guilty or not guilty. Only if the defendant contests the accusation by pleading not guilty is a trial held. Otherwise, the court pronounces the defendant guilty as charged and goes on to determine the penalty. With few exceptions (as in Spain), continental law does not provide for such shortcuts to sentencing. Rather, a trial must be held even if the defendant has confessed guilt from the outset.

Publicity of the trial. Trials, as opposed to pretrial investigation, must be accessible to the public. This principle, embodied in the constitutions of several countries, is meant to protect the defendant; in the United States it is also based on the freedom of the press. Publicity does not mean that broadcasting of trials must be permitted; in most countries, it is not allowed.

In spectacular cases, great publicity can influence the court and work to the detriment of defendants. Most legal systems, therefore, permit the court to exclude the public from the trial (or from parts thereof) or to change the location in which the trial is to be held if either measure is necessary to protect the trial process from undue interference.

Presentation of evidence. In Anglo-American law the presentation of evidence is left to the parties. Witnesses are examined and cross-examined by counsel, not by the court. The function of the trial judge is to enforce the rules governing evidence and to ask supplementary questions if he feels that the parties have failed to clarify the facts. The defendant may testify as a witness if he chooses to, but he is not examined by the judge. Under continental law, by contrast, the presiding judge typically dominates the process of taking evidence. He is responsible for establishing the relevant facts by calling and questioning witnesses and for introducing real evidence. The judge also interrogates the defendant unless the latter chooses to remain silent. Attorneys for the prosecution and the defense ask additional questions of witnesses and summarize the evidence at the end of the trial.

Finding the verdict. A basic principle of both Anglo-American and continental procedures is that the defendant is presumed innocent unless and until his guilt has been established beyond a reasonable doubt. The burden of proof, therefore, rests upon the prosecution. On the Continent, this is true even in cases involving insanity, drunkenness, self-defense, or necessity. Anglo-American law regards these as "affirmative defenses" and requires the defendant to provide at least some evidence that they were a factor.

A basic principle of criminal justice

Courts in continental legal systems are not bound by any legal rules concerning the evaluation of evidence presented; rather, they are to follow their conscience in establishing guilt or innocence. The same is generally true for juries in the Anglo-American system; however, since juries are thought to be easily distracted from the real issues of the case, there is a complicated set of legal rules determining what evidence can be presented to juries. (See below *The law of evidence.*)

In the United States, jury verdicts must be unanimous; if the jury is unable to agree, a new trial before another jury can be held. In England, majority votes by margins of 10 to two or nine to one are acceptable after the jury has deliberated for at least two hours. As a corollary of the presumption of innocence, many continental systems require a specified majority of the judges to vote for a finding of guilty.

Sentencing. In continental systems, the court decides, on the basis of a single comprehensive trial, both on the guilt or innocence of the defendant and on the penalty if he is found guilty. Sentences are conclusively determined

by the court, with prison terms being subject to conditional release.

Anglo-American law provides for separate sentencing hearings, which typically take place a few weeks after the defendant has been found guilty of the charges. In the interim, social workers gather information on the offender's psychological and social background, which they present to the court. Usually, a single professional judge determines the sentence after hearing the defense (and, in the United States, the prosecution). In the United States, juries in several states make a recommendation with respect to capital punishment in cases where the death penalty is available as a sentence.

POST-CONVICTION PROCEDURE

Common law. In Anglo-American legal systems, a convicted defendant may move in the trial court to arrest judgment, or he may file a motion for a new trial. The legality of the conviction may also be challenged by appeal to a higher court. Criminal appeals were unknown in the traditional common law, but today they are universally granted by statute. In the United Kingdom, the Criminal Appeal Act of 1907 established an elaborate system of appellate procedure, proceeding from Magistrate's Courts all the way to the House of Lords, the supreme court of England. Extraordinary remedies available in English procedure include the writ of habeas corpus (determining the legality of holding the prisoner in custody) and the orders of mandamus (compelling an official to perform an act required by law), certiorari (requiring a lower court to present the trial record to a higher court), and prohibition (by which a higher court prohibits a lower court from exceeding its jurisdiction).

Mechanisms of appellate review

In the United States, a defendant convicted in a state or federal court can appeal to that state's (or the appropriate federal) appellate court. Subject to certain restrictions, the defendant can turn to the federal court system when his rights under the U.S. Constitution have been violated in state court. Review by the U.S. Supreme Court is discretionary; the court grants it only in cases of general significance by issuing a writ of certiorari to the court whose judgment is to be reviewed. Even after the regular avenues of appeal are exhausted, defendants in custody can at any time apply for a writ of habeas corpus, challenging the prison warden's right to keep the petitioner in custody and demanding his release. Since the warden's right usually depends on the validity of the criminal judgment, habeas corpus constitutes an indirect method of review. Legislation in the 1970s curtailed access to federal courts on the basis of habeas corpus.

While defendants enjoy a liberal right to appellate review in criminal matters, the prosecution generally cannot appeal an acquittal. This is due to a strict interpretation of the concept of double jeopardy, which forbids a defendant to be tried twice for the same act.

Appellate courts do not take evidence but only decide points of law on the basis of the record. Since juries do not give reasons for their verdicts, appeals are usually based on allegations of faulty procedure (in particular, the admission and exclusion of evidence) and on erroneous statements on the applicable law in the judge's instruction to the jury. The sentence is also subject to review in Britain and Canada but not in most of the United States.

Civil law. Appellate procedure on the Continent follows quite different rules. Most important, the prosecution as well as the defense can appeal a judgment, including the sentence. In some countries (*e.g.*, Germany) it is possible to demand a new trial in a higher court if the original trial was held by a single judge. In other cases, appellate courts review only matters of substantive or procedural law, including the question of whether the lower court did everything necessary to find the relevant facts. Continental trial courts usually write elaborate reasons for their judgments, and it is these reasons that form the objects of the appellate courts' scrutiny.

When appellate review is waived or exhausted, judgments are deemed final and can be executed. Final judgments can be overturned only if significant new evidence is found indicating that the decision was wrong. (H.-H.J./T.We.)

The law of evidence

To the end that court decisions are to be based on truth founded on evidence, a primary duty of courts is to conduct proper proceedings so as to hear and consider evidence. The so-called law of evidence is made up largely of procedural regulations concerning the proof and presentation of facts, whether involving the testimony of witnesses, the presentation of documents or physical objects, or the assertion of a foreign law. The many rules of evidence that have evolved under different legal systems have, in the main, been founded on experience and shaped by varying legal requirements of what constitutes admissible and sufficient proof. Although evidence, in this sense, has both legal and technical characteristics, judicial evidence has always been a human rather than a technical problem. During different periods and at different cultural stages, problems concerning evidence have been resolved by widely different methods. Since the means of acquiring evidence are clearly variable and delimited, they can result only in a degree of probability and not in an absolute truth in the philosophical sense. In common-law countries, civil cases require only preponderant probability and criminal cases, probability beyond reasonable doubt. In civil-law countries so much probability is required that reasonable doubts are excluded.

THE EARLY LAW OF EVIDENCE

Characteristic features of the law of evidence in earlier cultures were that no distinction was made between civil and criminal matters or between fact and law and that rational means of evidence were either unknown or little used. In general, the accused had to prove his innocence.

Nonrational sources of evidence. The appeal to supernatural powers was, of course, not evidence in the modern sense but an ordeal in which God was appealed to as the highest judge. The judges of the community determined what different kinds of ordeals were to be suffered, and frequently the ordeals involved threatening the accused with fire, a hot iron, or drowning. It may be that a certain awe associated with the two great elements of fire and water made them appear preeminently suitable for dangerous tests by which God himself was to pass on guilt or innocence. Trial by battle had much the same origin. To be sure, the powerful man relied on his strength, but it was also assumed that God would be on the side of right.

Trial by ordeal and battle

Semirational sources of evidence. The accused free person could offer to exonerate himself by oath. Under these circumstances, in contrast to the ordeals, it was not expected that God would rule immediately but rather that he would punish the perjurer at a later time. Nevertheless, there was ordinarily enough realism so that the mere oath of the accused person alone was not allowed. Rather, he was ordered to swear with a number of compurgators, or witnesses, who confirmed, so to speak, the oath of the person swearing. They stood as guarantees for his oath but never gave any testimony about the facts.

The significance of these first witnesses is seen in the use of the German word *Zeuge*, which now means "witness" but originally meant "drawn in." The witnesses were, in fact, "drawn in" to perform a legal act as instrumental witnesses. But they gave only their opinions and consequently did not testify about facts with which they were acquainted. Nevertheless, together with community witnesses, they paved the way for the more rational use of evidence.

The influence of Roman-canonical law. By the 13th century, ordeals were no longer used, though the custom of trial by battle lasted until the 14th and 15th centuries. The judicial machinery destroyed by dropping these sources of evidence could not be replaced by the oath of purgation alone. With the decline of chivalry, the flourishing of the towns, the further development of Christian theology, and the formation of states, both social and cultural conditions had changed. The law of evidence, along with much of the rest of the law of Europe, was influenced strongly by Roman-canonical law elaborated by jurists in northern Italian universities. Roman law introduced elements of common procedure that became known throughout the

continental European countries and became something of a uniting bond between them.

Under the new influence, evidence was, first of all, evaluated on a hierarchical basis. This accorded well with the assumption of scholastic philosophy that all the possibilities of life could be formally ordered through a system of a priori, abstract regulations. Since the law was based on the concept of the inequality of persons, not all persons were suitable as witnesses, and only the testimony of two or more suitable witnesses could supply proof.

The formal theory of evidence that grew out of this hierarchical evaluation left no option for the judge: in effect, he was required to be convinced after the designated number of witnesses had testified concordantly. A distinction was made between complete, half, and lesser portions of evidence, evading the problem posed by such a rigid system of evaluation. Since interrogation of witnesses was secret, abuses occurred on another level. These abuses were nourished by the notion that the confession was the best kind of evidence and that reliable confessions could be obtained by means of torture.

Despite these obvious drawbacks and limitations, through the ecclesiastical courts Roman-canonical law gained influence. It contributed much to the elimination of nonrational evidence from the courts, even though, given the formality of its application, it could result only in formal truths often not corresponding to reality.

COMPARATIVE SURVEY OF MODERN PRINCIPLES

A comparison of the principles of evidence under different legal traditions can best be made by examining the rights and obligations of the plaintiff and the defendant in civil proceedings and of the prosecutor and the accused in criminal proceedings. The position of the judge is also crucial. Historically, two systems developed.

Civil law and common law

The first, which follows what may be called the inquisitorial principle, had its origins in medieval Roman-canonical proceedings. It is distinguished by the active part played by the judge, who, by virtue of his office, himself searches for the facts, listens to witnesses and experts, examines documents, and orders the taking of evidence. In continental European countries and those other countries that derive their law from them, this system has generally been retained for criminal proceedings. The prosecutor and the accused, of course, give their recital of the facts and indicate their evidence for specific assertions. But, by virtue of his role in the case, the judge must make further investigations if he deems them necessary to obtain the truth. In some western European countries, there is a definite inclination toward employing this inquisitorial system in all legal proceedings that have, or could have, a substantial public legal impact—*e.g.*, matrimonial, status, administrative, social, labour, and financial matters.

The second system, which employs what are usually called accusatorial or adversary principles, is used in the common-law countries for all civil and criminal cases. In this system, the parties and their attorneys are primarily responsible for finding and presenting evidence. The judge does not himself investigate the facts. Only if the efforts of the parties are incomplete must the judge make inquiries with regard to questions that have remained unanswered.

In civil matters, most continental European countries follow a mixed system of both inquisitorial and adversarial principles. In some of these countries, the judge can, for example, hear witnesses who have not been designated by the parties, and in all countries he can, by virtue of his office, hear the parties and experts and order documentary evidence or the actual inspection of evidence. In contrast to criminal cases, the continental European judge is always bound by the motions and assertions of the parties.

Oral proceedings. Under both systems of presenting and obtaining evidence, oral proceedings are generally accepted. The written proceedings favoured during the Middle Ages have been abolished, although the parties prepare their lawsuits through briefs, and parts of the preliminary proceedings can be handled in writing. The interrogation of witnesses, however, is oral. Most civil-law countries do not permit any exceptions, while other countries, such as Germany, permit written statements by witnesses in special cases and with the consent of the parties. In the common-law countries an exception is made to the principle of oral proceedings for certain types of affidavits, and, particularly in civil cases, the practice has steadily gained in importance.

Direct interrogation of witnesses by the deciding court is an aspect of the law of evidence closely connected with oral proceedings. Generally, in continental European countries, witnesses are interrogated by the judges who decide the verdict, but a number of countries have an investigation procedure according to which another judge, or only one member of the judging body, interrogates the witnesses. Under both the inquisitorial and the accusatorial systems, the principle of direct interrogation is of special importance in the free consideration of evidence. In the common-law countries the function performed by the judge in this regard is handled by attorneys for the prosecution or defense, with the judge's role restricted almost entirely to overseeing the questioning.

Influence of the jury

One major influence that has shaped the law of evidence has been the jury system. At least one writer has said that the law of evidence is the child of the jury. Oral proceedings, direct interrogation, and the public trial are much less problematic under the Anglo-American system than under the civil-law system to the extent that evidence is heard before the jury. But this system has spawned a large number of regulations for the admissibility of evidence in order to guarantee due process and fair procedure and to protect the jury from being misled. The initiative of the parties determines the handling of these regulations, for they must raise objections if, in their opinion, any of the numerous exclusionary rules is being violated. The judge then rules on the objection. By the complex working of this arrangement, the Anglo-American system has become more formalistic in many respects than the continental European system.

The burden of proof. The burden of proof is a manifold and somewhat ambiguous concept in the law of evidence.

The burden of producing evidence means that in general the party that cites specific facts for the substantiation of its claim also has the burden of producing the evidence to prove these facts. This burden depends on the substantive law governing the claim. Permissible presumptions and legal rules can shift the burden in various situations.

The burden of conviction, on the other hand, comes into play at the end of the hearing of evidence, if doubts remain. This is simply to recognize that the evidence is not sufficient to convince the jury or the judge and that, in general, the party having the burden of pleading and producing facts favourable to itself and of giving evidence also carries the so-called burden of conviction.

Difference between the burden of proof in civil and criminal proceedings

Whereas in civil proceedings it is generally the plaintiff who has the burden of proof for facts supporting a claim, unless this burden has been shifted to the defendant through rules or presumptions, in criminal proceedings it is the prosecution that bears the burden of proof for all relevant facts. What this means is that the defendant cannot be found guilty as long as proof has not been supplied or as long as doubts still remain. In continental European law, no distinction is made between civil and criminal cases with regard to the standard of proof. In both, such a high degree of probability is required that, to the degree that this is possible in the ordinary experience of life itself, doubts are excluded and probability approaches certitude. In the common-law countries the degree of probability required in civil cases is lower than that called for in criminal matters.

Relevance and admissibility. In civil proceedings in the common-law countries, evidence is both ascertained and simultaneously restricted by the assertions of the parties. If the allegations of one party are not disputed or contested by the other, or if the allegations are even admitted, then no proof is required. Proof would, in fact, be irrelevant. Evidence offered to prove assertions that are neither at issue nor probative of the matter at issue would also be irrelevant. The only evidence that is, therefore, relevant, is evidence that to some degree advances the inquiry and has a probative value for the decision. While continental European judges, in ordering the hearing of evidence or in

deciding on evidence, indicate the facts to be proved and thereby strictly eliminate irrelevant facts, Anglo-American judges first give the parties an opportunity to furnish any evidence that they deem suitable. If, during the hearing of witnesses, irrelevant questions are put, they are rejected after the adversary has objected to them.

It has been said that relevance depends on logical considerations and that admissibility depends on the law. In contrast to civil law, the common law has developed a large number of rules governing the admissibility of evidence. Relevant evidence is not admissible, for example, if the witnesses are excluded from testifying because of incompetency, or if they are protected by privileges against self-incrimination, or in instances in which they would have to divulge confidential or professional communications that have a privileged status or government secrets, or, again, when the evidence is excluded by the rules against hearsay (see below *Sources of proof: Witnesses*).

In criminal cases in civil-law countries, relevance relates to such questions that are so far removed from the case that they have no evidence value at all. Admissions and confessions do not exclude further evidence. According to Anglo-American law, the accused may be a competent witness under the admissibility rules, but, in contrast to an ordinary witness, he has the privilege of not taking the witness stand. According to continental European law, the accused is neither a party nor a witness. He can be heard, but he cannot be forced to answer questions of fact. In general, Anglo-American rules of admissibility apply to criminal proceedings much as they apply to civil cases.

The free evaluation of evidence. Freedom to evaluate all the evidence produced was established in Roman law but fell into disuse as a principle during the time of the formalistic Roman-canonical law of evidence that characterized the Middle Ages. Remnants of the medieval formal theory of evidence survive in various countries.

Conviction intime: the judge's opinion

In countries where remnants of the medieval formal theory of evidence are still preserved, the principle of free evaluation of the evidence by the judge generally dates from the French Revolution. The French introduced the concept of the judge's *conviction intime* (inner, deep-seated conviction) in contrast to rules of formal evidence that prescribed exactly when the evidence amounted to proof. The primacy this gave to the personal conviction of the judge meant that it was not even necessary to state the reasons for the inner conviction. This total dependence on the judge's discretion aroused a great deal of criticism, and, as a result, various judicial codes prescribed that, in giving the grounds on which judgment was based, the judge had to specify in writing why he was convinced in each case. *Conviction intime* in its original sense is limited to the testimony of witnesses and experts and to the explanations of the parties. Both kinds of formal oaths made by parties to a case, the supplementary oath and the tendered oath, are still valid in civil-law countries, and both may lead to formal solutions, since the judge must follow the legal consequences of the oath. But these survivals of medieval formal evidence theory have been weakened. In France, for example, the judge's latitude under the principle of *conviction intime* has been extended to allow him to pass on the affirmation oath of the party, which formerly had to be given a certain value, regardless of his opinion of its worth. In other states, such as Austria, Germany, and the Scandinavian countries, the formal oath of the parties was abolished and replaced by the free depositions of the parties. Even if the parties take an oath on their testimonies during this process, the judge is not bound by it but may still make his own evaluation of the evidence. In addition, some remnants of the formal evidence theory have been preserved with regard to documentary proof where rules of procedure contain presumptions as to the conclusiveness of certain documents. Since reliance on documentary evidence prevails in some countries, these formal evidence rules are still of special importance.

In Anglo-American law the problem of free evaluation of evidence can be understood through the institution of the jury. Obviously, the evidence must be convincing to the common sense of the jury members, who form their judgment on the basis of free conviction. The function of the jury, however, is to decide questions of fact rather than questions of law, which are left to the judge. The jury's verdict can be overturned by the judge if it is inconsistent with the evidence or with his instructions as to the law governing the case. The judge's relationship to the jury therefore plays a role in the decisions, and there are difficult questions in which it is unclear whether the jury or the judge should consider the evidence. Some formal rules of evidence survive in Anglo-American law. In some cases evidence must be corroborated before it can constitute proof. In homicide cases, for example, a confession must be supported by additional evidence. In addition, evidence by witnesses is sometimes excluded by rules of admissibility.

SOURCES OF PROOF

According to Anglo-American law, the classic means of proof are witnesses, documents, and real evidence (derived from the actual inspection of objects). As a result of historical development, the status of witness was accorded to experts and to the parties in a civil lawsuit, and even to the accused in criminal proceedings. The development of continental European law has taken a different course. Parties cannot be witnesses, and evidence by experts is subject to special procedural rules. Consequently, there are essentially five separate sources of evidence: witnesses, parties, experts, documents, and real evidence.

Witnesses. The oral testimony of witnesses competes in a sense with documentary evidence to the extent that one may exclude or supplement the other. Under Anglo-American law, almost anyone can be a witness, including the parties and experts; even insane persons, children, and convicted felons may testify. Grounds once used for excluding such persons as witnesses are now used only to impeach their credibility. Continental European countries, as has been said, do not treat either the parties or experts as competent witnesses, and they are still suspicious of interested witnesses. Some of them, influenced by the Roman-based school, deny, on the whole, the capacity of those persons having a certain degree of relationship to the parties. Some consider insane persons incompetent to testify, others grant them the competency but exclude their testimony on the grounds of credibility. The capacity to be a witness does not depend on whether or not the person can testify about questions relevant to the specific case. In general, the tendency has been to utilize all persons who can testify about facts that will help to establish the truth. Competency as a witness has therefore been extended to as many persons as possible. On the other hand, many persons are protected by law from being forced to testify. This type of protection derives either from privilege, or from the right to refuse to give evidence, either case distinguishable from incapacity to testify. Whereas privilege or the right to refuse to give evidence may be either requested or waived, incapacity to testify takes effect automatically; *i.e.*, it must always be officially considered by the court.

Competent witness

Privileges. Privileges under Anglo-American law must be distinguished from the right to refuse to give evidence under particular circumstances as it exists in continental European practice. The latter is granted to witnesses for either personal or objective reasons. The personal reasons are the same as those that result in incapacity to testify—*i.e.*, relationship, affinity, and marriage. The objective reasons concern persons who, as a result of their profession (for example, clergymen, physicians, attorneys, journalists), have been put in possession of confidential facts. Such confidants have a limited right to refuse to give evidence so long as the person protected does not give his consent (the German solution). In some cases they are not admitted as witnesses without the consent of the protected person (the Swedish solution). Thus the Swedish judge officially decides whether the protected person has given his consent, whereas the German judge leaves the decision whether to testify up to the confidant. In addition, witnesses might refuse to testify if their testimony were to cause direct financial damage to themselves or to their families, or if it were to publicly disgrace them or expose them to criminal prosecution. All persons may make their own decision to testify, but judges are obliged

to inform them about their specific rights in the matter. These procedural regulations have developed in order to avoid the situation in which the person protected becomes caught in a conflict between the truth and his personal interests. The interests of the protected person—perhaps partly out of realism—are thus given a higher value than the search for the facts.

The Anglo-American privileges differ from the continental European right to refuse to testify insofar as privileged persons cannot decide whether or not they wish to testify. They may only cite their privileges, and the judge decides if they must testify. Under a system that stresses the free evaluation of evidence, the obligation to testify is subject to only a very few exceptions.

Self-incrimination

The privilege against incriminating oneself has a twofold nature in Anglo-American law because, in civil proceedings, parties may appear as witnesses and, in criminal proceedings, the accused may appear as a witness. The privilege of an ordinary witness is considerably limited. He must submit to being designated and sworn in as a witness in all instances and must answer all questions except those that are self-incriminating. Consequently, either he or his attorney must sift out the incriminating questions that will evoke the privilege. This is not always easy, particularly since it is only the witness, and not the party or the party's attorney, who may cite the protecting privilege. Critics have called this privilege a sentimental institution, but it is worth noting, in this regard, that the privilege against self-incrimination is included in the U.S. Bill of Rights.

It has already been pointed out that in the common-law system, the accused in a criminal trial no longer lacks competence as a witness but may exercise the privilege of refusing to be called or sworn as a witness. Unlike ordinary witnesses, he may invoke this privilege with considerable latitude, but once he does decide to step into the witness box, he renounces his privilege and may be interrogated as if he were an ordinary witness. The question then arises whether the waiving of the privilege against self-incrimination is limited to testimony concerning crimes of which he presently stands accused, or whether he must answer all questions regarding criminal acts. It appears to have become fairly well established that the prosecutor can, in fact, interrogate the defendant about previous criminal offenses. In civil cases the parties have the same privilege for protection from self-incrimination as other witnesses; *i.e.,* they need not answer incriminating questions.

Privileges deriving from personal and professional relationships are generally not granted on principle, though historically a privilege for the protection of marital communications has developed. In England an 1853 law decreed that a husband could not be forced to testify concerning information that his wife may have given him during the course of the marriage. This, naturally, also applies to the wife. In the United States the courts contended that laws concerning testimony on matrimonial communications contained only a statement of the common law. Only the beneficiary of the privilege may cite it, and it is not applicable where criminal offenses by one spouse against the other or against the children are concerned or in the case of a divorce proceeding.

Professional confidentiality

Attorneys are considered to be under an obligation to refuse to testify about confidential communications with their clients. The privilege, however, protects the client, not the attorney, and, therefore, the client may waive it. This privilege applies principally to the adversary system, in which, so to speak, the attorney is the client's champion.

Clergymen are likewise under obligation to refuse to answer questions concerning information given them in the secrecy of the confessional by believers. Again, the privilege protects the believer. This custom has been sanctioned by legislation in many U.S. states. In England, however, there is no common-law rule for this privilege.

Physicians, as a rule, must answer all questions since there is no common-law privilege regarding confidential information furnished by the patient. In some U.S. states an appropriate privilege has been created by legislation; again, it is the patient who is protected, and only he may waive the privilege.

Journalists, like physicians, occupy a position that is not entirely clear. In some jurisdictions they may refuse to testify about their sources of information, and in a number of U.S. states such a privilege has been specifically created by statute. In other U.S. states and in England the question does not yet seem to have been settled.

Swearing. The oath, perhaps the oldest means for encouraging truthful testimony, forms a link between court proceedings and religious belief since, in its usual form, witnesses swear by Almighty God that they are speaking the truth. Though the effectiveness of such an act has certainly diminished in secular societies, this appeal to God has for centuries been considered the surest means of obtaining truth. There are two kinds of oaths, the preliminary and the subsequent. In Anglo-American practice the witness is sworn in before testimony. Under German and other continental procedures, the swearing-in may occur after testimony as well. The latter method allows the judge to use his own discretion in individual cases as to whether or not the witness should be ordered to swear. In current German practice, very few witnesses are sworn in for testimony in civil proceedings, whereas in criminal proceedings all witnesses have to swear. Some continental European countries allow witnesses who object to oaths to substitute a solemn affirmation, and Denmark has abolished all oaths in legal procedures. The oath of a witness does not have the formal effect of binding the judge or the jury. They must evaluate it and the testimony freely.

Examination and cross-examination. Judges and attorneys in common-law courts regard the opportunity to cross-examine as a guarantee of the reliability and completeness of testimony by a witness. Under the perfect operation of the adversary system it is not the judge but rather the parties or their attorneys who interrogate the witnesses. The plaintiff's attorney begins the "examination in chief," which is subject to a number of restrictions. Leading, misleading, and argumentative questions, for example, are not permitted. After the plaintiff's attorney concludes his interrogation, the defendant's attorney may cross-examine the same witness. This cross-examination generally consists of leading questions posed with the intent of weakening or invalidating the impression created by the direct testimony of the witness. The cross-examination must ordinarily be limited to subjects covered during direct interrogation. There is a recognizable tendency, however, for cross-examination to become as open-ended as possible. The plaintiff's attorney has the option, finally, to reestablish the credibility of his witness by reexamination. These interrogations are formally regulated and require a great deal of skill and experience on the part of the attorneys. Such formal questioning of the witness is unknown to the continental European rules of procedure, even though cross-examination is common. Continental rules of procedure require the judge to interrogate the witness first. Frequently, the witness begins with a free narration. Then, after the judge has finished his interrogation, the attorneys of both parties may question the witness. All this is done in an informal manner, and almost any question is permitted. In some countries the interrogation of witnesses is, however, rather formalistic because it is generally limited to questions concerning allegations specified in the evidence judgment. But here too, there is a tendency for the court to allow questions at its discretion.

Purpose of cross-examination

Scientific examinations of witnesses are especially common in paternity and status proceedings with regard to blood-typing. These methods have now been so much improved that the suspicion of paternity may be definitely dismissed in many cases. In Germany and elsewhere, opinions based on biologic and hereditary evidence are used for these same purposes. The use of fingerprint and ballistics evidence, among other types, has become quite customary in criminal cases. In the United States, there are varying opinions about the admissibility of lie-detector tests as evidence. The results of such tests are not yet admissible in the continental European countries.

The hearsay rule. Hearsay is testimony based on what a witness has heard others say. The hearsay rule limiting this type of testimony is perhaps the most characteristic feature of the Anglo-American law of evidence. It has

also been said that, next to trial by jury, the hearsay rule constitutes the most important and original contribution of this system's practice.

Notwithstanding the obvious dangers involved in its use, free evaluation of the evidence furnished by hearsay testimony continues to be characteristic of continental European law. This somewhat surprising fact may be explained by reference to the historical development already traced here. Until the 19th century the medieval theory of formal evidence strictly prescribed when the judge had to be convinced by the testimony of a witness. Moreover, there was no jury in the continental countries to be protected by rules of evidence and therefore no need to introduce rules of hearsay. When the formal evidence theory was replaced by the requirement that the judge freely consider the evidence, his discretion naturally extended to hearsay testimony.

The creation of a body of rules for the exclusion of hearsay evidence was motivated by the arguments that such testimony could tend to mislead the jury, that the hearsay observer, unlike the legal witness, was not under solemn oath and was inaccessible to cross-examination, that such testimony furnished third-hand evidence, and that it violated the best evidence rule (the rule that the best version possible of a written document be submitted as evidence).

The exceptions to the rule against hearsay

Over the years, exceptions to the prohibition of hearsay testimony had to be permitted, however, and these have become so numerous that the opinion has sometimes been expressed that no exhaustive list of such exceptions could even be compiled. The judge must decide in each case whether testimony based upon hearsay is admissible under an exception to the rule—a further indication that regulations governing the admissibility of evidence are far more important in Anglo-American law than in continental law. The most commonly cited exceptions to the rule of hearsay relate to statements made by dead or absent persons, statements in public documents, and to confessions and admissions by parties.

Confessions and admissions. Confessions, as a source of evidence, are distinguished from admissions. Whereas a confession is a complete acknowledgement of guilt in criminal proceedings, an admission is a statement of fact in either a civil or a criminal case. In former times, the confession was considered the ultimate form of evidence. As soon as the accused confessed—often under duress—no further proof was required. In time, involuntary confession came to be rejected as evidence under English law, and the burden of proving that a confession was voluntary lay with the prosecutor. In the United States the federal rule that confessions are inadmissible if obtained while the defendant was unlawfully detained has not gone quite so far, though the law is still in a state of considerable flux. Involuntary confessions, however, are not admissible for any purpose under Anglo-American law. In continental European law, on the other hand, confessions of the accused are always freely considered by the judge.

Differences between criminal and civil proceedings regarding admissions result mainly from the adversary principle governing civil proceedings. In Anglo-American procedure, if one party in a civil suit admits facts contrary to his interest, such an admission is conclusive and obviates the need for further evidence on the point. The same result follows in German or Swedish courts. Under the Roman-based laws of such countries as France, Italy, and Spain, an admission made before the court is a form of evidence that leads to conclusive proof binding upon the court. But admissions made out of court are subject to free evaluation by the judge and do not exclude further evidence.

Party testimony. Oral testimony by the parties in civil proceedings was introduced in Austria in 1895. Norway followed suit in 1915, Denmark in 1919, Germany in 1933, and Sweden in 1948. Party testimony is generally heard in the same way as the evidence of witnesses, but there are some essential differences. In some countries, the interrogation of parties is a subsidiary source of evidence to be used only when all other means have been exhausted; in others (*e.g.,* Norway, Sweden, Austria, Brazil), parties are heard before witnesses. In some countries, both parties must be heard; in others, only one party may be heard upon motion of the opponent. The judge decides whether the parties are to be heard; this contrasts with the procedure with witnesses, who are heard only after having been nominated by the parties.

In most cases, the parties do not have to confirm their testimony by oath, but the court may decree that one of the parties must swear. In Swedish law, for example, the parties must solemnly declare that they have told the truth.

Expert evidence. Expert witnesses must have specialized knowledge, skill, or experience in the area of their testimony. For the most part, they do not testify concerning facts but draw inferences from them. With a few exceptions, they are treated in Anglo-American law as ordinary witnesses and are brought before the court by the parties in the same manner as other witnesses. Although ordinary witnesses are generally allowed to testify only concerning facts and not to express opinions, an exception to this rule is made for the expert, who must, of course, be allowed to give his opinion.

Qualifications of an expert witness

Generally speaking, anyone with special knowledge may be an expert in his respective field. In Anglo-American law, the expert is designated by the party, while in continental European law the court decides who may be an expert, generally selecting from a list on file in the court so as to guarantee that the experts designated are impartial. Experts may not, therefore, be cited by the parties.

The oral interrogation of experts is customary in Anglo-American law and proceeds, with a few exceptions, under the same rules for the interrogation of ordinary witnesses.

Under continental rules of procedure, on the other hand, expert opinions are generally given in written form. Experts are allowed a rather wide scope of discretion, especially when the opinion involves scientific findings that often cannot be checked by the judge. But under some continental European rules, the parties or their attorneys may request that the experts testify before the court to defend their written opinion and tell how they arrived at it.

Documentary evidence. Documentary evidence is in many respects considered better than the evidence furnished by witnesses, about which there has always been a certain amount of suspicion. Documentary evidence differs considerably from the evidence of witnesses and is dealt with under special rules.

Criteria for establishing the authenticity of documents are only important if authenticity is contested. This is often impossible, however, if a presumption favouring the authenticity of a public document exists—which it frequently does under continental European law. Under Anglo-American law, a party may serve the adversary with a written request to corroborate the authenticity of any relevant document. Direct evidence of authenticity may be gotten through the testimony of persons who signed the original documents. This is often impossible, however, and in this case circumstantial evidence is permitted. In some civil-law countries, documents are proved genuine by special proceedings. In other continental European countries, a document may be proved genuine by any type of evidence.

The obligation to present documents in the Anglo-American system derives from the best evidence rule. If the original document is in the hands of a third person or the opponent, the party that must supply proof can ask the court for a writ of *sub poena duces tecum* compelling the third party to produce the document in court. If the original is not produced after this, second-hand evidence of its existence is then permitted. In continental law, there is no similar obligation to produce documents. The adversary or third persons can only be ordered to do so if there is a positive obligation under the substantive law. Among European countries, only Sweden has developed any extensive obligation for the parties to produce documents.

Proof of documentary evidence

Extrinsic proof of the contents of documents in Anglo-American law is admitted only in special cases, since oral evidence is inadmissable to vary, contradict, or add to the terms of a written agreement—a rule that makes many documents conclusive as evidence. The method of Anglo-American law in this particular area is consequently neg-

ative, since evidence outside the content of the document is in principle not admissible. Continental law follows the medieval method, by attributing a certain value as evidence to particular documents, which is binding on the judge.

The consideration of documentary evidence by the judge therefore tends to be restricted, since the document itself furnishes conclusive proof if evidence by reference to facts outside the document is inadmissible. In most continental laws, judges are bound by presumptions in this respect, and only in Swedish law are there no provisions restricting free judicial consideration of documentary evidence.

Real evidence. The remaining form of evidence is so-called real evidence, also known as demonstrative or objective evidence. This is naturally the most direct evidence, since the objects in question are inspected by the judge or jury themselves. Problems arise in this area over who is obliged to present objects for inspection or to actually undergo inspection. The use of the jury system in Anglo-American law has made it necessary that any real evidence be shown to be both relevant and completely genuine before it may be admitted as proof. The exhibit of real evidence may sometimes be directly connected with the case (for example, when a weapon is shown to the court), or it may involve something used to illustrate testimony, as, for example, a model or skeleton to clarify testimony about an injury. In any case, real evidence may not be accepted as legal proof unless it is authenticated by the testimony of witnesses. (H.N.)

Conflict of laws

Both civil and criminal procedure may have a variety of international aspects: the plaintiff and defendant may be foreign citizens or may reside outside the country of the court (called the forum state); evidence may have to be taken in a foreign country; or, finally, a decision rendered in one country may have to be enforced in another. Growing international activities, primarily for business purposes but also of a private nature, have decisively increased the practical relevance of these international aspects of procedural law.

Legal problems arising from the international aspects of civil and criminal proceedings are a result of the multiplicity of different sets of courts and different systems of law in the world. Each nation maintains its own set of courts in complete independence of every other nation, and each nation has its own set of laws, written or unwritten. The rules and provisions that deal with the international aspects of various national legal systems are called the law of conflict of laws.

DIVERSITY OF LAWS

Diversity of national and provincial laws

Diversity within countries. While in such countries as France, Sweden, Peru, or Japan one single system of law obtains for the whole country, diversity exists in many others, especially nations organized upon a federal pattern, such as the United States, Australia, Canada, and, to a minor degree, Germany, Switzerland, and Mexico. The law of Illinois is not the same as that of New York, Louisiana, or Indiana; that of Quebec differs from that of Ontario or Newfoundland; that of Chihuahua is not quite the same as that of Michoacán. In Germany and Switzerland the systems of private law are by and large uniform, but minor differences still exist among the *Länder* of Germany and among the Swiss cantons.

Even in countries whose political structure is of the unitary rather than the federal pattern, differences can be found. In the United Kingdom, for example, considerable differences exist between the laws of England, Scotland, the Isle of Man, the Channel Islands, and Northern Ireland.

Diversity of laws develops where a country is divided, as, for example, in Korea. Where a new country is formed, or where territory is annexed, legal unity may not be brought about at the same time. After the reannexation of Alsace-Lorraine by France in 1920, for example, German private law remained in effect there for a good number of years; and when after World War I Poland was formed out of parts of old Russia, Germany, and Austria, legal uniformity was not brought about until after the end of World War II.

Diversities of law within one country may also exist on an ethnic or religious basis. Such a situation has commonly existed in most countries of the Middle East; the laws concerning matters of the family, including succession upon death, remain different in India for Hindus, Muslims, Parsees, Buddhists, and other sects, and in Lebanon or Israel for Muslims, Jews, and the various groups of Christians. In the United States and Canada, American Indians are in several respects subject to their own tribal laws.

Diversity between countries. Because of the spread of Western civilization over the entire planet, the laws of modern nations present a considerable measure of similarity, at least with respect to business transactions between individuals and private enterprises. Owing to the endurance of social traditions or religious convictions that are still quite different in many parts of the world, there is much less harmony between the rules on personal status, family matters, and succession. The same is true for the rules of criminal law. In addition, economic regulations differ considerably, as is indicated by the contrasts between free-market economies and planned economies.

Rules on the conflict of laws. Wherever there is diversity of laws, be it within or between countries, rules are required that must deal with, and seek to mitigate, the consequences of that diversity. Although terminology is not uniform, in most countries these rules are generally called conflict of laws. However, within this broad field, which may cover all branches of the law, it is important to distinguish between two principal branches of the conflict of laws: private international law (dealing with civil procedure) and international criminal law.

PRIVATE INTERNATIONAL LAW

Public and private international law

The name private international law, which is generally used in countries of European-continental tradition, and occasionally also in the United Kingdom, seems to indicate that it is a part of international law—that is, that system of law that is superior to all sovereign states and that, at least in theory, is uniform throughout the world. This view was commonly held for many centuries, and when the name private international law was coined in the 19th century it was meant to signify that the supranational body of international law consisted of two parts, public and private international law. While the former would determine the proper conduct of sovereign nations toward each other in both peace and war, the latter would, in a uniform way, tell all nations in what cases their courts ought or ought not to take jurisdiction, under what conditions foreign judgments were to be enforced or otherwise recognized, and in what cases the laws of one nation were to be applied rather than those of another.

Since the latter part of the 19th century, however, such a view has been considered an ideal rather than a true description of reality. Today, it is generally recognized that each nation determines not only what is to be its substantive law (its law of property, contracts, torts, family relations, succession, corporations, etc.) but also in what cases its courts are to have jurisdiction, under what conditions foreign judgments are to be recognized, and which country's law is to be applied in any particular case.

As on other matters, nations may, of course, conclude treaties, bilateral or multilateral, in which they assume in relations with each other the duty to deal with certain problems in an agreed way. Treaties of such a kind have been concluded between numerous states, especially among countries of Latin America and of continental Europe. The creation of various regional associations in western and eastern Europe, in the Middle East, and in Latin America has led to the conclusion of new multilateral conventions between the member states of these unions. The United States has concluded many bilateral treaties granting substantive or procedural rights to the citizens of each contracting state within the territory of the other. The countries of the Commonwealth are parties to numerous treaties with one another and with other nations, concerning foreign judgments and mutual rights of owning, disposing, and taking of property. In those

numerous areas not covered by treaties, the rules of the conflict of laws of each nation are relevant. These rules differ from country to country since each state is sovereign in fixing and amending them. Even in France, Germany, or Latin America, where the bulk of private law is contained in codes and other statutes, the statutory provisions on private international law are fragmentary, and for large parts of the field the law must be sought in the decisions of the courts. In all countries the writings of scholars have been of considerable influence.

Among the rules of private international law, three important issues of international civil procedure arise in practice: (1) the problem of jurisdiction—that is, under which circumstances a case may be brought before the courts of a particular country or province; (2) international elements in the various stages of a judicial proceeding; and (3) the recognition and enforcement of foreign judicial decisions—that is, what weight, if any, is to be given in one country or province to the judgments and decisions of the courts of other countries or provinces.

Jurisdiction. If a person wishes to bring a civil lawsuit against another, he might conceivably bring the action in any country of the world. If, however, a citizen and resident of the United States, for example, were to sue a citizen and resident of Canada in Panama, a judgment obtained in Panama would be of no use to him unless the Canadian owned property in Panama that, if he did not pay, the U.S. citizen might attach there, or if the Panamanian judgment could be enforced in such other country or countries in which he happened to hold property. For this practical reason the problem of where to bring suit is thus tied up with that of the enforceability of foreign judgments. Even if a judgment might be of practical value to the plaintiff, however, he might find that the courts of the country in which he wished to bring his action would not receive it. As a matter of fact, all countries have limited their jurisdiction—that is, the scope of actions that they allow their courts to handle. Countries do not wish their courts to deal with lawsuits with which they have no proper contact, which might clog the calendars of their courts, or against which it would be unfair to compel a person to enter a defense on pain of having judgment by default rendered against him. Each country determines for itself when its courts should decide a civil lawsuit.

The limits of jurisdiction in civil suits

In composite countries, such as the United States, the United Kingdom, Canada, and Switzerland, rules also are necessary to determine in which of the several constituent states, provinces, or other parts a civil lawsuit may be brought. In some countries (for instance, Germany) this determination is made by the national law. It may be left, however, to each of the constituent states or provinces to determine for itself the scope of litigation that it will allow its courts to decide. Such, at least on general principle, is the situation in the United States, where the state's freedom of determination is limited, however, by the "due process" clause of the Fourteenth Amendment to the federal Constitution, which in effect prohibits the state from exercising civil jurisdiction where it would be grossly unfair to do so. In the countries of the Commonwealth, the jurisdiction of the courts is also determined for each constituent part by its own law, but the principles of such determination do not differ widely from one another.

As a general principle, most countries or states agree that a case may be tried in their courts if both parties have consented to their jurisdiction. The plaintiff's consent simply appears from his commencing his action in the country or state in question; the consent of the defendant is presumed when, rather than objecting to the jurisdiction, he confesses judgment or begins to litigate on the merits of the controversy. Some countries, nevertheless, close their courts to a litigant whose case has no more substantial connection with them than the parties' consent. French courts, for instance, will not try a lawsuit between foreigners unless it arises out of a controversy that has some real connection with France, such as the breach of a contract to be performed in France, or a tort committed in France, or title to land situated in France. As another example, the courts of New York regard themselves as an "inconvenient forum" for suits between nonresidents concerning a tort committed outside New York. With few exceptions, Anglo-U.S. courts will not try controversies concerning title to, or trespass upon, land that is situated outside the state.

Generally, however, the problem of jurisdiction does not become acute unless the defendant objects to having the case tried in the country or province of the plaintiff's choosing, or unless he fails to appear. Different approaches to this problem of jurisdiction are followed in the continental European countries of the civil-law tradition and in those of the common-law, or Anglo-American, tradition. The former start from the idea that the proper place for a person to be sued is his domicile or residence. Apart from this principal venue, however, several others are available. For example, contentions over title to land must be sought where the land is situated. A suit arising out of an alleged tort may be brought in the place where the tort is alleged to have been committed, and a suit based upon breach of contract may be brought in the place in which it is alleged that the alleged contract was to be performed.

Some countries—for instance, Germany—allow an absent defendant to be sued in their courts if he owns any property within the country. France keeps its courts open for suits of any kind brought by a French national against a foreigner. A large number of countries, including those adhering to the common-law tradition, allow a civil suit to be commenced by the attachment of property owned within the territory, the enforcement of a default judgment obtained being limited, however, to the assets thus attached.

In their general approach to the problem of jurisdiction, the common-law countries still proceed from the long-obsolete notion that a civil suit could be commenced only by the defendant's arrest by the sheriff. Consequently, an action can still be brought in any place in which the defendant is personally served with process, even though he may be there only for a few minutes to change airplanes. In modern times it has come to be widely held, however, that personal service upon the defendant is no longer an indispensable requirement of jurisdiction and that an individual may be sued in the country or state of his residence, even if the summons is not personally pressed upon him. A corporation can always be sued in the country or state in which it has been incorporated.

It is required, however, that an honest effort be made to give the defendant actual notice that a lawsuit is about to be brought against him. The mere publication of the summons in a newspaper or at the bulletin board of the court is not sufficient unless the address or identity of the defendant cannot be ascertained upon a reasonable effort.

States of the United States are now coming to allow their courts to exercise jurisdiction in cases having almost any kind of contact with the state. Generally, a corporation may be sued in any state in which it is simply "doing business," even though the case in question is totally unconnected with the state.

Special rules for judgments in rem

In both civil-law and common-law countries special rules apply for suits in which the plaintiff aims at a "judgment in rem." Rather than ordering a defendant to pay a certain sum of money or ordering him to do, or not to do, a certain act (such as deliver a deed to a piece of land or refrain from using a trademark), a judgment in rem produces by its own effect a change of the legal situation (for instance, the foreclosure of a mortgage, the removal of a cloud on a title to land, the dissolution of a marriage, the creation of an adoptive parent-child relationship). Lawsuits aiming at the court's changing the title to a piece of land can universally be brought nowhere but in the country or province in which the land is situated. Actions arising out of transactions connected with shipping can generally be brought in the port in which the ship in question happens to find itself. In the United States a suit for divorce can be brought in the state of the plaintiff's domicile or residence, for the establishment of which periods varying between a few weeks and several months in length are prescribed. In the British countries the traditional rule of exclusive jurisdiction of the domicile of the husband is weakening. Civil-law countries generally keep their divorce courts open to their nationals even if they reside abroad.

Proceedings. Even if a civil action can be properly brought before a court having jurisdiction, international aspects may have an impact on the course of the proceedings.

Deviations from a purely domestic proceeding may arise especially if one of the parties resides outside of the forum state. In order to commence an action, the plaintiff's complaint must be served upon the defendant. The question then arises how such service can be effected if the action is to be brought, for example, before an English court when the defendant resides in France. In this case, formal service of judicial documents is entrusted to state officers or bailiffs and is therefore regarded as an act of national sovereignty. Since no court or state authority may act outside the area or state of the court, service in another state requires assistance by public authorities in that state.

Judicial assistance

The authorities, especially the judicial authorities of the state in which the defendant resides, must be requested to assist. Such international judicial assistance is usually rendered on condition of reciprocity; *i.e.*, only if the requesting state is prepared to honour a request for similar assistance. More certainty exists if both the requesting and the requested state are parties to a multilateral treaty on service of judicial documents.

A related issue will arise if the plaintiff cannot raise the funds for bringing his action or pursuing the proceedings. If he is a citizen of the forum state, he may be entitled to proceed *in forma pauperis* ("in the manner of a pauper"—*i.e.*, exempt from the usual costs of proceedings) or to obtain legal aid. However, the forum state is often not willing to bestow this benefit upon foreigners or persons residing abroad. Multilateral conventions seek to remove these difficulties by facilitating "international access to justice."

Another major difficulty to be overcome is in taking evidence outside the forum state when, for example, a witness living in Australia has to be heard in a divorce suit pending before a German court. The problem and the solutions that have evolved are quite similar to those described above for the formal service of judicial documents outside the forum state.

Foreign judgments. If a creditor has obtained against his debtor a judgment for $1,000 in Mexico or in Michigan, and his debtor does not have sufficient property in that country or state, can he enforce it in Illinois, where the debtor owns land, keeps a bank account, or owns other assets? If someone has brought and lost a lawsuit in New York, can he start it all over again in California or in Peru? If the marriage of Mr. and Mrs. Smith has been terminated by a decree of divorce of a court in Nevada, or by an act of the parliament of Canada, or by the order of a district governor in Norway, and Mr. Smith wishes to remarry in Wyoming or in South Africa, will he be given a license? If he remarries will his new marriage be valid or does he have to go to jail as a bigamist? If a citizen of the United States residing in Wisconsin adopts a child of German parents residing in Germany and the adoption has been confirmed by a court of Wisconsin, will the child inherit on the adopter's death a piece of land situated in Indiana or an account in a bank in Germany or in Switzerland?

The recognition and enforcement of foreign judgments

Unless countries have bound each other by treaty mutually to enforce their civil judgments, each country is free as to whether or not, and, if at all, under what conditions, it wishes to enforce or otherwise recognize foreign judgments of the types indicated by the questions above. The attitudes of the several countries vary considerably in this respect, and the enforcement of foreign money judgments is not the same as the recognition of a judgment as a bar to the starting of a new suit all over again (res judicata effect), or the recognition of the termination of a marriage by a decree of divorce or of other changes of private legal relationships brought about by judicial act.

If, for example, an Illinois judgment for money is not promptly paid by the debtor, it can be enforced in Illinois by the attachment and sale of his property, the proceeds being turned over to the creditor. Such enforcement is generally the task of a public officer, such as a sheriff, who is empowered, where necessary, to break resistance with physical force. Although a sheriff knows well enough the looks of a judgment of his own country or province, he cannot be expected, or even allowed, to go into action simply upon the basis of a paper purporting to be the judgment of a foreign country with whose judicial system, language, or even script he cannot be expected to be familiar. For the protection of the citizen as well as of himself, it is indispensable that, before the sheriff or other enforcement officer goes into action, the foreign judgment be transformed into a domestic one. Some countries, such as The Netherlands or Sweden, simply limit enforcement to domestic judgments. Even if the creditor has obtained a judgment abroad, he must start regular proceedings all over again, and the only advantage that the foreign judgment provides for him lies in the fact that the Dutch or Swedish court will be inclined to regard it as good, although in no way conclusive, evidence that his claim is well founded. In most other countries, however, a domestic judgment will be supplied by a domestic court without a reopening of the dispute about the merits of the creditor's claim. All that the domestic court will inquire into is the regularity of the proceedings in which the foreign judgment was obtained. For this transformation of a foreign into a domestic judgment, the majority of the civil-law countries provide a kind of special proceeding (exequatur) that is supposed to be, but is not always, simpler and less expensive than an ordinary civil lawsuit. In the common-law countries it is necessary to bring upon the foreign judgment an action that in outward form is a regular civil lawsuit but that is, at least in the normal case, simple and speedy. In the United Kingdom and the Commonwealth a simplified mode of domestication is furnished by agreements and statutes providing, in certain cases, for the simple registration in one law unit of judgments rendered in another. In the United States a similar method exists in the relations between those states that have adopted the Uniform Enforcement of Foreign Judgments Act. In the European Community and in Scandinavia, multilateral treaties oblige signatory states to recognize and enforce judgments from other member states and provide for a simplified procedure for the domestication of the foreign judgment.

When a foreign judgment is not sought to be enforced by attachment of the debtor's property or similar measures, but when its res judicata effect is raised as a defense in a domestic lawsuit, or when the question is that of recognition of its law-changing effects, such as the termination of a marriage by a decree of divorce, it would seem to be unnecessary to require the formal transformation of the foreign into a domestic judgment by any special proceedings. Some countries (for instance, Italy and, to a more limited extent, France) nevertheless require such formal domestication for judgments purporting to affect the personal status of their nationals.

In the United States the Constitution provides that "full faith and credit shall be given in each state to the public acts, records and judicial proceedings of every other state." Under this clause the states, and by statute the territories, are obliged mutually to enforce their money judgments and to recognize the res judicata and law-changing effects of their judicial acts, provided the state by which the judgment was rendered was acting within the scope of its jurisdiction as defined by the Supreme Court of the United States. The only other defenses that might be raised are grave irregularity of the proceedings in which the judgment was obtained and, in certain cases, lack of finality.

Prerequisites for recognition of foreign judgments

In countries that follow the general principles of the common law, a foreign judgment usually is willingly enforced and otherwise recognized unless (1) the country by which it was rendered lacked jurisdiction according to the notions prevailing in the place where recognition is sought, or (2) the proceedings in which the judgment was obtained were tainted with fraud or were otherwise grossly unfair, or (3) the recognition or enforcement of the foreign judgment would seriously interfere with an important public policy of the country or state where recognition or enforcement is sought. In addition to these requirements, most civil-law countries (except, of course, those few in which foreign judgments as such are not enforced at all)

also demand reciprocity with the country seeking to have its judgment recognized.

Nowhere will a foreign judgment be enforced or recognized unless the country by which it was rendered had jurisdiction to do so under the notions obtaining where recognition is sought. These limits are sometimes wider, however, than those that a country will concede to others for the exercise of their jurisdictions. Whereas France, for instance, holds its courts open for all suits of a Frenchman against a foreigner, a U.S. or English court will not recognize a default judgment obtained in such an action unless the defendant was served with process in France or was a resident of France or had some other contact with that country that justifies his being sued in France.

In matters affecting personal status, especially divorce, civil-law countries generally recognize judgments rendered by the courts of the country of which the parties are nationals. Under the common law of England a decree of divorce will not be recognized unless it was rendered by the state of the domicile of the husband. After World War II, however, there were enacted in some parts of the Commonwealth statutes under which a wife living separately from her husband may also sue for divorce in the country or province of her residence, and a decree thus obtained is likely to be recognized in the other parts of the Commonwealth. Since 1971, the United Kingdom even recognizes a foreign divorce decree rendered in the country of which one of the spouses is a national.

In the United States the Supreme Court has determined that a divorce granted in one state must be recognized in all others if the state by which it was granted was the state of the true residence of the plaintiff or if the defendant actually participated in the proceedings without contesting the plaintiff's allegation of residence.

INTERNATIONAL CRIMINAL LAW

This young and less-developed branch of the conflict of laws has seen a tremendous growth in modern times. The growing importance of international criminal law is due, first, to the increased international mobility of people, giving rise to more criminal acts with a foreign element, such as traffic violations involving foreigners as offenders or victims. In addition, shrewd criminals and organized crime have discovered the increasing gap between the territorial limitations of police, prosecution, and court powers on the one hand and the easy and quick trans-border communication and movement of persons and assets on the other. The more important general issues of modern international criminal law are comparable to those of the conflict of laws of civil procedure and are therefore of three types: (1) jurisdiction—*i.e.,* the question of which authorities of which country may prosecute a criminal and bring him into court; (2) the international aspects of a criminal court proceeding; and (3) the recognition and enforcement of foreign judgments in criminal matters—*i.e.,* the value attached in state B to a judicial decision rendered in state A. Because of the marked differences between criminal and civil procedure, the problems and solutions of international criminal law differ from those of private international law in many important respects.

Three principal issues

Jurisdiction. May a public prosecutor start a prosecution or may a criminal court open a judicial proceeding dealing with an offense that involves foreign elements? For instance, may a Swiss investigative or judicial authority open proceedings against a Frenchman? Does it make a difference whether this alleged offender resides in Switzerland, in France, or in Italy? Is it relevant whether the alleged act took place in Switzerland or outside Switzerland? Should one distinguish between a traffic accident, a theft, and a murder?

The basic principle for determining jurisdiction in criminal matters is that the authorities and courts of the state or province in which the offense was committed are competent to investigate and adjudicate it. This so-called principle of territoriality can be justified by both general and specific considerations. It is an important aspect of sovereignty that the authorities of a territory are responsible for preserving law and order in their area by protecting the integrity of the inhabitants and of their property against attacks and by punishing offenders. In addition, it is most practicable to investigate and adjudicate offenses where they have occurred because local circumstances can easily be taken into account, and witnesses and other means of evidence usually are located at or near the place of commission. Territoriality is generally recognized as the defining principle of jurisdiction in international criminal law. Some countries, especially in the English-speaking world, even regard it as the exclusive basis of jurisdiction. Other countries extend the basic rule by a few narrow and specific additions for particular offenses, such as drug offenses, terroristic acts, or war crimes.

The principle of territoriality

Two conventions in force for some members of the Council of Europe empower a contracting state A, on whose territory a person residing in contracting state B has committed an offense, to request the authorities of B to prosecute and adjudicate that offense under its criminal law. Paradoxically, these conventions were concluded in the interest of offenders, because they allow the authorities in the country of the offense to release the offender and request the country of residence to take over the prosecution and trial. Without the conventions, the police in the country of the offense would be compelled either to ask for securities or, in aggravated cases, to arrest the offender, at least until the trial was over and possibly even until he had served the sentence imposed for the offense.

Jurisdiction in criminal matters also has a personal aspect. Criminal trials, as distinct from trials of civil actions, will usually be continued only if the defendant appears before the court. His presence is almost indispensable for a fair trial. Since, for instance, a U.S. court or police officer cannot arrest a suspect outside the United States, the assistance of other countries for seeking, arresting, and delivering an alleged offender is required. For this purpose, bilateral and multilateral treaties provide, under certain conditions, for the extradition of suspects to the requesting country, which is usually the place where the offense was committed. One important restriction of the duty to extradite on which many countries insist protects nationals of the country requested; thus, the German constitution prohibits the extradition of Germans to foreign countries. The purpose of this prohibition is not to make Germans immune from criminal trials for offenses committed abroad but to ensure them a trial before a German court. In order to achieve this, jurisdiction of German courts is extended beyond the basic principle of territoriality to include all offenses committed by Germans, even those committed abroad.

Proceedings. In a criminal proceeding, there are fewer complications caused by foreign elements than there are in civil proceedings, usually because the defendant is present at the trial. Nevertheless, he may require legal aid if he is too poor to pay for his defense. Or a foreign defendant or foreign witness may need an interpreter. These problems of access to the court are usually settled under the forum country's domestic rules of criminal procedure.

International cooperation is especially necessary where evidence has to be taken in a country outside the forum state. The most reliable way to obtain the necessary judicial assistance of foreign states is through treaties concluded between states. A convention of the Council of Europe, for example, makes judicial assistance in criminal matters obligatory for 18 member countries. Its opening provision binds the contracting states to afford one another, within the terms of the convention, "the widest measure of mutual assistance in proceedings in respect to offenses."

Foreign judgments. Generally speaking, the issues arising out of the international effects of a criminal judgment are comparable to those involved in determining the weight to be attached to a foreign judgment in civil matters. If a person has been convicted and sentenced in Turkey but before the end of the trial has fled the country, can he be brought back to Turkey to serve his sentence there? If he cannot be brought back to Turkey, will Italy, his present country of abode, be willing to enforce the payment of a penalty or even a prison sentence? Will the answer be different if the offender is an Italian residing in Italy? Less drastic, but certainly no less relevant, for the defendant is the question whether after an acquittal in

France he may be prosecuted and convicted in Denmark for an offense based upon the same facts.

Enforcing a sentence imposed by a foreign court

The stronger the intended effects of a foreign judgment, the less are states willing to give effect to it. The strongest effect of a criminal judgment is the enforcement of a sentence imposed by the court, and, contrary to the treatment of foreign judgments in civil matters, states are generally quite unwilling to enforce foreign prison sentences or even penalties. This situation can only be remedied by the conclusion of bilateral or multilateral treaties, of which many exist. The most common type provides for extradition of the convict to the country in which he has been convicted. Nineteen member countries of the Council of Europe, for example, are bound by a multilateral convention on mutual extradition. A more recent type of convention allows a contracting state A, where a resident of contracting state B has been convicted, to request the latter state to take over the person and make him serve in B the prison sentence imposed in A. Such a request usually serves the interests of the offender, especially if the penitentiary system of state A is not as well developed as that of state B. Such a transfer also helps to preserve the social and familial contacts of the offender and to facilitate his social rehabilitation. Usually state B will agree to enforcement of a foreign criminal sentence only if the offense is one that is punishable under the law of B also.

Lesser effects than enforcement of sentences are usually more acceptable to foreign states. The fact that a suspect has been acquitted or has been convicted and has served his sentence in one country is, in many other countries, recognized as a bar to a new prosecution or conviction for the same offense. In this way the double jeopardy rules of many national laws, barring a second proceeding on the same facts, is extended to the international level.

An earlier conviction for an offense in state A may have an adverse effect for a defendant who is later convicted for an offense of the same type in state B. This is true if, under the law of B, penalties may or must be increased for repeated offenses of the same type. In this case, too, a domestic rule (that penalizing recidivism) is extended to the international level, this time to the disadvantage of the defendant. (M.Rh./U.M.D.)

BIBLIOGRAPHY

Historical growth of procedural law. See LEOPOLD WENGER, *Institutes of the Roman Law of Civil Procedure,* rev. ed. (1940, reprinted 1986), the classic on the topic; JOHN P. DAWSON, *A History of Lay Judges* (1960), in disagreement with some of Wenger's conclusions, although not limited to Roman law, and *The Oracles of the Law* (1968, reprinted 1986), a complementary work on the history of professional judges; ARTHUR ENGELMANN *et al., A History of Continental Civil Procedure* (1927, reprinted 1968; originally published in German, 1901), the classic text in English; ROBERT W. MILLAR, *Civil Procedure of the Trial Court in Historical Perspective* (1952), which discusses the history of Anglo-American trial procedure; and THEODORE F.T. PLUCKNETT, *A Concise History of the Common Law,* 5th ed. (1956), a general treatment of English legal history.

Civil procedure. M. CAPPELLETTI (ed.), *Access to Justice,* 40 vol. in 6 (1978–79), is a cooperative work by several contributors discussing problems and possibilities involved in attempts to secure access to justice for all; M. CAPPELLETTI (ed.), *Civil Procedure,* vol. 16 of *International Encyclopedia of Comparative Law* (1973–), issued in fascicles, is a scholarly discussion of all aspects of civil procedure by contributors from many countries; and CHARLES E. CLARK, *Procedure: The Handmaid of Justice* (1965), is a collection of significant essays.

Elements of civil procedure. American civil procedure is treated in KEVIN M. CLERMONT, *Civil Procedure* (1982); JACK FRIEDENTHAL, MARY KAY KANE, and ARTHUR R. MILLER, *Civil Procedure* (1985), a substantially more detailed discussion, mainly intended for students; FLEMING JAMES, JR., and GEOFFREY C. HAZARD, JR., *Civil Procedure,* 3rd ed. (1985), a discussion in some depth; JOSEPH H. KOFFLER and ALISON REPPY, *Handbook of Common Law Pleading* (1969), a discussion of earlier procedure; and HARVARD LAW REVIEW, *Essays on Civil Procedure* (1961, reissued 1967), a collection of valuable essays in the field.

Civil procedure in other countries is treated in M. CAPPELLETTI and JOSEPH M. PERILLO, *Civil Procedure in Italy* (1965), one of the best available treatments in English; RUTH BADER GINSBURG and ANDERS BRUZELIUS, *Civil Procedure in Sweden* (1965), an indispensable work; TAKAAKI HATTORI and DAN FENNO HENDERSON, *Civil Procedure in Japan* (1983), a valuable discussion of modern procedure, published in loose-leaf format; and WILLIAM B. ODGERS, *Odgers' Principles of Pleading and Practice in Civil Actions in the High Court of Justice,* 22nd ed. by D.B. CASSON and I.H. DENNIS (1981), a standard work on England's civil procedure.

Treatments of special topics in civil procedure include EDWIN M. BORCHARD, *Declaratory Judgments,* 2nd ed. rev. (1941), the classic work; M. CAPPELLETTI, *Procédure orale et procédure écrite: Oral and Written Procedure in Civil Litigation* (1971), a comparative study based on reports from several countries, with a summary in English; and ROBERT C. CASAD, *Res Judicata in a Nutshell* (1976), an introduction to problems of effect of judgments. (P.E.H./Ma.E.O.)

Criminal procedure. Texts on English criminal procedural law and practice include JOHN FREDERICK ARCHBOLD, *Pleading, Evidence, and Practice in Criminal Cases,* 42nd ed. edited by STEPHEN MITCHELL and P.J. RICHARDSON (1985); and CELIA HAMPTON, *Criminal Procedure,* 3rd ed. (1982), an introductory text. Criminal procedure in the United States is detailed in FRANCIS WHARTON, *Wharton's Criminal Procedure,* 12th ed. by CHARLES E. TORCIA, 4 vol. (1974–76), with annual cumulative supplements; WAYNE R. LAFAVE and JEROLD H. ISRAEL, *Criminal Procedure,* 3 vol. (1984), a standard textbook; JOSEPH G. COOK, *Constitutional Rights of the Accused,* 2nd ed., 3 vol. (1985–86), a treatise on procedural law under the U.S. Constitution; and JAMES E. BOND, *Plea Bargaining and Guilty Pleas,* 2nd ed. (1983), published in loose-leaf format. Texts on the law of criminal procedure in other countries include, for France, JEAN PRADEL, *Procédure pénale,* 4th ed. rev. and enl. (1987); for Italy, GIAN DOMENICO PISAPIA, *Compendio di procedura penale,* 3rd ed. (1982); and for Germany, JOHN H. LANGBEIN, *Comparative Criminal Procedure: Germany* (1977); and CLAUS ROXIN, *Strafverfahrensrecht,* 20th rev. ed. (1987).

(H.-H.J./T.We.)

The law of evidence. Texts outlining U.S. practice include C.T. MCCORMICK, *Law of Evidence* (1954); and GRAHAM C. LILLY, *An Introduction to the Law of Evidence,* 2nd ed. (1987). Works treating evidence law in England include G.D. NOKES, *An Introduction to Evidence,* 4th ed. (1967); and RUPERT CROSS, *Evidence,* 5th ed. (1979). For a comparative study of the laws of evidence in Germany, England, France, Italy, Spain, Sweden, and the former Soviet Union, see HEINRICH NAGEL, *Die Grundzüge des Beweisrechts im europäischen Zivilprozess* (1967).

For information on the law of evidence in Argentina, see LINO ENRIQUE PALACIO, *Manual de derecho procesal civil,* 7th ed. updated (1987); in Austria, HANS W. FASCHING, *Lehrbuch des österreichischen Zivilprozessrechts* (1984); in Brazil, ARRUDA ALVIM, *Manual de Direito Processual Civil,* vol. 1, *Parte Geral,* 2nd ed. rev. (1986); in France, PETER E. HERZOG and MARTHA WESER, *Civil Procedure in France* (1968); in Mexico, NICETO ALCALÁ-ZAMORA Y CASTILLO, *Estudios procesales* (1975); in Spain, VÍCTOR FAIRÉN GUILLÉN, *Estudios de derecho procesal civil, penal y constitucional* (1983); in Sweden, PER OLOF EKELÖF, *Rättegång,* 4th ed., 5 vol. (1974–80), with vol. 1 and 5 also available in a 6th ed. (1980, 1987); and in Germany, LEO ROSENBERG and KARL HEINZ SCHWAB, *Zivilprozessrecht,* 14th rev. ed. (1986). (H.N.)

Conflict of laws. *Private international law:* Literature on individual national systems of law is too numerous to be cited. Broad comparative treatments are offered by ISTVÁN SZÁSZY, *International Civil Procedure: A Comparative Study,* trans. from Hungarian (1967); and ERNST RABEL, *The Conflict of Laws: A Comparative Study,* 2nd ed., 3 vol. (1958–64), the only scholarly analysis of the conflict of law on a worldwide scale.

International criminal law: M. CHERIF BASSIOUNI, *International Criminal Law,* 3 vol. (1986–87); STEFAN GLASER, *Introduction à l'étude du droit international pénal* (1954), with a supplemental volume (1959), and *Droit international pénal conventionnel,* 2 vol. (1970–78); F. MEILI, *Lehrbuch des internationalen Strafrechts und Strafprozessrechts* (1910), the first classic; GERHARD O.W. MUELLER and EDWARD M. WISE (eds.), *International Criminal Law* (1965), especially ch. 1 and 4; DIETRICH OEHLER, *Internationales Strafrecht,* 2nd rev. and enl. ed. (1983); ANTONIO QUINTANO RIPOLLÉS, *Tratado de derecho penal internacional y internacional penal,* 2 vol. (1955–57); and EDWARD S. STIMSON, *Conflict of Criminal Laws* (1936).

(U.M.D.)

Propaganda

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

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

This article is divided into the following sections:

Propaganda and related concepts 171
- Connotations of the term propaganda
- Related terms
- Signs, symbols, and media used in contemporary propaganda

Evolution of the theory of propaganda 172
- Early commentators and theories
- Modern research and the evolution of current theories

The components of propaganda 174
- Goals
- Present and expected conditions in the world social system
- Present and expected conditions in subsystems
- The propagandist and his agents
- Selection and presentation of symbols
- Media of propaganda
- The reactors (audiences)
- Measurement of the effects of propaganda
- Countermeasures by opponents
- Measures against countermeasures

Social control of propaganda 178
- Democratic control of propaganda
- Authoritarian control of propaganda
- World-level control of propaganda

Bibliography 179

PROPAGANDA AND RELATED CONCEPTS

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

Agitprop: agitation and propaganda

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

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

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

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

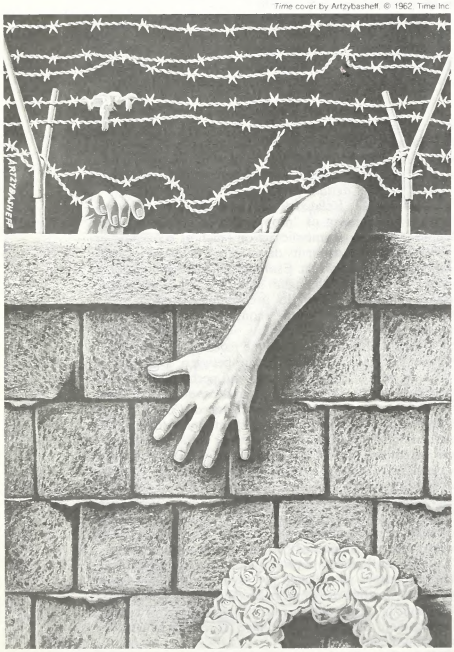
Brainwashing

Still another related concept is that of brainwashing. This term usually means intensive political indoctrination. It

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

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

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



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

The nature of symbols

A symbol is a sign having a particular meaning for a given reactor. Two or more reactors may of course attach quite different meanings to the same symbol. Thus, to Nazis the swastika was a symbol of racial superiority and the crushing military might of the German folk; to some Asiatic and North American peoples it is a symbol of universal peace and happiness. Some Christians who find a cross reassuring may find a hammer and sickle displeasing and may derive no religious satisfaction at all from a Muslim crescent, a Hindu cow, or a Buddhist lotus.

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

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

The nature of organizational media

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

EVOLUTION OF THE THEORY OF PROPAGANDA

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

The systematic, detached, and deliberate analysis of

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

The Greco-Roman tradition of rhetoric

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

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

Hindu and Chinese traditions

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

The spread of all complex political systems and religions probably has been due very largely to a combination of earnest conviction and the deliberate use of propaganda. This mixture can be detected in the recasting in various times and places of the legends of the Judaeo-Christian messiah, of heroes of the Hindu *Mahābhārata*, of the Buddha, of the ancestral Japanese Sun Goddess, of the lives of Muḥammad and his relatives, of the Christian saints, of such Marxist heroes as Marx, Engels, Lenin, and Stalin, and even in the story of George Washington and the cherry tree.

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

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

Sample surveys and market research

Large quantities of such information on consumers and voters are stored and statistically processed by computers and are drawn upon for nationwide and international advertising campaigns costing billions of dollars annually. Such advertising—including political advertising—occupies a very high percentage of radio and television time and of newspaper, magazine, and billboard space in countries where it is permitted. By conservative estimates $140,000,000 was spent in the U.S. presidential election of 1952, $155,000,000 in that of 1956, $175,000,000 in 1960, and $200,000,000 in 1964. On paid media the Republican Party was estimated to have spent more than $23,000,000 and the Democratic Party over $25,000,000, for their presidential and vice presidential candidates in 1984. Critics have argued that advertising expenditures on such a scale, whether for deodorants or presidents, tend to waste society's resources and also to preclude effective competition by rival producers or politicians who cannot raise equally large amounts of money. A rising tide of consumer resistance and voter skepticism is leading to various attempts at consumer education, voter education, counterpropaganda, and proposals for regulatory legislation.

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

and followers, as were Walter Lippmann's *Public Opinion* (1922) and *The Phantom Public* (1925).

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

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

Propaganda specialists

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

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

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

THE COMPONENTS OF PROPAGANDA

The contemporary propagandist employing behavioral theory tends to analyze his problem in terms of at least 10 questions:

1. What are the goals of the propaganda? (What changes are to be brought about? In whom? And when?)
2. What are the present and expected conditions in the world social system?
3. What are the present and expected conditions in each of the subsystems of the world social system (such as international regions, nations, lesser territories, interest groups)?
4. Who should distribute the propaganda—the propagandist or his agents?
5. What symbols should be used?
6. What media should be used?
7. Which reactors should the propaganda be aimed at?
8. How can the effects of the propaganda be measured?
9. By what countermeasures can opponents neutralize or suppress the propaganda?
10. How can such countermeasures be measured and dealt with?

In the present state of social science, this 10-part problem can be solved with only moderate confidence with respect to any really major propaganda campaign, even if one has a great deal of money for research. Yet if the propagandist is to proceed as rationally as possible, he needs the best answers that are available.

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

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

Particularism versus universalism

Present and expected conditions in subsystems. In many times and places in the past, the propagandist could profit handsomely by ignoring the welfare of a nation or the world and appealing to extremes of religious, racial, political, or economic fanaticism. This paid off very well, in the short run at least, within many subsystems. Today, however, this kind of propaganda can prove to be useless and even dangerous. The prudent propagandist has therefore to decide what mix of universalistic and particularistic symbolism will best serve his purposes at given times in

given places. The choice is never an easy one: parochial or class-conscious or national groups may be aroused to the highest passions; and they are numerous and diverse and often highly incompatible with one another and with the imperatives of the nation or the world.

Propaganda "fronts"

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

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

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

Image makers

> There are now four essential ingredients to a professionally managed political campaign: political polls, data processing, imagery, and money. The polls discover what the voter already believes, and data processing interprets and analyzes the depth of voters' attitudes. After that, an image of the candidate is tailored to meet the voters' demands and desires, and the whole package is then sold by massive expenditures of money in the advertising media, particularly television.
>
> The candidate has become relatively unimportant as long as he can be properly managed. The candidate must be bright enough to handle the material furnished to him, but not too intelligent, because there is always the danger that an intelligent candidate may come up with unpopular or controversial ideas of his own, and thereby destroy a carefully contrived campaign strategy. [Excerpt from a public address by Zolton Ferency, chairman and gubernatorial candidate of the Democratic Party of Michigan, June 1970.]

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

Predispositions and inducements

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

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

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

There is substantial agreement among psychoanalysts that the psychological power of propaganda increases with use of what Lasswell termed the triple-appeal principle.

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

Role models and parental symbols

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

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

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

Communication overload

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

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

Use of reference groups

For reasons just indicated, the most effective media as a rule (for messages other than the simplest of commercial advertising) are not the impersonal mass media like newspapers and television but rather those few associations or organizations (reference groups) with which the individual feels identified or to which he aspires to relate his identity. Quite often the ordinary man not only avoids but actively distrusts the mass media or fails to understand their messages; but in the warmth of his reference groups he feels at home, assumes that he understands what is going on, and feels that he is sure to receive a certain degree of emotional response and personal protection. The foremost reference group, of course, is the family. But many other groups perform analogous functions—for instance, the group of sports fans, the church, the trade union, the alumni group, the clique or gang, the Communist cell. By influencing the key members of such a group, the propagandist may establish a "social relay" channel that can amplify his message. By concentrating thus on the few, he increases his chances of reaching the many—often far more effectively than he could through a plethora of mass

meetings, paid broadcasts, handbills, or billboards and at much lower cost. Therefore, one important stratagem involves the combined use of mass media and reference-group channels—writing up materials for such media as news releases or broadcasts in ways designed specifically to reach certain groups (and especially their elites and leaders), who can then relay the messages to other sets of reactors.

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

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

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

Use of experiments

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

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

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

Use of interviews and observation techniques

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

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

Use of "content analysis"

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

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

Very elaborate means of coding and of statistical analysis have been developed by various content analysts. Some count symbols, some count headlines, some count themes (sentences, propositions), some tabulate the frequencies with which various categories of "events data" (newspaper accounts of actual happenings) appear in some or all of the leading newspapers ("prestige papers") or television programs of the world. Some of these events data can be counted as supporting or reinforcing the propaganda, some as opposing or counteracting it. Whatever the methodology, content analysis in its more refined forms is an ex-

pensive process, demanding long and rigorous training of well-educated and extremely patient coders and analysts. And there remains the intricate problem of developing relevant measurements of the effects of different contents upon different reactors.

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

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

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

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

SOCIAL CONTROL OF PROPAGANDA

The democratic ideal and propaganda

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

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

Equal time and equal space

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

World-level control of propaganda. One of the most serious and least understood problems of social control is above the national level, at the level of the world social system. At the world level there is an extremely dangerous lack of means of restraining or counteracting propaganda that fans the flames of international, interracial, and interreligious wars. The global system consists at present of a highly chaotic mixture of democratic, semidemocratic, and authoritarian subsystems. Many of these are controlled by leaders who are ill educated, ultranationalistic, and religiously, racially, or doctrinally fanatical. At present, every national regime asserts that its national sovereignty gives it the right to conduct any propaganda it cares to, however untrue such propaganda may be and however contradictory to the requirements of the world system. The most inflammatory of such propaganda usually takes the form of statements by prominent national leaders, often sensationalized and amplified by their own international broadcasts and sensationalized and amplified still further by media in the receiving countries. The only major remedy would lie, of course, in the slow spread of education for universalist

humanism. A first step toward this might be taken through the fostering of an energetic and highly enlightened press corps and educational establishment, doing all it can to provide the world's broadcasters, newspapers, and schools with factual information and illuminating editorials that could increase awareness of the world system as a whole. Informed leaders in world affairs are therefore becoming increasingly interested in the creation of world-level media and multinational bodies of reporters, researchers, editors, teachers, and other intellectuals committed to the unity of mankind.

BIBLIOGRAPHY. Annotated listings of books and articles of many times, countries, and languages, with respect to public opinion and the theory and practice of communication (including propaganda) appear in HAROLD D. LASSWELL, RALPH D. CASEY, and BRUCE LANNES SMITH (eds.), *Propaganda and Promotional Activities: An Annotated Bibliography* (1935, reissued 1969); BRUCE LANNES SMITH, HAROLD D. LASSWELL, and RALPH D. CASEY, *Propaganda, Communication, and Public Opinion: A Comprehensive Reference Guide* (1946); and HAROLD D. LASSWELL, DANIEL LERNER, and HANS SPEIER (eds.), *Propaganda and Communication in World History,* 3 vol. (1979–80). Further listings on general and international propaganda are in BRUCE LANNES SMITH and CHITRA M. SMITH, *International Communication and Political Opinion: A Guide to the Literature* (1956, reissued 1972). More recent periodical literature may be found in *International Political Science Abstracts* (bimonthly); *Psychological Abstracts* (monthly); and *Sociological Abstracts* (5/yr.).

General works of considerable significance include QUINCY WRIGHT (ed.), *Public Opinion and World-Politics* (1933, reissued 1972); LEON FESTINGER, *A Theory of Cognitive Dissonance* (1957, reissued 1967); KARL W. DEUTSCH, *The Nerves of Government: Models of Political Communication and Control* (1963); LEWIS ANTHONY DEXTER and DAVID MANNING WHITE (eds.), *People, Society, and Mass Communications* (1964); JACQUES ELLUL, *Propaganda* (1965, reissued 1973; originally published in French, 1962); LEONARD W. DOOB, *Public Opinion and Propaganda,* 2nd ed. (1966); and ANTHONY R. PRATKANIS and ELLIOT ARONSON, *Age of Propaganda: The Everyday Use and Abuse of Persuasion* (1992).

Analyses of the role of propaganda during World Wars I and II can be found in HAROLD D. LASSWELL, *Propaganda Technique in the World War* (1927, reprinted 1972); DANIEL LERNER, *Sykewar: Psychological Warfare Against Germany, D-Day to VE-Day* (1949, reissued as *Psychological Warfare Against Nazi Germany: The Sykewar Campaign, D-Day to VE-Day,* 1971); ALEXANDER L. GEORGE, *Propaganda Analysis: A Study of Inferences Made from Nazi Propaganda in World War II* (1959, reissued 1973); K.R.M. SHORT (ed.), *Film & Radio Propaganda in World War II* (1983); and RICHARD W. STEELE, *Propaganda in an Open Society: The Roosevelt Administration and the Media, 1933–1941* (1985). Studies of the role of propaganda in the conduct of foreign policy include DANIEL LERNER (ed.), *Propaganda in War and Crisis* (1951, reissued 1972); ROBERT T. HOLT and ROBERT W. VAN DE VELDE, *Strategic Psychological Operations and American Foreign Policy* (1960); and RALPH K. WHITE, *Nobody Wanted War: Misperception in Vietnam and Other Wars,* rev. ed. (1970).

The role of propaganda in the creation of the British empire is discussed in JOHN M. MACKENZIE, *Propaganda and Empire: The Manipulation of British Public Opinion, 1880–1960* (1984). Soviet propaganda efforts are discussed in FREDERICK C. BARGHOORN, *The Soviet Cultural Offensive: The Role of Cultural Diplomacy in Soviet Foreign Policy* (1960, reprinted 1976), and *Soviet Foreign Propaganda* (1964). Also worth reading is PHILIP SELZNICK, *The Organizational Weapon: A Study of Bolshevik Strategy and Tactics* (1952, reissued 1979). PETER KENEZ, *The Birth of the Propaganda State: Soviet Methods of Mass Mobilization, 1917–1929* (1985), thoroughly discusses the origins of the Soviet propaganda efforts. Chinese propaganda is examined in FREDERICK T.C. YU, *Mass Persuasion in Communist China* (1964).

The impact of propaganda upon individuals is considered in the following texts: ELIHU KATZ and PAUL F. LAZARSFELD, *Personal Influence* (1955, reissued 1965); IRVING L. JANIS *et al., Personality and Persuasibility* (1959, reprinted 1982); JOSEPH T. KLAPPER, *The Effects of Mass Communication* (1960); and GARTH S. JOWETT and VICTORIA O'DONNELL, *Propaganda and Persuasion,* 2nd ed. (1992).

The role of propaganda in creating political and symbolic realities is considered in HAROLD D. LASSWELL *et al., Language of Politics* (1949, reissued 1968); HAROLD D. LASSWELL, DANIEL LERNER, and ITHIEL DE SOLA POOL, *The Comparative Study of Symbols* (1952); and PHILO C. WASBURN, *Broadcasting Propaganda: International Radio Broadcasting and the Construction of Political Reality* (1992).

The relationship between propaganda and the political process is treated by BERNARD R. BERELSON, PAUL F. LAZARSFELD, and WILLIAM N. MCPHEE, *Voting: A Study of Opinion Formation in a Presidential Campaign* (1954, reprinted 1986); LUCIAN W. PYE (ed.), *Communications and Political Development* (1963); WILBUR SCHRAMM, *Mass Media and National Development: The Role of Information in the Developing Countries* (1964); JOE MCGINNISS, *The Selling of the President, 1968* (1969, reprinted 1988); and TED J. SMITH, III (ed.), *Propaganda: A Pluralistic Perspective* (1989). Propaganda aspects of diplomatic negotiation include HAROLD NICOLSON, *Diplomacy,* 3rd ed. (1963); and FRED CHARLES IKLÉ, *How Nations Negotiate* (1964, reprinted 1982).

Texts on legal aspects and social control of propaganda include HAROLD D. LASSWELL, *Democracy Through Public Opinion* (1941); L. JOHN MARTIN, *International Propaganda: Its Legal and Diplomatic Control* (1958, reissued 1969); JOHN B. WHITTON and ARTHUR LARSON, *Propaganda Towards Disarmament in the War of Words* (1963); B.S. MURTY, *Propaganda and World Public Order* (1968, reissued as *The International Law of Propaganda,* 1989); JOHN W. BURTON, *Conflict & Communication: The Use of Controlled Communication in International Relations* (1969); and TERENCE H. QUALTER, *Opinion Control in the Democracies* (1985), an advanced study. The sociological impact of propaganda is considered in MAUREEN HONEY, *Creating Rosie the Riveter: Class, Gender, and Propaganda During World War II* (1984).

Propaganda issues within the media include ALEX P. SCHMID and JANNY DE GRAAF, *Violence as Communication: Insurgent Terrorism and the Western News Media* (1982); and DAVID L. PALETZ and ALEX P. SCHMID (eds.), *Terrorism and the Media* (1992).

(B.L.S.)

Property Law

Like all law, property law seeks to articulate the principles, policies, and rules by which disputes are to be resolved and by which transactions may be structured so that disputes may be avoided. What distinguishes property law from other kinds of law is that the principles, policies, and rules of property law deal with the relationships between and among members of a society with respect to "things." The things may be tangible, such as land or a factory or a diamond ring, or they may be intangible, such as stocks and bonds or a bank account. Property law, then, deals with the allocation, use, and transfer of wealth and the objects of wealth. As such, it reflects the economy of the society in which it is found. Since it deals with the control and transfer of wealth between spouses and across generations, property law reflects the family structure of the society in which it is found. Since it deals with such fundamental issues as the economy and the structure of the family, property law reflects the politics of the society in which it is found.

This article outlines the major systems of property law that have existed historically and that exist today. The principal focus is on the two major Western systems of law that have become dominant in the industrialized world: the Anglo-American system, derived from the English common law, and the "civil-law" system, which was developed on the European continent on the basis of Roman law. The focus is particularly on the different ways in which these systems attempt to resolve the conflicts that result from their historical tendency to regard absolute individual ownership of property as normative.

For coverage of related topics in the *Macropædia* and the *Micropædia*, see the *Propædia*, sections 551 and 553, and the *Index*.

The article is divided into the following sections:

Definition and scope 180
  The problem of definition 180
  Property law and the concept of private property 181
    Etymology
    The Western tendency to agglomerate
    Basic tensions within the tendency
  Scope of the article 181
    Relation to other branches of law
    Geographic scope
History and theory in the West 182
  The anthropology of property 182
  The origins of the Western idea of property 182
    Rome
    England
    The Continent
    Explaining the origins
  Property law and theory in the early modern period 184
    The classical theories of property
    Possessive individualism and the law
    Marxism, liberalism, and the law
    The social situation
    The fundamental tendency revisited
  Recent history of the Western concept of property 186
Objects, subjects, and types of possessory interests in property 186
  Objects 186
    Classification of "things"
    Within and without commerce
    Possession of tangible things
    Possession of intangible things
    Movable and immovable property
  Subjects 189
    Single individuals
    Groups
  The concept of ownership 190
    Everything must have an owner
    Ownership and possession
    Ownership as the absolute right to possession
  Divisions of ownership 192
    Spatial divisions
    Temporal divisions
    Divisions as to rights, privileges, and powers
  Protection of property interests 195
    Public law protections of property
    Private law protections of property
Use of property interests 196
  Nuisance law and continental parallels 196
  Private land-use control: servitudes 197
    Easements and profits
    Real covenants
    Equitable servitudes
    Civil law
  Public regulation of land use 198
    Public nuisance
    Direct regulation
    Eminent domain
    Limitations on government action
Acquisition and transfer of property interests 200
  Original acquisition 200
    First possession
    Accession
    Adverse possession, prescription, and expropriation
    Privileges conferred by public authorities
  Derivative acquisition 200
    Contract and conveyance
    Registration and recordation
    Sales
    Gifts
    Succession
    Nonconsensual transfers
The Western concept of property: assessment 203
Bibliography 204

Definition and scope

THE PROBLEM OF DEFINITION

Property is frequently defined as the rights of a person with respect to a thing. The difficulties with this definition have plagued speculative jurists in the West for the better part of a century. Speaking of the relationship of a person to a thing is of limited usefulness in legal discourse because a thing cannot bring or defend a lawsuit. The law does not deal with rights, privileges, powers, their correlatives and opposites, in the abstract; it deals with relationships between and among people.

With this in mind one may redefine the law of property as the complex of jural relationships between and among persons with respect to things. It is the sum of rights and duties, privileges and no-rights, powers and liabilities, disabilities and immunities that exist with respect to things.

Definition of property law

What distinguishes property law from all other jural relationships is that the jural relationships of property law deal with things.

For purposes of this article, all tangible things are included within the realm of property law, even if a specific legal system denies the classification "property" to certain kinds of tangible things. Many, but not all, legal systems that recognize a separate category of property law also include within that category some intangible things, like stocks and bonds, but not other intangible things, like claims for compensation for wrongs (tort, delict). The definition of property law used here includes those intangible things that the legal system under discussion classifies as property.

This descriptive definition of property law makes it possible to say that there is no known legal system that does not have a law of property. A legal system may not have

a category that corresponds to property in Western legal systems, but every known legal system has some set of rules that deal with the relations among persons with respect at least to tangible things.

PROPERTY LAW AND THE CONCEPT OF PRIVATE PROPERTY

Etymology. The descriptive definition of property law adopted for this article is far removed from what the word *property* means in normal English usage: "an object of legal rights," or "possessions" or "wealth" collectively, frequently with strong connotations of individual ownership. The English word *property* derives either directly or through French *propriété* from Latin *proprietas,* which means "the peculiar nature or quality of a thing" and (in post-Augustan writing) "ownership." The word *proprietas* is derived from *proprius,* an adjective meaning "peculiar" or "own," as opposed to *communis,* "common," or *alienus,* "another's." *Proprius* itself is of uncertain etymology but is probably related to the Indo-European root that appears in Latin *pro, prae,* and *prope,* Greek *pró* and *prín,* and Sanskrit *pra.* The meaning of the root is suggested by the meanings of its derivatives: it is the core of meaning within a group of words that may be translated as "in front of," "before," "close to," and "on behalf of." Thus, even before it comes to be a legal term, "property" in the West expresses what distinguishes an individual or a thing from a group or from another. It is the face of one to other(s), what separates me from thee and ye, what lies in a person's view, or what has priority in time.

The Latin root

When the word *property* first appears in the English language, it is regularly used in either of the two senses of the Latin *proprietas,* or one easily derivable from them. Before the 17th century it is rare to find the word in its modern sense of an object of legal rights, and even where it is so found, the context almost always suggests a thing or group of things owned by an individual. The present descriptive definition of property law, then, abstracted from a relatively modern usage of the word property the body of law surrounding its object and then stripped the word of its normal connotation of individual ownership.

The Western tendency to agglomerate. If property law in the descriptive sense exists in all legal systems, the extraordinary diversity of the property systems of non-Western societies suggests that any concept of property other than the descriptive one is dependent on the culture in which it is found. Even in the West, as the discussion of the English word property shows, the concept has varied considerably over time.

Definition of "agglomerative tendency"

Nonetheless, one tendency seems to characterize the legal concept of property, in the descriptive sense, in the West: a tendency to agglomerate in a single legal person, preferably the one currently in possession of the thing that is the object of the inquiry, the exclusive right to possess, privilege to use, and power to convey the thing. In the technical language of jural relationships, Western law tends to ascribe to the possessor of the thing: (1) the right to possess the thing with a duty in everyone else to stay off, (2) the privilege of using the thing with no right in anyone else to prevent that use (coupled with a right in the possessor to prevent others from using the thing), (3) a power to transfer any or all the possessor's rights, privileges, powers, and immunities to anyone else (who would in the technical language be described as liable to the exercise of the power), and (4) an immunity from change by anyone of those same rights, privileges, and powers (so that everyone else is disabled from changing them).

Basic tensions within the tendency. The tendency of Western property law to agglomerate in the possessor of a thing all rights, privileges, powers, and immunities concerning the thing has never been more than a tendency, because the tendency carries with it two inherent tensions. First, the exercise of the power to convey by any individual cannot begin to be full unless he can limit the power to convey of the individual to whom he conveys. The Earl of Arundel, for example, knows that his eldest son, the heir apparent of the earldom, is insane and likely to die childless. The earl wishes to give the barony of Grostok to his second son but to transfer the barony from his second son to his third son if his second son should become earl. If the earl cannot do this, his power to convey the barony is less than full. If he can do this, the power of his second son to convey the barony will be less than full. (These are basically the facts of *The Duke of Norfolk's Case* [1682] 3 Ch. Cas. 1.)

The second tension is similar to the first. The privilege of one person to use his things, if exercised to the fullest, is likely to interfere with another's privilege to use his things. Two absolute privileges of use cannot coexist in proximity. The concept of proximity, moreover, has gained new meaning in the late 20th century, applying, for example, to a situation in which emissions of smoke may cause acid rain thousands of miles from the source of the emission.

Both tensions are likely to arise in situations in which the current possessor of the thing is being sued. The Duke of Norfolk's case will normally arise after both the earl and his heir are dead and the third son seeks to recover the barony from the second son, who has now inherited the earldom. The agglomerative tendency will favour the current possessor and may obscure the fact that ruling for the current possessor, the second son, will diminish the power to convey of all property holders. Similarly, cases of incompatible land use will normally result in the passive land user suing the active. For example, the victim of the acid rain will sue the factory emitting the smoke. Because the plaintiff in Western law generally has the burden of proof, the burden will be on the plaintiff to justify a limitation on the privilege of use of the defendant. If the issue is framed in these terms, the agglomerative tendency will favour the defendant and may even obscure the fact that the privilege of use of the plaintiff is equally at stake.

SCOPE OF THE ARTICLE

That the law has a tendency to agglomerate rights, privileges, and powers over a thing in a single individual is a commonplace on the European continent, where it has long been believed that the tendency was inherent in Roman law. That it is also a tendency in Anglo-American law is more controversial, because Anglo-American law recognizes more types of interests in things than does civil law. The first task, then, must be to show that the tendency exists generally in the West and to trace its history down to the present.

Second, the various rights, privileges, and powers associated with the term property must be analyzed in order to describe how the law categorizes the objects, subjects, and types of property interests, particularly possessory interests and interests denominated ownership; how it deals with privileges of use and limitations on such privileges; and how it deals with the acquisition and transfer of both possessory and use interests. The discussion closes with a theme adumbrated in the historical discussion—the future of the Western concept of property.

Relation to other branches of law. In nonindustrialized societies the complex of jural relations with respect to tangible things is intimately connected both with the economy of the society and with the family, at least the family as broadly conceived. In most such societies the economy and the family are also intimately connected. The tendency in industrialized societies, by contrast, is to divorce production from consumption—that is, the economy from the family. As a result, in the legal systems of the industrialized world, property used in production and property used within the family are divorced functionally and tend to be treated by different bodies of law.

Thus, although the descriptive definition of property employed in this article encompasses the American law of real estate, trusts and estates, family law (in part), sales of goods, negotiable instruments, secured transactions, securities (stocks and bonds), pensions, environmental protection, and patents, trademarks, and copyrights, only American lawyers who regularly deal with land or decedents' estates normally think of themselves as property lawyers. Those who deal with the other topics tend to think of themselves as family lawyers or commercial lawyers or corporate lawyers or environmental lawyers. Although some consideration is given here to all the branches of law that treat of tangible things and those intangibles that the system in question regards as property, the focus of

discussion is on the two areas of property law that have the deepest roots in the West, the law concerning land and that concerning the allocation and transmission of property interests within the family.

Geographic scope. As noted above, the Anglo-American and civil-law systems have become dominant in the industrialized world, and property systems derived from one or the other of these systems are found in many countries in which industrialization is just beginning. Former British colonies and members of the British Commonwealth tend to use some version of the Anglo-American system, frequently with additional rules designed to accommodate indigenous religions, family structures, and systems of land tenure. Such, for example, is the situation on the Indian subcontinent. Countries that were colonized by continental Europeans tend to have a civil-law system, with the French or the closely related Spanish model tending to dominate. Such, for example, is the situation in most of Central and South America. Even countries that were never colonized tend to have a Western property system. Japan's, for example, is based on the German Civil Code. China's property system, on the other hand, remains relatively free of influence from the West, although even China's has been influenced by the codes of other socialist countries, and these are largely based on civil law with separate rules for property used in industrial and agricultural production.

Spread of the Anglo-American and civil-law systems

The principal focus, then, is on the Anglo-American and civil-law systems that originated in western Europe. Specific examples from Anglo-American law come from the United States and England; specific examples from civil law come from France and Germany. Something is said about socialist legal systems where they differ notably from the two main systems. Occasional examples from some of the great religious legal systems are given and some generalizations made about other property systems of non-Western peoples. Here, however, illustrative examples must suffice. The focus is on land and property in the family in the Anglo-American and civil-law systems.

History and theory in the West

THE ANTHROPOLOGY OF PROPERTY

The evolutionary anthropology of the 19th century tended to see the emergence of absolutist individual property rights as a necessary and inevitable product of human development. Modern anthropology is considerably more cautious. The few generalizations that can be made on the basis of what is now known comparatively about the property practices of various societies, both historical and contemporary, do not seem to support a general evolutionary theory. But they can provide some starting points that help to explain certain features of the property systems that have emerged in the West.

First, almost all societies give individual members considerable control over some kinds of things. This normally includes the right to exclude most, if not all, other members of the society from those things, the privilege of using those things in a wide variety of ways, and the power to convey those things by gift or sale, at least inter vivos—that is, between living persons.

Second, every society has some kind of system for dealing with the things that were in the control of members who have died. These systems are frequently quite complicated. Some items may be buried or burned with the deceased; some may be taken by those in authority; but almost all societies, perhaps all, transfer at least some of those things from the control of the deceased to that of a member or members of his or her family (broadly conceived).

Generalizations about property systems

Third, there is a relationship between the economic system of the society and the types of things about which it has carefully worked out rules of property. A pastoral society, for example, is likely to have quite complicated rules and practices about the possession, use, and transfer of herds but may not have very complicated rules about land. On the other hand, an agricultural society is likely to have quite complicated rules about land but may devote correspondingly less attention to the rules about most kinds of movables.

Fourth, very few, if any, non-Western societies generalize about property in the way that Western legal systems do. Hence, the fact that most non-Western societies have some categories of things about which they have rules that look like the private-property rules of the West does not mean that the Western concept of private property is universal. What distinguishes the Western concept of private property from that found in the property systems of most, if not all, other societies is the fact that it is a default category; *i.e.*, Western legal systems regard individual ownership as the norm, derogations from which must be explained.

Modern anthropology has difficulty generalizing about history. Many non-Western societies have relatively short discoverable histories. The most dramatic change in the history of many non-Western societies has been their relatively recent contact with the West and with Western ideas. The fact that many non-Western societies have, as they have come in contact with the West, changed their property systems to make them more like those of the West does not indicate that such a change was inevitable, or even functional. It may be simply the result of the political and economic dominance of the Westerners who have come in contact with non-Western societies.

THE ORIGINS OF THE WESTERN IDEA OF PROPERTY

Deprived by modern anthropology of any notion that the Western idea of property can be shown to be an inevitable development of the forces of progress or civilization, one must look again at the emergence of the agglomerative tendency, which, as noted above, is characteristic of property law in the West.

Rome. In classical Roman law (*c.* AD 1–AD 250) the sum of rights, privileges, and powers a legal person could have in a thing was called dominium, ownership, or, less frequently, *proprietas* (though frequently enough for it to be clear that the two words were synonyms as legal terms). The classical Roman jurists do not say that their system tends to ascribe *proprietas* to the current possessor of the thing but that it did is clear enough. A number of Roman legal rules deny the label possession to the person who is in fact, though not legally, in possession in order to keep legal possession in the *proprietarius*. Further, the person legally in possession is presumed to be the *proprietarius*. This is clear enough from the procedural rules that require a person who is not peaceably in possession of a thing to establish affirmatively that his title is better than that of the peaceable possessor.

Once the Roman system had identified the *proprietarius*, it was loath to let him convey anything less than all the rights, privileges, and powers that he had in the thing. Thus, full-use rights divorced from ownership (*usufructus*) could be given only to a living person, and that person could not convey those rights to another. The ability of an owner to agree to legally binding restrictions on his privilege of use (servitudes) was sharply limited. Moreover, anyone who found himself owning a thing jointly with others could require that thing be divided ("nemo invitus ad communionem compellitur"; "no one is forced to have common property with another").

One might argue that the tendency toward absolute individual property rights in Roman law was more apparent than real. For example, classical Roman law never developed a remedy whereby an individual could, upon proof of ownership, specifically recover a thing. The owner could obtain a judicial declaration of his right to the thing, but the defendant could respond by paying damages. The Roman law of persons put extraordinary power over things in the hands of the head of the household (paterfamilias); indeed, this power was so extraordinary that an elaborate system (peculium) was necessary to allow slaves and sons in the power of their fathers to make binding legal transactions with things that were in fact but not in law their own. Moreover, land outside of Italy was owned not by individuals but by the Roman people collectively or by the emperor; yet individuals who had use rights in such land came to have control over it not far different from that of the owners of Italic land, even though they were not called owners. Finally, the sharp cleavage in Roman law between

public law and private law prevented the Romans from ever developing a legal notion of protection of property as against the state. This meant not only that property rights were not so absolute in Roman law as it might first seem but also that nothing prevented many of the sorts of conflicts about land use that in the later Anglo-American legal system were traditionally the subject of private tort suits or private agreements from being dealt with in Roman law as legislative or administrative matters.

Roman inheritance

One may question, therefore, how different was the sum of Roman property rules in the descriptive sense of the term from that of legal systems where the agglomerative tendency in property law is less manifest. Nonetheless, the tendency itself existed to a marked extent in Roman legal thought about property. It is notable not only in the ways outlined above in which Roman legal thought focused on the interests of the owner of a thing to the expense of those of others but also in the fundamental separation that Roman law made between property law and the law of obligations (contract and delict). This latter separation was to become characteristic of all the Western legal systems, while the specific decisions that the Roman jurists made about what was to be characterized as a necessary part of ownership became characteristic of many Western legal systems, particularly the civil-law systems.

The existence of the agglomerative tendency in Roman legal thought has no obvious explanation in Roman political or philosophical thought other than the broadest of connections with general ideas of individual worth. Further, present-day knowledge of the relationship of Roman law to Roman society makes hazardous any attempt to explain the existence of the tendency on the basis of social causes. That the tendency, coupled with the Roman law of persons, favoured the property-holding classes, seems obvious. That it was a product of their power, particularly in the Republican period, is likely. A number of its manifestations, however, cannot easily be attributed to class interest, notably the law's refusal to allow family settlements of any but the most short-lived variety, the paucity of land-use control devices, and the failure of the law to develop any notion of protection of property against the state.

England. In medieval English law, the procedural system prevented any clear separation between property and obligation. It was not until the abolition of the forms of action in the 19th century that Anglo-American law was fully able to implement the distinction in the way the Romans had. It is therefore remarkable that English law prior to the abolition of the forms of action tended at critical junctures to move in directions similar to the Roman—namely, to agglomerate property rights in a single individual

In England a notion of property in land emerged at the end of the 12th century from a mass of partly discretionary, partly customary, feudal rights and obligations. The way in which this happened was extraordinarily complex. What began as essentially an appellate jurisdiction, offered by the king in his court to ensure that a feudal lord did right by his men, ended with the free tenant being the owner of the land, in a quite modern sense, with the lord's rights limited to receipt of money payments.

Legislation at the end of the 13th century (statute De Donis, 1285) allowed a conveyor of land to limit its inheritance to the direct descendants of the conveyee and to claim it back if the conveyee's direct line died out (fee tail). In one of their few deviations from the principle of consolidating the power to convey in the present possessor of land, the English courts extended the scope of this legislation in the 14th century. In the middle of the 15th century, however, the courts reversed the trend and allowed the present possessor of entailed land to extinguish the interests of his descendants and of the conveyor (docking of entails by common recovery).

In the 16th century the process that had operated at the end of the 12th century to consolidate ownership rights in the free tenant was replicated for the copyholder, the descendants of those who held land by unfree tenure. The royal courts opened appellate jurisdiction to copyholders wronged by the unjust behaviour of their lords' courts, and the end result was that the copyholder became the owner of what had heretofore been the lord's land in the eyes of the king's law. Once again, the lord's right in the land was reduced to the receipt of money payments.

Medieval and Tudor land-use law

The history of English land-use law in the medieval and Tudor periods has not been well explored. A more static society may have produced fewer land-use conflicts than does modern society. Such conflicts as there were may have been resolved in courts other than the central royal courts. The cases in the central royal courts, however, suggest, somewhat surprisingly, that the agglomerative tendency was working on behalf of the plaintiff and not the defendant in disputes over land use (nuisance actions).

What protection a citizen's property received against the state in medieval England is a question that requires considerable translation for a world that knew neither citizens nor the state in the modern senses of the words. There are, however, intimations of a notion that property ought to receive some legal protection against the king. This may be an idea that underlies Magna Carta (1215). The statutes dating from the reign of Henry VIII (1509–47) that take property for public works are important both for their recognition of the state's power of eminent domain and for their provision for compensation to the owner whose property was taken.

The earliest manifestations of the agglomerative tendency in 12th-century England seem to have announced a fundamental change in the English social system. The man who was seised (*i.e.*, put in possession) of a freehold emerged as not far different from what one would nowadays call the owner of the property, and the rights of the lords of freeholders became more like those of taxing authorities. The rights of the nonfreeholders who held land of the free tenant, however, became obscured by the fact that they were not protected in the king's courts.

Current understanding both of the purpose and of the effect of this legal development is obscured by the fact that it is not possible to date precisely when the changes either in law or in society took place. It may be, as the legal historian S.F.C. Milsom has suggested, that the purpose of the intervention of the king's courts was to shore up a system that was weakening and that the intervention, by shifting jurisdiction from the lords' courts to the king's courts, had the unintended effect of destroying the previous system.

The tendency in the later Middle Ages to give greater power to convey to the present holder of the land at the expense of his power to convey in such a way as to deprive future generations of that same power has been seen as favouring the interests of lords. But that view does not explain the whole of a complex reality. Preservation of feudal revenues is certainly important to the understanding of such developments as the abolition of subinfeudation (statute Quia Emptores, 1290) and the requirement that legal title to land follow the right to its use (Statute of Uses, 1535), but it does not explain why the courts first extended and then restricted the power of the landowner to entail his land.

The Continent. The history of property law on the continent of Europe has not been as well explored as that in England. The collapse of Roman and then of Carolingian power led, in most areas on the Continent, to a situation not unlike that which prevailed in England before the emergence of the central royal courts in the late 12th century. As in England, land was bound up in a mass of partly discretionary, partly customary, feudal rights and obligations. England, however, was precocious in developing central royal courts as early as it did. In most areas of Europe lords' courts remained a significant force for a much longer period, even for free tenants.

The influence of Roman law

The Roman idea of property was revived on the Continent as an intellectual matter before it came to have much practical force. Beginning in the 12th century, the study of Roman law in the universities led to a renewed awareness of Roman conceptions of property, and in many areas a mixture of Roman law and canon law, known as jus commune ("common law"), came to be authoritative in the absence of local law. The commentators on Roman law in the 14th and 15th centuries developed ways of

describing the prevailing feudal forms of property-holding in terms of the Roman law (dominium utile; *quasi-possessio iuris*), and this process led to making these forms of property-holding more like the Roman. Further, Roman ideas were influential both because they were part of the equipment of every university-trained jurist and because they were part of the jus commune. By the end of the Middle Ages the property law of most European countries was still far from that of the Romans, but it was heading in that direction. Civil law was thus displaying the same agglomerative tendency noted in more detail for England.

Explaining the origins. So far a tendency has been identified that manifested itself with sufficient frequency to be worthy of being described as a fundamental tendency of property law in the West. How can this fundamental tendency be accounted for? Clearly, it antedates the philosophy of possessive individualism of the late 17th century, and it long antedates 19th-century liberalism, although it may not antedate certain general individualist trends in Western thought. Social causes may partially account for the tendency in Roman law; but if the dominus of Republican Rome was the most important element in the political system, the same can hardly be said to have been true of the ordinary freeholder of 13th-century England and even less of the copyholder of the 16th century.

To explain the tendency, something more basic is needed than the influence of a particular philosophical idea or the dominance of one social interest over another or even the product of a balancing of social interests. Both the Roman and the Anglo-American legal systems began as mechanisms for resolving disputes. Both systems began with possession of a thing by an individual. The convenience of assuming that that possessor had all the other rights, privileges, and powers one might have in a thing may go a long way to account for the presence of the tendency in both legal systems. The tendency began as an allocation of a burden of producing evidence of ownership. A dispute arose about a thing. Both systems began by determining who is possessed of it. They then assumed that that person had all the rights, privileges, and powers that go along with property until someone else could show that that was not the case.

Although Western legal systems are not unique in beginning as dispute-resolution mechanisms, the Western concept of property is, if not unique, certainly unusual. One may speculate that what makes this dispute-resolution device operate in favour of the individual property holder in the West is an accident of chronology: the coincidence of the emergence of systematic legal thinking (associated with professionalization) and a state of society that saw one individual's connection with a thing more clearly than it saw any group's connection with the thing. Thus, the notion of individual property emerges in both Roman and English law at a time when family ties to property were weakening and legal professionalization was occurring. In the Rome of the earliest jurists (*c.* 100 BC) the gens was ceasing to have the power over family property that it probably once had. In late 12th-century England, heirs were losing their power to control their ancestor's alienations. At the same time neither society was yet prepared to see the connections with the property of such groups as slaves and sons in paternal power in Rome or serfs in England. What caused this state of society, and how important it was that there existed general notions of individual autonomy and worth before the idea developed, is impossible to say. History reveals more of what happened to the tendency after it started than it does of how it started.

PROPERTY LAW AND THEORY IN THE EARLY MODERN PERIOD

Beginning in the 17th century, developments in property law both in England and on the Continent can be related to developments in speculative jurisprudence. Although general speculation about the nature of property is at least as old as Plato and Aristotle and property is considered, sometimes quite critically, in the writings of Church Fathers of the Latin West and medieval theologians, these writings had relatively little direct effect on the secular law.

The classical theories of property. In the early 17th century the Dutch speculative jurist Hugo Grotius announced the theory of eminent domain. On the one hand, according to Grotius, the state did have the power to expropriate private property. On the other hand, for such a taking to be lawful, it had to be for a public purpose and had to be accompanied by the payment of just compensation to the individual whose property was taken. The idea was not new with Grotius, but he stated it in such a way that it became a commonplace of Western political thought. The theory of eminent domain

In the late 17th century Samuel von Pufendorf refined a theory of the origins of property rights that had been in existence since ancient times. Property, Pufendorf said, is founded in the physical power manifested in seizing the object of property (occupation). In order, however, to convert the fact of physical power into a right, the sanction of the state is necessary. But the state cannot, Pufendorf seems to suggest, make a property right where physical possession is not present. Thus, both occupation and state sanction are necessary conditions for the legitimacy of property.

Pufendorf's contemporary John Locke had a different theory, again one that had considerable antecedents. What gives a man a right to a thing, according to Locke, is not simply his seizing of the object but rather the fact that he has mixed his labour with the thing in making it his own. This right to a thing arising out of labour is a natural right. It does not require state sanction in order to be valid. It should, however, be protected by the state. Indeed, property is fundamental to the contract that people make in forming the state, and for the state to deny the right to property is a breach of this contract.

The end of the 18th and the beginning of the 19th century saw the emergence, particularly in England, of a new set of ideas about property, influenced by ideas prominent in the Scottish Enlightenment but given a new and powerful twist by the utilitarian political philosopher Jeremy Bentham. Property, according to Bentham, is nothing but an expectation of protection created by the legislator and by settled practice. It is, however, an expectation that should be carefully respected. Since the function of the legislator is to maximize the sum of human felicity, he should know that rarely does any interference with property produce more felicity than it destroys.

Bentham's follower John Stuart Mill associated property with liberty and suggested that security of property is essential for man to maximize his potential for liberty. Modern economic theories of property that justify property on the ground that there must be an initial allocation of resources to allow the market to operate and on the ground that individual property rights minimize transaction costs derive from the tradition of Bentham and Mill.

On the Continent, thought about property took a somewhat different turn. Building on Immanuel Kant's categorical imperative that persons must always be treated as ends in themselves rather than as means, Georg Wilhelm Friedrich Hegel suggested that the same imperative applies to a person's property. The reason for this, according to Hegel, is that when someone extends his will to a thing, he makes that thing a part of himself. Protection of property is thus intimately connected with protection of the human will. Hegel

The middle of the 19th century saw the first concerted attacks on the institution of property since the time of the early Christians. *The Communist Manifesto* (1847) of Karl Marx and Friedrich Engels holds that property is nothing but a device in the social warfare between the capitalist and proletarian classes, the means by which the capitalist expropriates the labour of the proletarian and keeps him in slavery. Reform, according to Marx and Engels, would not come until the revolution, when property would be abolished.

Possessive individualism and the law. The 17th- and 18th-century theories of property, particularly Locke's, have been called possessive individualism. The importance of property in these theorists' overall conception of man and the state is undeniable. The extent to which possessive individualism manifested itself in the development of property law in the late 17th and 18th centuries is not so clear.

In England, Sir William Blackstone's concept of property was curiously bifurcated. In a famous passage he lists property along with life and liberty as one of the absolute rights of Englishmen, but when he comes to restating the rules derived from this absolute private right, the tendency to absolutism is far less manifest. Blackstone's encomium of property reflects much that had been expressed before in the philosophy of possessive individualism, but it is not until the Fourth and Fifth Amendments to the U.S. Constitution in 1791 that one finds these ideas embodied as specific legal principles. And it was not until the second half of the 19th century that they received any substantial elaboration in American federal jurisprudence.

More careful research might be able to trace the influence of the ideas of the 17th- and 18th-century continental theorists of property in areas other than international law, about which most of them wrote. Prior to the French Revolution, however, it is hard to see any other specific legal manifestations of their theories.

What one does see in writings about property law on the Continent from the 16th through the 18th century, however, is the increasing dominance of ideas drawn from Roman law. In the 16th and 17th centuries local customary law was redacted into brief statements of rules. The redaction process itself had a tendency to magnify the influence of the jus commune, because what could not be found in the local custom was filled in with the jus commune. The local customs were then subjected to academic commentary that made use of the categories of Roman law, including the category of property. Finally, particularly in Germany, the entire corpus of scholarship on Roman law was brought to bear on local law, integrating it more fully into the European mainstream (a movement called *usus modernus pandectarum*).

Marxism, liberalism, and the law. Not surprisingly, relatively little of Marx's theory of property showed itself in property law until a Marxist revolution took place in Russia in the early 20th century. For utilitarianism and Hegelianism, and their combination in various forms of liberal thought, the evidence of influence is more pronounced as the 19th century progressed.

Nineteenth-century legal change. The beginning of the 19th century saw the promulgation in France of the Napoleonic Civil Code (1804), a systematic and comprehensive code of private noncommercial law that was to have great influence in the European codification movement which followed. The code is notable for its reluctance to recognize interests in property other than the owner's. This is probably to be accounted for by the fact that the codifiers were determined to abolish what they called feudalism—hereditary class privileges and all that was associated with them—and one of the characteristics of feudalism in the codifiers' minds was the recognition of multiple outstanding interests in land. While it is chronologically possible that the code's concept of property was influenced by utilitarian ideas, most of what is in the code can be adequately explained by the concept of property that the drafters of the code found in Roman law and in continental writing in the Roman law tradition.

Liberal conceptions of property seem to have influenced legal thought later in the 19th century. On the Continent the pandectists, a group of systematic jurists prominent in Germany, took the agglomerative tendency inherent in the Roman conception of property and developed it to a point that most modern commentators find goes far beyond what the Roman sources themselves suggest. Their ideas were embodied in the German Civil Code (effective 1900) and had substantial influence on the codes of other countries.

The pandectists

The tendency toward absolutism in matters of property is also remarkable in Anglo-American law on both sides of the Atlantic. The English Settled Land Acts (1882, 1890, 1925) gave considerably more power to the present holder of settled land than the common law had given him. The Married Women's Property Acts in both countries (England from 1857; United States from 1839) divided property within the marital unit by assigning it to one or the other spouse. The examples could easily be multiplied.

In the United States the influence of liberal ideas on the development of the law of property can be seen most strongly in the substantial jurisprudence that emerged concerning the protection of property against the state. On the federal level, this development came after the passage of the Fourteenth Amendment to the U.S. Constitution in 1868. This amendment prohibits, among other things, any state from depriving a citizen of property without due process of law. Under this rubric, many types of economic and land-use regulations were struck down by the U.S. Supreme Court on the ground that they interfered with citizens' property rights.

As in the case of Roman law, so too in the case of 19th-century Anglo-American law, focusing on what is called property law may obscure the reality of the total law with respect to the possession, use, and conveyance of things. Thus, while the rights, privileges, and powers of the owner seemed to expand, other legal developments undercut them. For example, while the types of permissible easements (a property interest in land other than the owner's) were restricted in the 19th century, the English equity courts in the same period created a new form of obligation, now called equitable servitudes, which served the same function as easements and were not subject to the same restrictions (*Tulk* v. *Moxhay,* 2 Ph. 774 [Ch. 1848]). While the agglomerative tendency affected the law of trusts, the express trust resisted any attempt to defeat its basic division of legal from equitable title. Indeed, trusts were increasingly used as a form of holding newly important aggregations of personal property. Similarly, the increasing complexity of land-use conflicts led to an ever growing body of local land-use regulations, and the first comprehensive planning act was passed in England at the beginning of the 20th century (Housing, Town Planning, etc. Act of 1909).

Nineteenth-century legal thought. Perhaps because the category property told little of the real extent of the property holder's rights, privileges, and powers, the analytic jurists of the 19th and early 20th centuries were not so successful in constructing a system of "scientific" property law as they were in other areas of private law. Analytic jurisprudence in the Anglo-American realm tended to focus on the aspects of property law to which its technique was peculiarly adapted. Thus, the laws relating to common-law estates in land and future interests became the system of estates and future interests; the doctrine of perpetuities became the rule against perpetuities; and the doctrine about the three types of private land-use controls—easements, covenants, and equitable servitudes—was systematized. The first *Restatement of the Law of Property* in the United States (1936–44) limits itself to the two broad areas of estates and future interests (including the rule against perpetuities) and private land-use controls. Notably missing from the restaters' efforts were conveyancing law (except to the extent that it was included in estates and future interests), landlord and tenant law, public control of land use, and the law of natural resources.

Similarly on the Continent, the concept of possession received loving elaboration in the 19th century. In the hands of some pandectists it yielded striking echoes of Hegel's will theory. The various schemes of marital property were carefully delineated. Conveyancing, considered a matter for notaries rather than academic jurists, was largely ignored. Land-use regulation was left to public law and, hence, outside the ambit of those concerned with property.

The social situation. The social situation within which the law was operating changed markedly over the course of the 17th through the 19th centuries. In England in the 17th century the abolition of most of the lord's rights to receive money from his freehold tenants relegated the tension between the free landowner and his lord to the background, and a new tension emerged between the established country landowner seeking to perpetuate his family on his land and the newly wealthy man of commerce seeking to buy country land. While lordship remained a force on the Continent throughout the 18th century and in Germany into the 19th century, the same tension between noble and bourgeois was also apparent. The restrictions on the power to tie up land for long periods of time, which the law either invented or extended in

this period, favoured commercial interests. That the law stopped short, considerably short in the case of the Anglo-American law, of abolishing all restrictions on the power of the conveyee himself to convey is a product not only of a recognition of the tension in the fundamental tendency to agglomerate but also of the power of the countrymen and their conveyancers to influence the course of legal development.

Industrial versus agricultural land uses

By the 19th century the conflict between the established landowner and the merchant gave way to a conflict between industrial and agricultural land uses, and the key issue was not the power to alienate land but the privilege to use it. In Anglo-American law the 19th-century courts gave broad scope to industrial development of land at the expense of adjoining residential and agricultural uses. The courts, however, at no time recognized an absolute privilege of land use, and this may be seen as a product both of the inherent tension in the fundamental tendency and also of the fact that the agricultural classes retained some social power. The developments on the Continent are less well known, principally because of the civil-law tendency to relegate land-use control to the area of public law and hence outside the purview of property.

The fundamental tendency revisited. As the need arose for a category to describe the sum of the rights, privileges, and powers that an individual could have with respect to a thing, the Romans, followed by the English, chose a noun derived from an adjective that means "own." The category at once described the concept and also the tendency. As time went on, the tendency took on an independent life. Western law excluded from the category property certain rights, privileges, and powers with respect to a thing because they existed in someone other than the property holder. In modern legal systems, though not in the Roman, property came to represent one of the rights of the individual against the state, perhaps originally because property had come to rest in the freeholder and not in his lord, and the king was the lord of all. The tendency was reinforced by philosophical ideas; social forces not only attached themselves to it but also arranged themselves around tensions within it.

In classical Roman law and in the late 19th and early 20th centuries, when the agglomerative tendency was at its strongest, the sum of rules pertaining to things undercut the tendency, but if one looked solely to the rules labeled property, the system of categorization obscured this fact. The question even was raised with regard to Roman law whether the strength of the tendency made any difference. One suspects that it did. In the United States in the 19th century, the categorization process clearly allowed the tendency to work in a more unfettered fashion than it would have if more conflicting interests had had to be reconciled under one category.

RECENT HISTORY OF THE WESTERN CONCEPT OF PROPERTY

State interference with property rights

The first half of the 20th century saw an extended debate in the United States about the extent of permissible state interference with property rights. One by one the more extreme doctrines of the late 19th and early 20th centuries were struck down. By the mid-1930s federal and state regulation of the economy could no longer be challenged on the ground that it constituted a deprivation of property rights without substantive due process of law, and direct restrictions on the use of property in the form of comprehensive zoning and planning ordinances had been sustained. Thus, the trend seemed clearly away from use of the property clauses of the Constitution to protect citizens against the state.

Neither the United Kingdom nor the countries of the Continent make as much use of judicial review of public regulation as does the legal system of the United States, and neither the United Kingdom nor the Continent ever went as far as the United States in disabling the state from acting in areas that affected property rights. Both the United Kingdom and the Continent have strong traditions of protecting property rights against state interference, however, and despite these traditions the 20th century has seen increasing state regulation of property. The debates tend to be legislative rather than judicial, and, although the regulators do not always win, over the course of the century they have won more than they have lost.

As part of their effort to permit greater regulation of property rights, the legal realists in the United States in the 1920s began to attack the categorization process itself. In both the public and the private law of property they found an easy mark, if only because the systematizers had been less successful there than in other areas of private law. Thus, in the late 20th century, the concept of property in the United States is a matter of controversy. The fundamental tendency has received a major setback in the field of public law, and the whole process of categorization with which the tendency has been associated is discredited.

In the United Kingdom and on the Continent, because property is less of a constitutional concept, it is less necessary to take a position about it. The growth of public regulatory law has meant that the concept of property may remain more or less intact but is of less relevance to the sum of rights, privileges, and powers with respect to things. Thus, in these countries modern controversies that strongly affect property in the descriptive sense are not always seen as involving property in the legal sense in the way that they are in the United States.

Objects, subjects, and types of possessory interests in property

Property law has been defined in this article as the sum of jural relationships with respect to things and things as all tangible things and those intangibles that the legal system in question classifies as property. This definition creates difficulties when one comes to describe property systems generally because it mixes a definition of things that is external to a given legal system with a definition of things that is dependent on the legal system in which it is found. The difficulty becomes more acute as one moves from general definition to subordinate levels of classification, for at these levels of classification, almost all the terms are specific to a given legal system or group of systems.

Most of these difficulties of bias in favour of one legal system or another can be avoided by acknowledging in advance where a particular scheme of classification comes from. The one used in this article is based on that employed in Western legal systems. Where the Anglo-American and the civil-law systems diverge, the categories of the Anglo-American system are considered first and are then compared functionally with those in the civil-law system.

The discussion begins by identifying the objects (things) and subjects (persons and groups) of the jural relationships with regard to things in Western legal systems generally. There follows a treatment of possession and ownership, categories that are closely related historically in the West. Then the discussion deals with divisions of ownership and in so doing contrasts the divided ownership system of the Anglo-American law with the devices in the civil-law system that achieve many of the same practical results while employing a quite different set of concepts. The section closes with the procedural protection of property interests.

OBJECTS

Classification of "things." The Roman jurist Gaius (*c.* AD 160) includes in his treatment of the law of things the acquisition and loss of individual things, succession (both testate and intestate, including legacies, with some references to what today would be called bankruptcy), and the acquisition and extinction of obligations (contracts and delicts). If one views Gaius' category of the law of things as equivalent to the descriptive definition of the law of property used in this article, then his category "property" is very broad indeed.

Property patrimony, and obligations

In civil-law systems today the term "property" applies to those tangible things that can be conveyed inter vivos and to a very few, if any, intangibles. Patrimony is a broader term. It includes everything that is the object of the law of succession. Obligations are defined in contradistinction to property. They do not receive the procedural protection that property interests receive, and the state is not as constrained in dealing with them as it is in dealing with property interests.

Anglo-American law is generally less concerned with matters of definition than is the civil law. Except in the United States, where defining something as property automatically entitles it to constitutional protection, there is less discussion in the Anglo-American legal system of whether a given interest or a given thing should be classified as property or not. Nonetheless, Anglo-American law shows broadly the same characteristics as the civil law. Almost all tangible things are conceived of as being capable of supporting property interests; some intangibles are treated the same as tangibles, and some are not.

Within and without commerce. Of all things, Gaius says, some are of divine right (*divini juris*) and some are of human (*humani juris*). Of those that are of human right, some things are of public right (*publici juris*), and some things are things of individuals (*res singulorum*). Other Roman jurists further divided some of the things in Gaius' category of public right into things that by natural law are common to all (*res naturali jure communes omnium*) and things that belong to no one (*res nullius*). Some medieval jurists combined elements in the categories of things of divine right, things of public right, things of no one, and things common to all in a single category of things outside commerce (*res extra commercium*). These schemes point in somewhat different directions, but they all deal with the same core idea: some tangible things are withdrawn from the normal operations of property law. Modern jurists have attempted to categorize these things more precisely, though with no more success than the older jurists.

Religious things. The Roman category of things of divine right has generally been abolished in the West, and things that are dedicated to religious use are normally conceived of as the property of a state-established or state-recognized religious group. The religious group may, in turn, have internal rules that require that things associated with the religion be treated differently from other kinds of things. Roman Catholic canon law, for example, allows a diocesan bishop to relegate a church to profane uses only if he determines that serious reasons suggest that the church no longer be used for worship and that the good of souls will not be impaired; he must consult his presbyteral council and obtain the consent of those legitimately claiming rights in the church.

Non-Western legal systems, particularly religious legal systems, tend to retain a special category for things dedicated to religious or charitable use. Thus, in Hindu law the deity represented by an idol is regarded as owning the property dedicated to its use. Similarly, assets in an Islāmic waqf (foundation, charitable trust; see below) are regarded as belonging to God.

Things that cannot be possessed. A rather large group of tangible or at least perceptible things are withdrawn from commerce in most Western legal systems because of the difficulty of imagining how one could possess them. The ambient air, for example, is normally not the object of property. The effect of this doctrine has been to make air a free good, subject to pollution by anyone who has the capability of polluting it. Modern concerns with air pollution have led some commentators to suggest that recognizing property rights in the air might lead to more efficient control of air pollution.

Wild animals. All Western legal systems have special property rules about wild animals. Roman law, followed with some modifications by English law, held that wild animals belonged to no one until they were captured. Once captured the animal belonged to its captor unless the animal escaped and returned to its natural liberty. The capture of fish and game is regulated everywhere in the West, and some legal systems hold that such animals belong to the state unless and until they are captured in a manner conforming with the regulatory scheme.

Water. Water and the land under and bordering on water are everywhere in the West treated differently from other kinds of property. In Roman law navigable water and beaches were common to all. As in the case of wild animals, modern law in the West tends to give substantial power over water and land near water to the state. Beyond that the regimes vary substantially from jurisdiction to jurisdiction.

Highways of the monarch

In England, navigable waterways are the highways of the monarch, open to all who have access to them. Access, however, is not guaranteed. Because private ownership extends from the low-water mark, the public has access to water and so may use it for navigation only if there is a public right-of-way leading to the water.

In the United States the same rules apply so far as access is concerned, but beaches may, depending on the jurisdiction, be subject to a public right of passage up to the high-water mark. The United States also has a well-developed law concerning the taking of water from a navigable or nonnavigable stream. In the eastern part of the United States the right to take water from a stream is dependent on ownership of lands adjoining the stream. In the western part of the country the right to take water tends to depend on having first taken it (prior appropriation). In both parts of the country public regulation has increasingly come to the fore, and the fact that water was historically not an object of private ownership has made it easier to introduce regulatory control over this resource.

Other natural resources. Other natural resources have, in some Western legal systems, been removed from normal private ownership. The tendency on the Continent is to make all minerals subject to state ownership or at least to extensive state control. Historically in England gold, silver, and lead were reserved to the crown. In the United States private ownership of minerals has been the rule, subject to considerable state regulation in the name of conservation. Just as the systems of private ownership with regard to water have tended to divide between those systems that award the water to the person who has it on his land and those that award it to the person who discovered or appropriated it, so, too, those Western systems that allow private ownership of minerals alternate between giving them to the landowner and giving them to the discoverer.

The human body. Throughout the West, the human body, living or dead, is not an object of private property. This fact has raised difficulties in many legal systems. For example, if the human body is not property, the question arises of what is happening when someone makes a gift of or sells blood or bodily organs or makes a testamentary disposition of his body for medical purposes. Many jurisdictions have special legislation on this topic, but the conceptual difficulty is by no means resolved.

Possession of tangible things. Possession of a tangible thing is, at least in the West, a concept that antedates conscious thought about law. Possession is a fact, the Roman jurists said, formed of an intention and a thing (*animus et corpus*). The thing was basically anything that was capable of being physically possessed; the intention was to hold it as one's own.

Seisin

English law also had to deal with a fairly complicated social fact, seisin, the process by which a lord put his man in possession of a tenement. In English law the concept of seisin was also applied to tangible things other than land, things that were not subject to lordship. Together the two concepts of seisin came to equal something quite close to the Roman possession.

Any legal system that begins its property law with a concept of possession is going to have a property law biased in favour of tangible things. It is easy for Westerners to conceive of possessing almost anything that can be touched. It is far more difficult to conceive of possessing an abstraction like a right, a privilege, or a power. Westerners who are not lawyers will say that they possess their watches or their land; they will rarely say that they possess their bank accounts or the power to convey their land.

Possession of intangible things. Civil law, following Roman, has tended to deny the possibility of legal possession of anything that cannot be touched. English law, by contrast, and, following English law, Anglo-American systems generally are more open to the notion that one may be possessed of a right, a power, or a privilege.

Because possession is so fundamental to property in both Anglo-American and civil law, the civil-law systems' greater reluctance to recognize possessory interests in intangibles has important consequences for the way the two systems conceive of property rights. In the case of land,

civil law tends to give possession to the owner of the land and to be reluctant to recognize property rights in anyone other than the owner. Anglo-American law, however, recognizes multiple possessory rights in land and hence tends to speak not of ownership of the land but of ownership of an interest in land—*i.e.,* of an intangible legal abstraction in a tangible thing.

Although the abstraction of interests in land came early in the development of the English law, it was not until quite recently that Anglo-American jurisdictions recognized the possibility of divided possessory interests in movable property, and the English legal system may not recognize it to this day. Abstraction of contract rights and duties was also slow in developing. It was not until the 19th century that contract rights became freely assignable, and it was not until the same century that methods were developed whereby they could be protected against interference by third parties. Because contract rights were not assignable and because they could only be enforced against the maker of the contract, it was difficult to conceive of them as property. Bank accounts and stocks and bonds are simply standardized forms of contract rights, and so it was as difficult to conceive of them as property.

Negotiable instruments

The conceptual breakthrough came with regard to negotiable instruments. A promissory note (an agreement to pay back a loan), for example, made payable to the creditor "or order," is freely assignable by the creditor and by anyone to whom the creditor may assign it. The assignee (holder in due course) is also entitled to recover the debt even if the original creditor could not recover it (because, for example, it was part of a larger transaction in which the creditor had failed to perform his part).

Since the negotiable instrument is freely assignable and since it is enforceable without regard to the underlying transaction, it has two of the general characteristics of property: it is freely alienable and enforceable without regard to the relationship of the right-holder and the one against whom enforcement is sought. Negotiable instruments are thus sufficiently like property that Anglo-American law came to regard them as a species of property. Other forms of contract rights, such as bank accounts, followed. Today most forms of contract rights are treated for many purposes as property in Anglo-American law.

On the Continent historical development since the Middle Ages has moved away from the Romans' thing-oriented conception of property. The development came first in the doctrines about possession. While the medieval jurists recognized that one could not physically possess a right, they nonetheless developed possessory remedies for those who had been deprived of intangible rights that they had long enjoyed (quasi-possession of rights). The development, however, was aborted. The concept of quasi-possession of rights was associated with feudalism, in the context of which it had developed. With the abolition of feudal forms of property at the time of codification, the treatment of intangibles as property tended to be suppressed. Negotiable instruments, although they were (and are) recognized on the Continent, were treated in different kinds of courts. When codification came, they were not incorporated in the civil codes but were left in the commercial codes, and their use was frequently limited to those who could qualify as merchants.

This development could not last for long in a world in which contract rights are the principal means for the storage of value. The French civil lawyers, for example, soon came to recognize that it was possible to have a property interest in an intangible, and French developments in this regard have tended to parallel the Anglo-American. The German civil code, however, proved more intractable. The code specifically excludes the possibility of property in intangibles; so the development of property interests in contract rights in Germany has been devious. By the middle of the 20th century, however, German jurists had developed mechanisms whereby intangible rights and privileges could be transferred and protected in ways quite similar to property interests, despite the fact that they cannot be called property.

Government-granted rights as property. The 19th and 20th centuries have seen a great expansion of the types of intangible rights granted by the government. The oldest of these are the exclusive rights given by the state to encourage and protect authors, inventors, manufacturers, and tradesmen. Copyright, the exclusive right to prohibit the copying of a piece of writing or a work of art or music, is almost universally regarded as a property right. In the Anglo-American system it may be enforced by an injunction—a characteristic of property rights generally. In most Western systems copyrights are freely assignable. They are normally protected against state interference in the manner of other forms of property. Patents, the government-granted right to the exclusive use of an invention, and trademarks, the government-granted exclusive right to market one's product with a given distinctive sign or symbol indicating its source, receive similar treatment in most Western countries.

The "new property"

The last generation has seen considerable debate in the West about whether property-like protection should be extended to other forms of government benefits. Modern Westerners are heavily dependent on the government. Most professions and many trades and businesses cannot operate without a government license. Citizens of virtually every Western country are protected against unemployment by a system of government insurance. Those who cannot work frequently subsist on government grants. Many of the elderly live on government pensions or other forms of social welfare. In many Western countries the government provides medical services or health insurance.

All these benefits have increased dramatically in number and importance over the course of the 20th century. By and large the law concerning them is treated by jurists as a matter of public or administrative law. While they are clearly as important or more important to the citizen than the traditional forms of property, it was not until 1964 that the American educator and lawyer Charles Reich suggested that the legal system should treat them like property. His suggestion was favourably received, at least for a period, by the courts in the United States. It then seemed to disappear from view.

In the United States it seems clear that the legislature may make a grant to an individual or group of individuals in such a way as to entitle that individual to property protection in the grant. The grant may then not be taken away without due process of law in a procedural sense. The grant may even be made in such a way that it cannot be taken without the payment of compensation. In other countries in the West the courts have been less involved in these public-law programs. It is perhaps all the more notable, therefore, that throughout the West there has been a tendency in recent years to make at least certain kinds of government grants more secure. As a general matter, government grants can be taken away for fewer reasons, and the process by which they can be taken away has become more elaborate.

The same tendency toward property-like treatment is also noticeable throughout the West with regard to certain kinds of arrangements between private citizens. Landlord-tenant law, for example, a traditional topic of property law at least in the descriptive sense, has tended to give greater security to the tenant (see below). Western law has also tended to give greater security to employees (who are not the holders of property rights even in the descriptive sense), requiring, for example, that an employer justify discharging a long-term employee. Employer-provided pension plans have come under increasing regulation. In the United States the Employees Retirement Income Security Act of 1974 (ERISA), for example, requires that after a period of time an employee's pension rights shall be deemed to have vested—*i.e.,* that the employee be entitled to a fixed amount from the pension plan, even if he quits his employment.

Movable and immovable property. If the distinction between tangible and intangible property has become increasingly blurred in Western law and if the category of intangible property seems to be increasingly expanding, the distinction between movable and immovable tangible things has remained relatively fixed. As noted above, Anglo-American property law began as a law concerning land. The actions that protected interests in land were real

Real and personal actions

actions, real both in the sense that the interest claimed was notionally good against the whole world and real in the sense that the remedy afforded was the recovery of the land itself or the interest claimed in it. Movable objects, by contrast, were protected by personal actions—personal in the sense that one had to allege that the defendant had committed some wrong in order to recover and personal in the sense that money damages, not specific recovery of the thing, were normally the only available remedy. Reflecting these two types of actions, immovable property came to be called real property, and movable property, personal property.

Beginning from a law that made a radical distinction between interests in land and all other kinds of property, modern Anglo-American law has gradually come to view both kinds of property as similar. There remain, however, in many jurisdictions distinctions between the two that are more the product of the historical development than they are of any modern functional distinction. In almost all Anglo-American jurisdictions, for example, different forms of conveyance are used depending on whether the property conveyed is real or personal. The types of interests that may be recognized in the two also vary in many Anglo-American jurisdictions.

Modern civil law, by and large, has followed Roman law in minimizing the distinction between movables and immovables. Certain types of privileges of use are recognized only in land, but these tend to be interests that could not be had in a movable good, such as a right-of-way or a privilege to build. Conveyance of land may be somewhat different, but not radically different, from conveyance of movables. Statutes of limitation or periods of prescription may be longer for land than for movables. On the whole, however, the differences are not so great as they are in Anglo-American law.

All Western legal systems recognize that things attached to the land, such as trees or buildings, can become part of the land, so that they are automatically conveyed if the land is conveyed. Jurisdictions differ as to the extent to which it is possible for a landowner and someone else to agree that something built on the land will not become part of the land and as to the extent to which it is possible to convey something attached to the land separately from the land. Generally, the Anglo-American law of fixtures (things permanently attached to the land) pays more attention to the intent with which the attaching was done than does the civil law and gives greater freedom to convey fixtures separately from the land.

The law of fixtures

SUBJECTS

The topic of the subjects of property rights has been greatly affected by the agglomerative tendency. Both Anglo-American and civil law sought a single legal person in whom the vast complex of property rights, privileges, and powers could be said to reside. Historical shifts in the law of persons (the recognition, for example, of more persons as being of equal status before the law) have created more persons to whom the agglomerative tendency could attach but have not defeated the tendency. The fact that modern law freely allows the creation of fictitious legal persons (corporations) has, if anything, exaggerated the tendency.

Single individuals. In both Anglo-American and civil law the paradigmatic holder of property is a single human person. The fact that in the West today far more wealth is held in some form of co-ownership or corporate ownership has not yet affected this paradigm.

Historically, property-holding capacity was quite limited, and the capacity freely to deal with property was even more limited. In classical Roman law, for example, only free adult male citizens not in paternal power were fully capable of holding and exercising property rights. Since paternal power lasted until the death of the parent or ancestor (there was no automatic emancipation), since a large percentage of the population were slaves, and since more than half the population were women and boys, very few people, in fact, qualified as subjects of full property rights. The situation in medieval England was not too different. Automatic emancipation at the age of 18 or 21 and the notion that adult unmarried women had property-holding capacity somewhat expanded the number of persons having property-holding capacity, but serfs, aliens, children, and married women had limited property-holding capacity or capacity to deal with property.

The gradual abolition in Anglo-American and civil law of principles and rules based explicitly on personal status has made it easy to forget the limitations that still exist on property-holding capacity and on the capacity to deal with property. Thus, many Western jurisdictions still limit, in some way, the property-holding capacity of noncitizens. Many of the Western countries that have indigenous non-Western peoples living among them have separate rules concerning these peoples' property-holding capacity. Such regimes exist, for example, for American Indians who reside on reservations, at least with regard to tribal land.

Limitations on property-holding capacity

Many citizens who are legally capable of holding property are not legally capable of dealing with it. In Western legal systems generally, children are recognized as capable of owning property, but they cannot deal with it without the consent of their parents or guardians. All Western legal systems have procedures whereby incompetent adults can be deprived of their capacity to deal with property. These procedures generally provide for the appointment of a guardian for the incompetent; the guardian is authorized to deal with the property on the incompetent's behalf.

Restrictions on both the property-holding capacity and the capacity to deal with property of competent adult women have largely been abolished in the West. Marital property regimes differ substantially, however, and although the trend is in the direction of equalizing the powers of husband and wife, full equalization of the power to deal with marital property has not been achieved in all Western jurisdictions.

The abolition of slavery and serfdom in the 18th and 19th centuries in the West has meant that with few exceptions competent Western adults are fully capable of holding and dealing with property. Having the capacity, of course, is not the same thing as having property. It is sometimes suggested that if the older law perpetuated inequalities by its explicit recognition of different statuses, the newer law may be doing the same thing by refusing to recognize real social inequalities.

Groups. Despite the tendency of Western legal systems to regard individual ownership as paradigmatic, all Western legal systems allow a number of different forms of group ownership. The categories offered below are not exhaustive, but they give some notion of the various forms of group ownership that may exist.

Concurrent individual owners. All Western legal systems recognize that a group of individuals may each have an undivided ownership interest in a thing. This is the norm, for example, when property is inherited by a group of siblings from a parent, but it is also possible for an individual owner to sell or give a piece of property to a group.

The two most commonly recognized forms of co-ownership in Anglo-American jurisdictions are joint tenancy and tenancy in common. In both forms of tenancy each tenant has the right to possess and the privilege to use the whole thing. If it is physically impracticable for them all to possess or to use the thing, they must agree among themselves who will have possession in fact, since all have possession in law. If they cannot agree, one or more of them may petition the court to have the thing partitioned among them. If partition in kind cannot be had, the court will order the thing sold and the proceeds to be divided among the erstwhile cotenants.

Joint tenancy and tenancy in common

The two forms of cotenancy differ when it comes to succession and to the power to convey. In joint tenancy, if one of the joint tenants dies, the remaining tenants succeed to his share (moiety). In tenancy in common, if one of the tenants dies, his heirs or devisees succeed to his moiety. In joint tenancy, if one of the joint tenants conveys his moiety inter vivos, the conveyance destroys the survivorship interest of his cotenants so far as that moiety is concerned. The conveyee takes not as a joint tenant but as a tenant in common with the other tenants. In tenancy in common, however, conveyance operates like succession. The conveyee takes the same undivided interest that the conveying tenant had.

Civil-law systems recognize a form of co-ownership similar to the Anglo-American tenancy in common (*co-propriété; Miteigentum*). It is not possible in the civil-law systems to hold property in a form in which one's cotenants automatically succeed to it. French law, like Anglo-American, allows co-owners to demand partition of a cotenancy and is hostile to attempts to restrict this power. German law, however, has a form of cotenancy (*Gesamthandeigentum*) in which the cotenants cannot partition the tenancy property, although they may alienate their shares. This form of cotenancy is used for many kinds of partnerships, including the partnership of coheirs that exists until the deceased's estate is settled and divided.

At English common law, partners held partnership property in their individual capacities. They were obliged to account to their partners for profits earned from it, but the ownership interest was in the partner individually, not in the partnership. The common-law rule prevails in England today. In many American jurisdictions, however, legislation allows the partners to hold partnership property in a form of cotenancy, known as tenancy in partnership, which is quite similar to the German *Gesamthandeigentum*. Roman law treated ownership by partners in a way similar to the English common law, but that rule has, in general, not survived in the modern civil law. Those civil-law countries that do not recognize a form of ownership like the *Gesamthandeigentum* tend, like the French, to recognize the property-holding capacity of the partnership itself. Thus, partnerships in these countries are treated like corporations (see below) for property-holding purposes.

Marital owners. English common law, after some hesitancy, adopted a regime of separate marital property. The wife had her property, the husband his. The only things that they owned together were things that had been conveyed to them together in a form of tenancy known as tenancy by the entirety (which still exists in a number of American jurisdictions). Tenancy by entirety is like joint tenancy in that the surviving spouse takes the whole of the property upon the death of the other spouse. It differs from joint tenancy in that it is not possible for one of the spouses to convey his or her interest so as to defeat the survivorship right of the other.

Tenancy by entirety

At common law, each spouse's separate property in land was subject to an expectancy in the other spouse. If the husband survived the wife and if a living child had been born to the marriage, the husband was entitled to curtesy, a life estate in all the lands of which his wife was seised during the marriage. If the wife survived the husband, she was entitled to dower, a life estate in one-third of all the land of which her husband was seised during the marriage.

The 19th-century Married Women's Property Acts had little effect on this basic system. The acts removed the husband's power to control his wife's property during the marriage, and they made it clear that the wife had the capacity to own personal property separate from her husband, but the basic scheme of separate property, curtesy, and dower remained.

What changed the system was the great increase in the amount of wealth held in personal property. Such property was the subject of neither dower nor curtesy. In the United States, although not in England, most jurisdictions provided for a forced share for widows. Under the forced-share system, the widow could waive her dower and any interest that she stood to take under her husband's will and take instead a share (normally a third or a half) of her husband's net estate.

The movement for equality for women in the latter half of the 20th century wrought another change in this system. Today, most American jurisdictions provide for a forced spousal share rather than simply a widow's share. Interests denominated dower and curtesy still exist in some jurisdictions, but they have been made sex-equal.

In the civil-law jurisdictions and in eight states of the United States, a different system of marital property prevails. As in the common-law system, husband and wife each have their separate property, but this is only the property they had prior to the marriage or property they received by gift or inheritance during the marriage. All property that is the result of earnings of either spouse during the marriage is community property, as are, in some of the civil-law jurisdictions, all movables. Separate property descends to the heirs of the spouse who holds the property, but community property is generally divided in half upon the death of the first spouse to die. Half of it goes to the surviving spouse and half of it to the heirs of the first-dying spouse. Other community-property jurisdictions give the first-dying spouse's portion of the community to the surviving spouse, at least in the absence of a testamentary disposition to the contrary.

Community property

Both the common-law and the community-property systems arose at a time when divorce was not as common as it is today. In common-law property jurisdictions the tendency now is to allow the judge wide discretion to divide the property of a divorcing couple without regard to who holds title to what. In community-property jurisdictions the tendency is to divide the community and to leave the separate property with the spouse who has title to it.

The importance of marital property for the concept of property in the West cannot be overestimated. Although spouses have some power to change their marital property arrangements by private agreement, most married people in the West today live under a regime either of community property or of separate property subject to division upon divorce and to a forced share in the surviving spouse. One might well question to what extent any Westerner who is married can be said to have individual property when his or her spouse has so much of a stake in it.

Corporate owners. Throughout the West the vast bulk of productive assets are owned by fictitious legal persons (corporations, companies, *sociétés, Gesellschaften*), created under general incorporation statutes that allow such fictitious legal persons to engage in a wide variety of profit making and, frequently, of eleemosynary endeavours. This development is of relatively recent vintage, but it is so common today that it needs hardly to be stated.

What does need to be stated is the consequence of this fact for property law. If one asks who has the right to possession, privilege to use, and power to convey property of a corporation, the legal answer is that the corporation does, just as if it were an individual. But a corporation is not an individual; it is a fiction. People act collectively through a corporation. The seeming simplicity of corporate ownership masks a variety of interests.

Throughout the West the model of the corporation is that of shareholders owning the corporation. In the most common types of corporation the shareholders choose the directors of the corporation, who, in turn, choose the management of the corporation, who direct its day-to-day affairs. But the shareholders do not own the property of the corporation. They own a piece of the corporation itself; they have neither the right to possess, the privilege to use, nor the power to convey the corporation's property. Only if the corporation is liquidated, and after the corporation's creditors have been paid, are the shareholders able to claim any portion of the corporation's property.

The model of a corporation

Nor do the directors own the property of the corporation. According to the model, they manage the affairs of the corporation in a fiduciary capacity looking after the interests of the shareholders. They do have the right to possess, the privilege to use, and the power to convey the corporation's property, but they are subject to an overriding obligation to the shareholders. (Increasingly, in the modern world they may be subject to overriding obligations to others than the shareholders: creditors, workers, or the public at large.)

The managers of the corporation would seem to be the least likely candidates for the label owners of the corporation's property. According to the model, the managers are but the hired servants of the corporation. They are not only bound by fiduciary duties to the corporation, but they are also subject to the directors' control. But the larger and more complex the corporation, the more likely it is that the effective right to possess, privilege to use, and power to convey the property of the corporation will be exercised by the management of the corporation, with the express or tacit consent of the directors and with the much more notional consent of the shareholders.

Community or state owners. In every Western legal

system certain tangible things, such as water, air, or wild animals (see above), are withdrawn from private ownership. Modern Western law tends to regard these things as belonging to the state or the community.

Furthermore, certain things that are not withdrawn from private ownership can at any time happen to belong to the community, to the state, or to some governmental entity. Some of these things, like public highways or public parks, may be open to the public generally, at least under certain conditions; some of them may be owned by the state in a manner quite similar to things that are in private ownership, like government office buildings or government-owned enterprises.

Finally, the community or the state may have interests in things that are owned privately by someone else. Offensive land uses may be abated by a public officer acting on behalf of the community or, in some situations, by any affected member of the community suing as private attorney general on behalf of the community. The number and types of land uses that are deemed offensive has increased notably throughout the West with the increase of concerns about the environment (see below).

The major difference between socialist legal systems and those of the West is that socialist legal systems reserve a large amount of property for community ownership.

Socialist legal systems

How much property is so reserved varies from socialist country to socialist country, and socialist legal systems are in a considerable state of flux. At the end of the 1980s, however, some or all of the following items were withdrawn from private ownership in most socialist countries: (1) the means of industrial production, or at least the basic means of production, (2) land (this varies considerably), (3) natural resources, (4) sources of energy, (5) the transportation infrastructure, (6) financial institutions (for example, banks and insurance companies), (7) cultural institutions (theatres, schools, the communications media), (8) rented houses, (9) public utilities, and (10) consumer goods manufactured by state-owned enterprises that have not yet been transferred to individuals.

The fact that this large amount of property is vested in the community does not mean that socialist governments manage it all. In all socialist countries much of this property is entrusted to entities that have some independence from the government, although they are subject to overall direction by the government. Such entities are generally of two types: state enterprises, which roughly correspond to government-owned corporations in the West, and cooperatives, which tend to operate in the fields of agriculture, fishing, crafts, consumer buying, or housing. In contrast to Western state-owned enterprises, socialist state enterprises generally do not have title to their property, but their ability to manage it makes their control over it not far different from that of Western state-owned corporations. In some socialist countries cooperatives may have title to the property that they use for their enterprise; in others it is simply entrusted to them.

THE CONCEPT OF OWNERSHIP

Everything must have an owner. It is tempting to define ownership as the sum of the possible rights, privileges, and powers that one might have in a thing. But it is rare in today's world for any one person to have all the legally possible rights, privileges, and powers in a thing. To say that one does not own a thing unless one has all possible interests in it would leave many things without an owner.

Western law has taken a different tack. It has tried to identify an owner or owners of everything. Many legal systems regard the state or the community as owner of those things for which no other owner can be identified. For those things that are the object of multiple and diverse interests, Western law will identify an owner or owners and then will say that the other interests are held by non-owners. The question, then, is how the owner is to be identified when the full panoply of property rights, privileges, and powers do not exist in one person.

Ownership and possession. Anglo-American law tends to identify ownership with the right to possession. Since Anglo-American law recognizes the possibility of multiple rights to possession, one must further qualify: the owner of property is the one who has the right to possession that is better than anyone else's right to possession. As stated above, it is possible that more than one person can have this right concurrently; for example, it is possible to speak of cotenants as co-owners of the thing. As will be seen below, it is possible for the right to possession to be divided temporally, so that, for example, one person may own a life estate in the property, while another owns a remainder (a right to possession after the death of the life tenant).

The civil-law systems have a more univocal conception of ownership, but it too is closely associated with the right to possession. As in the Anglo-American system, it is possible for more than one person to have the best right to possession, so that co-owners are possible in the civil-law systems as they are in the Anglo-American. Civil-law systems are reluctant, however, to allow the division of the right to possession along temporal lines, so that in the life-estate–remainder example given above, the remainderman is deemed the owner, and the life tenant (usufructuary) possesses on behalf of the owner.

Because Western systems connect ownership with the right to possess, it is possible that the ownership of property will shift when the right to possession and possession in fact are separated for a long time. If person A leaves the tract of land that he owns in the woods, person B may enter into possession of it. That second possession is wrongful as to person A, but person A must act to recover his possession from person B within the period set down in the statute of limitations. In most Anglo-American jurisdictions the statute of limitations on actions to recover land is quite long, 10 or 20 years. But if person A fails to act within the limitations period, his action will be barred.

Statute of limitations

One may ask who then owns the land. In most Anglo-American jurisdictions the peaceable possessor of land has the right to possess that land against all except those who can show a better right to possession. But if person A's right to possession is barred by the statute of limitations, then his claim is not better than that of the peaceable possessor. Thus, the person who has actual possession of land for the limitations period acquires a right to possession good as against the whole world, including the true owner whose claim is now time-barred. This adverse possessor, then, becomes the true owner by passage of time.

In the civil-law countries the vocabulary is different, but the results are similar. With the passage of time (somewhat longer than in the Anglo-American systems), the possessor is said to acquire title by prescription, because the right of the former owner to bring an action to recover his ownership is barred.

The logical result of these last paragraphs has not been accepted with equanimity, particularly where the result would be to give ownership of the land to someone who has no legitimate claim to it. Both the Anglo-American and the civil-law systems purport to require more of the adverse possessor or the person who seeks to acquire title by prescription—that he be in good faith, for example, or that he have a claim of right. If, however, the claim of the true owner is indeed time-barred, then the only other alternative would be to hold that the land somehow belongs to no one. That possibility would violate the principle that everything must have an owner. In general, then, once the true owner's claim is time-barred, ownership devolves on the person who has the best possessory right.

Ownership as the absolute right to possession. One may thus define ownership in the same way that the legal philosopher Felix Cohen defined property: "That is property to which the following label can be attached: To the world: Keep off X unless you have my permission, which I may grant or withhold. Signed: Private citizen. Endorsed: The state." Cohen, however, goes on to warn that all the terms of the definition "shade off imperceptibly into other things." Consider, for example, the large range of possibilities encompassed in the phrase "permission, which I may grant or withhold." In all Western legal systems there are a number of situations in which the law will either assume that permission has been granted or will require the private citizen to grant his permission. The situations tend to be dramatic: Firefighters, for example, are usually allowed to enter private property to prevent the spread

of a fire and frequently are authorized to destroy private property in order to prevent the spread of a fire.

Possessory rights versus civil rights

In the 1960s a number of U.S. Supreme Court cases starkly posed the conflict between the property owner's right to exclude and civil rights, in the context of "sit-ins" in restaurants that were excluding customers on racial grounds. These cases suggested, if they did not quite hold, that in this context the possessory right of the restaurant owner would have to yield to the civil-rights claim of those sitting in. In the same period a number of courts held that owners of farms could not exclude visitors from agricultural migrant labour camps.

The conflict in these cases between property rights and civil rights was made starker by the practice in the United States of treating social issues as constitutional controversies. The issue, however, of the use of property to discriminate against members of the society whom the property owner disfavours is present throughout the Western world. Ultimately in the United States the problem of restaurant sit-ins was resolved by national legislation that made it the duty of anyone providing food or lodging to serve all comers without regard to race. Similar legislation exists in many Western countries, as does legislation allowing access to premises in which workers are employed.

DIVISIONS OF OWNERSHIP

Spatial divisions. All Western legal systems allow the owner of property to divide it along spatial lines. Such divisions may be unwise, for example, where the resulting piece of land has no access to a public right-of-way (see below). In the case of land, public regulation may prevent the division. The basic principle, however, is uncontroverted. If person A owns a piece of land or personal property, he is normally permitted to divide it, giving a piece of it to person B, and either keeping the other piece for himself or giving it to still a third person.

A somewhat different set of problems arises when the desired division is vertical rather than horizontal. By and large Anglo-American law allows such vertical divisions, so that one person may own the mineral strata underneath land, another the surface of the land, and the third the air rights. The civil-law systems have had some difficulty with this type of division of ownership, because of the medieval maxim: "Cuius est solum eius est usque ad coelum et usque ad inferos" ("Whoever owns the soil owns all the way to heaven and all the way to hell"). In both systems modern legislation has made possible, for example, ownership of an apartment on the 30th floor of a building.

Condominium and cooperative ownership

Condominium ownership is more complicated, because the condominium owner owns not only the area within the four walls of his apartment or house but also access rights and privileges to use common areas and utilities. Cooperative ownership avoids this complexity by having each of the cooperators own a share in a corporation. The corporation, in turn, allows the cooperators to possess their dwelling units, while retaining the title to all the property.

Temporal divisions. Anglo-American law is notorious for the number and complexity of temporal divisions of ownership it allows. The English law on the topic was considerably simplified in 1925, when it became impossible to have legal ownership divided temporally other than between landlord and tenant. English law, however, continues to allow complicated temporal divisions of beneficial interests in trusts, allowing, therefore, a temporal division in the equitable but not the legal ownership (see below). In many of the remaining Anglo-American jurisdictions, temporal division of the legal ownership of land is still possible, although increasingly undertaken by way of trust.

Life estate and remainder. One of the possible temporal divisions of ownership in Anglo-American law, the life estate and the remainder in fee, has already been considered. In such an arrangement the life tenant has the right to possess the land for his natural life. He may use the property, but he may not impair its capital value (commit waste). He may convey his interest, but he may convey no more than what he has, an interest limited by his life. Hence, his conveyee receives an estate limited by the life of the conveyor (estate pur autre vie). Common-law dower and curtesy are types of life estates.

The remainderman has a right to possession that commences upon the death of the life tenant. He may not use the land until the life tenant dies but may sue the life tenant if the life tenant commits waste. Since the remainderman's interest is an interest in fee, his interest will pass to his heirs or devisees if he dies before the life tenant. The remainderman may also convey his interest inter vivos, subject to the life estate. If the original owner in fee conveys a life estate to someone else and retains the rest of the property in himself, the retained interest is called a reversion. For most purposes reversions have the same characteristics as remainders.

A number of variations on the basic pattern of life estate and remainder are possible in Anglo-American law. There may, for example, be successive life estates: "to my wife Edith for her life, remainder to my son George for his life, remainder after George's death to George's children."

Contingent interests. Not only is it possible to create successive interests in land in Anglo-American law, it is also possible to create interests that are subject to express contingencies. Thus, in the example given above, the donor could make the remainder in George contingent upon George's having attained a specific age, say 21, at the time of the death of the previous life tenant.

Not only is it possible to make future interests subject to contingencies, it is also possible in most Anglo-American jurisdictions to make present interests in fee subject to contingencies. Thus, it is possible, for example, to grant a fee interest subject to the contingency that the land be used for school purposes and to provide for a forfeiture of the interest if it is not so used (fee simple determinable, fee simple subject to a condition subsequent).

The rule against perpetuities. The rule against perpetuities limits the number of successive interests that may be created in property by requiring that all interests vest, if at all, within lives in being plus 21 years. The operations of the concept of vesting are complicated, but basically the rule requires that any contingency with regard to a future interest must necessarily occur or fail to occur within the perpetuities period or else the interest will be held void from the beginning. For purposes of the rule, a contingency is not only an express contingency, such as the one in the example above that George have reached the age of 21, but also an implied contingency, such as the identification of George's children.

The concept of vesting

Both of the contingencies in the example given above meet the requirements of the rule. George is already a life in being, so he will obviously become 21 during his lifetime or fail to become 21. The children of George are not necessarily lives in being at the time of the grant, but they will all be identified during George's life (plus a possible period of gestation, which is allowed to extend the lives in being).

If, however, the gift had said "remainder to my children for their lives, remainder after the death of my last surviving child to my children's children," then the interest in the donor's grandchildren would be void under the rule. The donor could have another child after the effective date of the grant, which child could have a child more than 21 years after the donor, his wife, and all the child's living sibs were dead. That grandchild's interest, then, would vest outside the period of the rule. In most situations, the consequences of voiding the gift to the grandchildren will be to leave a reversion in fee in the grantor or his heirs, following the two valid life estates.

There are numerous modern proposals for reform of the technical operations of the rule against perpetuities. The details of its operation have changed over time and are likely to change further in the future. Its principle, however, has remained quite constant since *The Duke of Norfolk's Case* (see above). It compromises the conflict between the power to convey of the present holder of the property and the power to convey of those to whom the present holder conveys by allowing the present holder to restrain the power to alienate of those who are alive at the time he makes the conveyance and, in effect, that of their children until they reach the age of majority (21). After that full power of alienation must be allowed.

Civil law. Some, although not all, of the arrangements

described above are possible in civil law. The major distinction between Anglo-American and civil law in this regard is that civil law normally does not regard such arrangements as involving divisions of ownership. Thus, the usufruct, the device in civil law that most closely corresponds to the life estate of the Anglo-American law, is regarded not as a form of ownership but as a right in the thing of another (*jus in re aliena*). Although the usufructuary normally does not have the right to possession in civil law, he is normally given possessory remedies against third-party wrongdoers. All in all one may question how different the practical position of the usufructuary is from that of the life tenant in Anglo-American law, despite the substantial conceptual differences between the two systems.

Usufruct

Even in the area of conditional gifts, the differences between the two systems are not as great as they might seem. True, in civil law the basic principle is that gifts cannot be conditioned. The donor must give outright or not at all. There are, however, exceptions in civil law that derive from the medieval Roman law of fideicommissary substitutions. The rules are complicated and vary from jurisdiction to jurisdiction. In the French system, for example, it is possible to make a will giving property to one's children and requiring that they turn it over to their children. In German law, it is possible to appoint successive heirs, so long as the succession occurs within 30 years of the death of the testator.

There is no equivalent in the civil law of the fee simple with a forfeiture clause. Thus, a grant subject to the condition that the land be used for school purposes is not possible in civil law, although there are ways of achieving similar results in civil law, at least for limited periods of time.

Landlord and tenant. In Anglo-American law present possessory interests less than the fee need not be limited to the life of the holder of the interest; they may also be limited to a specific term of years or to a renewable term. Such a transaction creates the relationship of landlord and tenant. The tenant may have a possessory interest for any specific term, such as a month, a year, 5 years, or 99 years. The tenant may also have an interest for a specific term that is renewed automatically unless the landlord or the tenant gives notice within a fixed period before the term expires (periodic tenancy). Thus, tenancies can be arranged, for example, from week-to-week, month-to-month, or year-to-year. It is also possible to have a tenancy for no fixed term but subject simply to the will of the landlord and tenant (tenancy at will). Either landlord or tenant may give notice to the other at any time to terminate the tenancy. (In many jurisdictions tenancies at will are subject to statutory regulation concerning the time of the notice to terminate, thus making them more like periodic tenancies.)

Similarly, civil law allows the creation of landlord-tenant relations. Although the categories of tenancies recognized in Anglo-American law do not exist in civil law, it is possible to create by private agreement most of the landlord-tenant arrangements that Anglo-American law recognizes. What is different in civil law is the conception of the relationship between landlord and tenant. In modern civil law, as in Roman law, the tenant does not have the right to possession, the landlord does. Since the landlord has a contractual obligation to allow the tenant to possess, the practical consequences of this conceptual distinction are not great. The most important area where the two systems differ is in the situation where the landlord sells his interest in the land to someone else. In Anglo-American law, the tenant has an enforceable possessory interest against his new landlord. In civil law, the tenant's remedy is against his old landlord. Even this difference has been narrowed by recent legislation in civil-law jurisdictions that allows tenants to sue third parties who interfere with their possession in fact.

Recent developments

Throughout the West in the second half of the 20th century there have been substantial changes in the law governing landlord-tenant relationships. These changes have most notably affected the law concerning residential tenancies, particularly tenancies in urban apartments. Some jurisdictions have also made extensive changes in the law governing agricultural tenancies. By and large, the law of commercial tenancies has been left to private agreement. Commercial leases, therefore, continue to use traditional forms, with the terms being negotiated between the parties.

In the United States the changes in residential landlord-tenant law were initiated by the courts. Legislation has followed, confirming and frequently going beyond what the courts have done. In the rest of the West, the initiative has been taken by the legislatures, although the end results have been quite similar.

The principal changes are in four areas: security of tenure, legally imposed terms in the lease (particularly concerning the maintenance of the premises), regulation fixing rents, and direct government ownership of residential premises.

The law of landlord and tenant that the West inherited from the 19th century gave the tenant security of tenure only so long as the lease lasted. The parties could, of course, renew the lease once it was at an end, but such renewals required the consent of both landlord and tenant. In fact, however, residential leases tended to be renewed, with or without some adjustments in rent as the market dictated.

World War II precipitated a housing crisis throughout the West. The wartime economy did not produce enough new housing, and the end of the war saw a great increase in population. The end of the war also saw massive shifts in population to urban areas and great changes in patterns of land use. Landlords responded to these market forces by refusing to renew leases without sharp increases in rents or by refusing to renew leases at all, so that they might convert housing to more profitable uses.

Different jurisdictions in the West responded to this crisis in different ways. In England and on the European continent, where much of the housing stock had been destroyed in the war, direct provision of housing by the government began to play an important role in the housing market. There was also a tendency in these countries to fix the price of private rental housing and to require that private landlords renew leases so long as the tenant continued to pay the rent thus fixed.

In the United States more reliance was placed on the private market. Government-subsidized financing of housing encouraged construction of new housing by the private sector. Fixed rents existed in some jurisdictions but were intended as a temporary measure. Home ownership was, and remains, more common in the United States than it is in Europe.

Landlord and tenant

In the United States, changes in the law of landlord and tenant came in the 1960s and early '70s, when the country came to focus on the deteriorated conditions of housing that existed for the urban poor, those who had been left behind as the middle class had moved to newly constructed housing in the suburbs. Urban housing for the poor was frequently maintained at levels far below what was required by local regulation (see below), but enforcement of these codes was sporadic.

Faced with this situation and with considerable legislative inertia, American courts came to read the provisions of the housing code into the lease. *Javins* v. *First Nat'l Realty Co.*, 428 F.2d 1071 (D.C. Cir. 1970), for example, requires that every residential lease have within it an unwaivable warranty of habitability, requiring the landlord to maintain the premises up to the standard of the local housing code. If the landlord does not maintain the premises up to this standard, the tenant may withhold rent and the landlord cannot evict him for nonpayment of rent.

Legislation in the 1970s tended to confirm the results of *Javins*. Other legislation and decisions gave residential tenants more security of tenure. A number of urban jurisdictions that had abandoned rent control reinstated it. While there are still considerable differences between the American and European patterns of regulation of the residential landlord-tenant relationship, the trend is toward convergence. Everywhere the essential terms of the relationship are increasingly fixed by law; the tenant's interest in his dwelling has become more secure, and the landlord is seen less as an owner than as the provider of a public utility. Recently a countertendency to rely more

on market forces is observable throughout the West, but this has so far only modified the previous trend, it has not reversed it.

Divisions as to rights, privileges, and powers. *Trusts.* Anglo-American law recognizes another possible division of ownership, that between the power to manage property and the privilege of receiving the benefits from it. This division, known as the trust, is of great practical importance in Anglo-American law. The trust device is used in a wide variety of contexts, most notably in family settlements and in charitable gifts. In the area of family settlements it has largely replaced the legal life estate and remainder.

Fundamental to the notion of the trust is the division of ownership between legal and equitable. This division had its origins in separate English courts. The courts of common law recognized and enforced the legal ownership; the courts of equity recognized and enforced the equitable ownership. The conceptual division of the two types of ownership, however, survived the merger of the law and equity courts. Thus, today legal and equitable interests are usually enforced by the same courts, but they remain conceptually distinct.

Legal and equitable ownership

The basic distinction between legal and equitable ownership is quite simple. The legal owner of the property (trustee) has the right to possession, the privilege of use, and the power to convey those rights and privileges. The trustee thus looks to all the world like the owner of the property. He so looks to all the world except for one person, the beneficial owner (beneficiary, *cestui que trust*). As between the trustee and the beneficiary, the beneficiary gets all the benefits of the property. The trustee has a fiduciary duty to the beneficial owner to exercise his legal rights, privileges, and powers in such a way as to benefit not himself but the beneficiary. If the trustee fails to do this, the courts will require him to pay over what he has earned for himself to the beneficiary, and may, in extreme cases, remove him as legal owner and substitute another in his stead.

Divisions between legal and beneficial ownership are normally created by an express instrument of trust. The maker (settlor) of the trust will convey property to the trustee (who may be an individual or a corporation, such as a bank or trust company) and instruct the trustee to hold and manage the property for the benefit of one or more beneficiaries of the trust.

Trust instruments can be quite complicated. They may provide for succession among the trustees and for succession among the beneficiaries. They may give the trustee considerable discretion in managing the property and in paying out the benefits to the beneficiaries. In many jurisdictions the beneficiary's interest may be insulated from the claims of his creditors (spendthrift trust, protective trust). There are only three basic limits on private trusts in Anglo-American law. The first is that there must be some diversity between the legal and the beneficial owners. If they are or become identical, the interests merge and the trust is dissolved. The second limitation is that the beneficiaries must be identifiable. The third is that the power of the beneficiaries to dissolve the trust cannot be suspended for longer than the perpetuities period. Thus, future interests in a trust, like future interests at law, must vest within a life or lives in being at the creation of the irrevocable character of the trust plus 21 years.

In the arrangement described above, the donor gave property to his spouse, Edith, for her life, remainder for life to his child George, if George reached the age of 21, with a remainder after George's death to George's children. This arrangement could be accomplished by means of a trust. The donor (settlor) would give the property in trust to a trustee with the following instructions: (1) pay the income from the property to Edith for her life, (2) upon Edith's death, accumulate the income or expend it on behalf of George until George reaches the age of 21, (3) when George has reached the age of 21, pay the income to George during his lifetime, and (4) upon George's death, divide the principal of the trust among George's then-living children.

Using a trust rather than the legal life-estate-remainder arrangement allows a separation of the management of assets from the enjoyment of them. Getting assets into the hands of professional or semiprofessional managers frequently allows them to be managed more competently than they would be by some or all of the beneficiaries. Further, the assets themselves are not locked into the arrangement. If it makes sense to sell a piece of land or to shift a portfolio from bonds to stocks, the trustee does this, and he can give good title to the assets. In the legal life-estate-remainder arrangement all the holders of interests in the assets have to consent to the conveyance of the property, a cumbersome procedure at best and particularly difficult when some of the beneficiaries are minors or are unascertained. Further still, use of the trust allows the beneficial interest in property to pass from generation to generation without the property having to pass through probate, an awkward and time-consuming process in many Anglo-American jurisdictions. Finally, use of the trust sometimes allows for a savings in taxes.

Civil-law parallels to the trust

There is no precise equivalent of the trust in civil law. (Islāmic law knows an institution, the waqf, that is somewhat like the Anglo-American trust.) Some modern civil-law systems have created an institution like the trust, but this has normally been by adapting trust ideas from the Anglo-American system rather than by developing native ideas.

Most of the uses to which the Anglo-American trust is put are achieved in civil law in other ways. For example, the charitable trust of Anglo-American law has a quite close analogy in the civil-law foundation (*fondation, Stiftung*). Of the purposes for private express trusts mentioned above, lawyers on the Continent get professional management for assets by turning them over to managers who are paid a fee for their services. Since the number of possible outstanding interests in a given piece of property is more limited in the civil law than it is in Anglo-American, it is less necessary to have a trustee who can give good title to the whole of the property. Probate avoidance is rarely an issue on the Continent, because civil-law systems of probate are normally far less cumbersome than the Anglo-American systems. Thus, as in so many other areas in comparative Western law, it turns out that some of the needs that the Anglo-American trust serves are not needs in the civil-law systems because of structural differences in the systems and that the remaining needs are served by other devices.

Security interests in property. Another division of the rights, privileges, and powers of ownership exists in all Western legal systems—the division that occurs when an owner makes use of his thing as security for a loan or other obligation. Both English common law and Roman law recognized a number of different ways in which property could be used to secure a transaction. In general, the devices used at common law transferred ownership or a real right in the thing to the creditor, while those used in Roman law kept the ownership of the thing in the debtor. The focus, then, of the development of the common law was on ensuring that the debtor got his ownership back if he discharged the obligation, while the focus of the development of Roman law was on ensuring that the creditor's rights were protected if the thing ended up in the hands of some third party.

From these two quite different starting points a long course of development in both the Anglo-American and civil-law systems has arrived at a point where, despite great differences in vocabulary and conceptualization about property used in a secured transaction, there is considerably less difference in practical result. Both systems recognize arrangements between debtor and creditor in which the ownership of the thing is nominally transferred to the creditor, but the creditor's ability to deal with the thing is limited in such a way that the ownership will revert to the debtor so long as the debtor discharges his obligation. Both systems also recognize arrangements in which the creditor does not receive an ownership interest in the property but where he receives sufficient rights against the debtor so that he is secure if the debtor does not discharge his obligation.

Mortgages

In both systems the most complicated, and historically the most important, security devices have to do with land—

the mortgage of the common law and the *hypothec* of the civil law. In the mortgage of the common law the debtor (mortgagor) conveyed his land to the creditor (mortgagee) subject to the condition that the land would automatically revert back to the debtor if the debtor discharged his obligation by a certain date. The debtor, however, remained in possession of the land, and the practice of allowing the debtor to remain in possession became an obligation of the creditor to allow the debtor to possess the land and finally a right in the debtor to possess the land so long as the debtor was not in default on the debt. If the debtor defaulted, the creditor's right to possess became perfected, and he could enter and use the land for himself or sell it as he wished. The debtor's interests were extinguished.

The equity courts intervened on the side of the debtor. Equity first gave the debtor a right to redeem the property by paying the amount that was owing, even if he had defaulted on the debt. In order to sell the property, creditors were forced to bring an action in equity to foreclose the debtor's equity of redemption. As a condition of foreclosure, equity gave the debtor a right to the proceeds of the sale to the extent that the sale realized more than the outstanding debt. Legislation in the 19th century extended the debtor's right to redeem even after the creditor had foreclosed. Finally, in some jurisdictions, legislation required that the creditor sell the property after he had foreclosed, and in some of these jurisdictions the sale had to be conducted by a public official.

At common law the debtor could not transfer legal title to his property to third persons because he did not own it. (He could, however, convey his equity of redemption.) This meant that a purchaser in good faith might end up with nothing even though the mortgagor looked to all the world like the owner of the property (he was in possession and could normally produce evidence that the property had been transferred to him by a previous owner). In order to protect third-party purchasers, most Anglo-American jurisdictions have public offices in which mortgage transactions can be recorded (see below). Under the prevailing legislation, the purchaser who takes in good faith from a mortgagor in possession acquires good title against the mortgagee unless the mortgage is recorded.

In most Anglo-American jurisdictions these developments in mortgage law have led to a reconceptualization of the mortgage itself. If the mortgagor has the right to possession, an equity of redemption, and the power to convey good title to the land (subject to the mortgage if the transaction is recorded), his interest looks more like that of an owner than do those of the creditor, despite the fact that the mortgage deed says that the creditor is the owner. Other jurisdictions retain the notion that the creditor is the owner subject to all the qualifications offered above. There is little practical difference in result in the two types of jurisdiction.

Starting from very different premises the civil-law systems have arrived at much the same result. The debtor has the right to possession and privilege of use of the property unless and until he defaults. If he defaults, the creditor may, depending on the jurisdiction, either take possession of the property or force a sale of it. The debtor's interest in the proceeds of the sale over and above the outstanding amount of the debt is everywhere protected. In some jurisdictions the debtor may also be given a grace period within which he can redeem the property after default. Registration of security interests is virtually universal. If the interest is registered, the creditor's interest survives any transfer of the property, even to a good-faith purchaser without actual notice of the security interest.

Security interests in movables

Security interests in movables have a somewhat different history. In the Anglo-American system security interests in personal property were developed largely by the equity courts, aided in the 19th and 20th centuries by legislation. The result is a quite complex branch of what is normally called commercial law. Suffice it to say here that it is possible to have arrangements much like a mortgage whereby the debtor retains possession of the property subject to a security interest in the creditor (chattel mortgage or conditional sale) or to have the creditor take possession of the property subject to the debtor's right to redeem it by paying the debt (pledge or pawn). In some jurisdictions, notably England, the debtor will lease the property from the creditor (who is also normally the seller), his title becoming absolute when the payments have been made (hire purchase). In the United States the differences among the various types of personal property security agreements have been considerably reduced by uniform legislation that deals with all of them under one heading.

On the Continent the pledge or pawn (pignus) was historically the chief security device for movables. Under this device the right to possession of the movable was in the creditor, although possession in fact might not be. Financing devices for merchants are handled in separate codes of commercial law, where the devices tend to be similar to those of the Anglo-American chattel mortgage or conditional sale. Modern consumer credit law has produced a number of devices, some of them representing developments from the civil law of pledge, some more closely resembling the English hire purchase.

PROTECTION OF PROPERTY INTERESTS

Public law protections of property. *Criminal.* If person A takes the property of person B without his permission and with the intent permanently to deprive him of it, that is theft, a concept that is universal in the West. In the Anglo-American systems today theft is everywhere a crime, the penalty normally being dependent on the value of the thing stolen and the circumstances under which the taking occurs. Modern Anglo-American criminal codes tend to subdivide theft in ways that reflect their common-law background. Larceny is the simple taking of personal property or money from the possession of another with the intent permanently to deprive the possessor of it. Burglary is larceny aggravated by the fact that it is achieved by breaking and entering premises in order to accomplish it. Robbery is larceny aggravated by the fact that it is achieved by the exercise of force or threats of force against the possessor. Embezzlement is a wrongful taking of property by someone (like a bank clerk) who is already rightfully in possession of it.

The civil-law criminal codes do not observe the Anglo-American distinction between larceny and embezzlement. Otherwise, the criminal prosecution of theft in civil law is quite similar to that in the Anglo-American systems. An intent to deprive (*animus furandi*) is required. The penalty will vary depending on the value of the thing stolen and will be aggravated if the theft is accompanied by wrongfully entering premises or by the exercise of force.

Land cannot be stolen in either Anglo-American or civil law. Wrongful entry onto land may be punished in Anglo-American law by statutes regulating criminal trespass. Deliberate damage to another's land may also be punished criminally, particularly under modern regulatory statutes concerning the environment.

Regulatory. The 20th century has seen the rise of an extensive body of regulatory law concerning the use of property, particularly of land. This law is treated in more detail below, but it is well to point out here that the effect of such regulatory law is to protect the property interests of those members of the community whose property would be adversely affected by the land use proscribed by the regulation. Thus, if an environmental law prohibits the emission of pollutants from a smokestack, that law protects the interests of those on whose land the pollutants would otherwise descend.

In some circumstances some Western jurisdictions allow those adversely affected by the violation of such regulations to sue the violators directly. In other circumstances and in other jurisdictions such standing to sue is not allowed, but the adversely affected individual may bring an administrative proceeding to compel enforcement of these regulations. Even if no private enforcement is allowed, the facts that the regulation exists and that its enforcement by public authorities can normally be expected changes the property interests, in the definitional sense, not only of the property owner whose privilege of use is limited by the regulation but also of those who are benefited by the regulation.

Private law protections of property. The protection of

property in civil procedure has a long history in both the Anglo-American and civil-law systems. Both procedures are strongly affected by the fundamental distinction that Roman law made between actions in personam and those in rem and by the distinction that the medieval civilians made between actions to establish ownership (petitory actions) and those to recover possession (possessory actions).

Anglo-American law. In the Anglo-American systems the basic action for vindication of an ownership interest in land is usually a modern action derived from the common-law action of ejectment. This action results in the successful plaintiff being restored to the physical possession of the land. After some controversy, still not completely settled, it was decided that the plaintiff in ejectment need not prove title good as against the whole world but simply a relatively better right to possession than the defendant. The operations of this action thus fit into the Anglo-American concept of ownership as a relatively better right to possession.

Ejectment

For the owner seeking a judicial declaration of his title to land, most of the Anglo-American systems provide an action derived from the equity action to quiet title. This results in a declaratory judgment as to the state of the title. The procedural difficulties of bringing this action make it a decided second best to ejectment, but sometimes it is the only remedy available (where, for example, the plaintiff is already in possession but the defendant claims ownership or some lesser interest and hence is hampering the market value of the plaintiff's land).

As a general matter, where the action of ejectment is not available, equity courts, or their modern descendants, will protect the plaintiff who has established that he has a property interest in land by issuing an injunction against the defendant who is interfering with the interest.

Because the action of ejectment tries the better right to possession, separate possessory actions for land are no longer a main feature of Anglo-American law. Most jurisdictions do, however, have a statutory possessory action, derived from the English statutes of forcible entry and detainer, in which an owner or prior peaceable possessor can recover possession from one who has taken or who detains possession without pretense of right. These actions are frequently used by landlords to recover possession from tenants who have held over after the terms of their leases have expired and are occasionally used by peaceable possessors who have been ousted from their possession by force.

Possession of land is also protected in the Anglo-American system by civil actions of trespass. Technically, trespass is a personal action, and the successful plaintiff recovers only money damages. Since such actions frequently rest on the right to possession, however, they were used in the past, and in some jurisdictions are used today, to try title.

Historically, the Anglo-American system had no real action to vindicate ownership of movables. Though still technically personal actions, actions concerning movable property have been expanded in Anglo-American law, so that today they serve most of the purposes of the old real actions of the land law. In England, conversion, a descendant of the common-law trover action, is used, coupled with the possibility that in some situations (normally in the case of unique movables) the court may specifically decree the restoration of the thing itself. In the United States the common-law action of replevin was changed to allow the same purpose to be achieved.

Trover and replevin

Civil law. Roman law had an action of *rei vindicatio*, whereby the plaintiff could claim ownership of any thing, movable or immovable. It could be brought only against someone who was withholding possession of the thing from the plaintiff, and it required the proof of ownership by derivation of title from the conceded owner or by demonstrating that the quite strict requirements of usucapion had been met. (Usucapion, a type of prescription, required that the person claiming it have uninterrupted possession of a movable for a year or of an immovable for two years, that he have acquired it in good faith, and that he have a just cause for his acquisition. This last requirement effectively limited usucapion to cases of defective sales and certain types of official orders.) Roman law also offered an *actio publiciana* to the person who had fulfilled all the requirements of usucapion except that the prescriptive period had not run out. A third action, the *actio negatoria*, protected the owner in possession against the claims of one out of possession, allowing the owner to obtain what was in effect a declaratory judgment as to his rights.

Roman procedure for protecting possession was sharply distinguished from the actions that protected ownership. The interdict *unde vi* protected the peaceable possessor of land who was dispossessed by force; the interdicts *utrubi* and *uti possidetis* protected the possessor either of movables or immovables whose possession was being disturbed by some third party.

Modern civil-law systems retain the distinction that Roman law made between petitory and possessory actions, but the tendency in both cases is toward a procedure of relative rather than absolute rights. Thus, for example, the modern French revendication, while still nominally an action that tries absolute ownership, has in practice become an action that tries relatively better title between the plaintiff and the defendant. Similarly, the French possessory actions of *réintégrande* and *complainte* are available to almost any peaceable possessor to recover something of which he was dispossessed by someone whose claim to possession is inferior to his. The results in the German system are similar, although the German scheme of actions is somewhat closer to that of Roman law. German law also knows an action to correct the *Grundbuch* (see below), which has a somewhat similar function to that of the Anglo-American quiet title action.

Use of property interests

The previous section focused on the right to possession of property. This section focuses on the privilege of use of property—the extent to which the law allows an owner or possessor of property to use the property and how an owner or possessor of property may grant privileges of use to others. The fact that person A's privilege of using his property inevitably conflicts with person B's privilege of using his, if their properties are located near each other, has led throughout the West to extensive limitations on the privilege of use, first in the area of private law and, increasingly today, in the area of public law.

NUISANCE LAW AND CONTINENTAL PARALLELS

At English common law the basic limitations on the privilege of use of property were incorporated in the law of nuisance, the action that a landowner could bring if his privilege of using his land was being interfered with. Historically, nuisance law seems to have been deeply conservative; existing land uses were protected against more recent ones. A hierarchy of land uses favoured residential uses over agricultural and agricultural over industrial. (Commercial uses were sometimes placed after residential, sometimes after agricultural.) The maxim "sic utere tuo ut alienum non laedas" ("use your own thing so as not to harm that of another") expressed this conservative tendency, though it hardly offered a precise solvent for difficult cases.

Today, nuisance law is still used in the Anglo-American system as a means of resolving land-use disputes. The hierarchy of land uses is still employed, tacitly if not expressly; the maxim is still occasionally quoted, and at least in close cases the land use that is prior in time will prevail over subsequent ones. What has changed about nuisance law is the fact that today the element of judicial discretion in resolving the basically unresolvable conflict between two equally privileged land uses is more frankly recognized.

Nuisance is defined as the substantial interference with the plaintiff's use of his land by the unreasonable conduct of the defendant. Each of the qualifying words in the definition can lead to an exercise of judicial discretion. One may ask, for example, whether the harm caused by the defendant's activity is substantial. A judgment is called for, aided, of course, by precedent, but always unique to the given case. Hazards to health, offenses to the sense of smell or hearing, and demonstrated economic loss are

Nuisance

frequently found to be substantial harms. Offenses to the sense of sight and injuries to peculiarly sensitive activities (such as maintaining a mink farm) are much less likely to be found substantial.

The second stage in determining that a nuisance exists requires a finding that the defendant's activity was unreasonable. Unreasonable conduct is a relative matter. It may be unreasonable to engage in heavy manufacturing in a residential area and perfectly reasonable to do so in an industrial area. The care with which the defendant conducts his activities is of relevance, but it is not decisive.

Once a nuisance is found, there still must be, in most jurisdictions, a "balancing of the equities" to determine whether the defendant will be enjoined from his activities or whether the plaintiff will have to content himself with money damages. In recent cases, economic considerations have come to the fore in making this determination. Thus, in a celebrated New York case, the court refused to enjoin the operations of a cement plant that represented a $45 million investment and a large number of jobs for a small community but instead awarded money damages to the nearby residents calculated on the basis of the reduction in the capital value of their houses that would result from the continued presence of the smoke-emitting plant (*Boomer* v. *Atlantic Cement Co.,* 26 N.Y.2d 219, 257 N.E.2d 870 [1970]).

Unlike the English common law, Roman law had no single action whereby a plaintiff could complain of his neighbour's interference with his land use. Various private actions did exist by which a plaintiff could complain of particular noisesome land uses, but the jurists never seem to have generalized about them.

The adoption by modern civil law of the Roman conception of ownership and of substantial parts of the Roman scheme of actions has meant that modern civil law also lacks a unified protection of the privilege of use like that of the Anglo-American nuisance law. In France this lack has been made up by the development of the concept of *abus de droit* ("abuse of right"). The concept has been extensively used in situations where the defendant has employed his land in a given way in order to interfere with his neighbour's land use. The paradigm case came from Colmar in the middle of the 19th century, when the defendant built a large and totally unnecessary chimney on the roof of his house in order to block the light to his neighbour's windows (2 mai 1855, D.1856.2.9). From there the concept has developed so that it may be used in situations where the motives of the defendant are not so obviously malicious as they were in the Colmar case, but it has never involved the French judiciary as much in land-use questions as has the Anglo-American. German law, on the other hand, has developed a concept similar to that of Anglo-American nuisance law, based on the general requirement in the code that one act in good faith and on a specific provision dealing with smoke and noise.

PRIVATE LAND-USE CONTROL: SERVITUDES

Both Roman law and English common law recognized that an owner of land could voluntarily part with a right or privilege with regard to his land so that a neighbour might use the land in a way that would otherwise be actionable. The classic case is the right-of-way, whereby an owner agrees to allow a neighbour to cross his land in order to allow the neighbour to reach his own land. What distinguishes the right-of-way and similar interests from the myriad types of enforceable agreements not to sue is that the right-of-way is a real right; that is, if it is properly created, the right-of-way will remain in effect even when the owner of the burdened land has transferred the land to another.

Right-of-way

Today the category of *jura in re aliena* ("rights in the thing of another") is broader in Anglo-American systems than it is in the civil law. The developments, however, were not entirely independent of each other. The similarity in the two bodies of law will become even more noticeable if, as has been proposed, American law comes to abandon its traditional distinctions between and among easements, profits, real covenants, and equitable servitudes and adopts instead, like the civil law, a general category of servitudes.

Easements and profits. An easement in Anglo-American law is a privilege to do something on the land of another or to do something on one's own land that would otherwise be actionable by one's neighbours (affirmative easement). Exceptionally, it is the right to prevent a landowner from doing something on his land that he would otherwise be privileged to do (negative easement). Examples of affirmative easements include rights-of-way, the privilege of using land for pasture, the privilege of using a wall as a party wall, the privilege of flooding land, and the privilege of maintaining a nuisance on one's own land (for example, a garbage dump or an airport). Examples of negative easements are more restricted. It is sometimes said that there are only four such easements: two being the right to prevent one's neighbour from obstructing the light and the air that normally come to one's property, the third being the right to prevent him from undermining the support for a building, and the fourth being the right to prevent him from changing the course of an artificial stream.

Historically, and in some jurisdictions today, easements had to be appurtenant: the land benefited by the easement (dominant tenement) and the land burdened by the easement (servient tenement), in most situations, had to be adjacent to each other. A separate category of privileges, known as profits, was recognized, and these could be held in gross; *i.e.,* the holder of the profit did not have to be the owner of the adjacent land. Profits gave the holder the privilege of taking something from the burdened land, like timber, game, minerals, or water. The recognition by most Anglo-American jurisdictions that affirmative easements can be held in gross has led to the gradual abandonment of the distinction between affirmative easements and profits.

Creation of easements

Easements may be created by grant, by implication, or by prescription. Normally, the owner of the burdened land will grant the easement expressly. Recordation may be necessary in order to have the grant bind third parties (see below). Where the owner has divided land in such a way that the conveyee has no convenient means of access except across the land retained by the conveyor, the conveyor will be presumed to have given the conveyee a right-of-way across the retained land (easement by implication). The same will often be presumed where the conveyor has left himself totally landlocked (easement by necessity). (In a few jurisdictions statutes compel the same result.) Implication will also be found where there were pipes or paths on the undivided parcel that suggest that the parties to the transaction that divided the parcel intended to subject one parcel to an easement in favour of another. Finally, the continuous and uncontested use of an easement for the period of prescription (normally, the statute of limitations for ejectment actions) can give rise to an easement by prescription.

Real covenants. The common law recognized that under certain circumstances a promise could be made to "run with the land," so that the owner of the estate burdened by the promise would have a duty to perform it. The promise could be either negative (a promise not to do something, like not using the land for commercial purposes) or affirmative (a promise to do something, like maintaining a fence or paying an assessment to a homeowners' association). The conditions under which such covenants would run with the land were, and perhaps still are, complicated. In many jurisdictions the precise contents of the doctrine are not clearly defined. This is because the enforcement of covenants by means of injunction, the equitable servitude discussed in the next section, has largely taken over, as a practical matter, for covenants that run with the land at law.

Equitable servitudes. The equitable servitude is an invention of the English equity courts in the 19th century. This device allows the enforcement of promises (usually negative promises) that neither fall within the traditional types of negative easements nor meet the traditional requirements of covenants that run with the land against the successors in title to the land owned by the original promisors. What is required is that the content of the promise "touch and concern the land," a requirement that allows the court to make a policy determination that a

particular class of promises should be permitted to burden the land and that the person against whom enforcement is sought took the burdened land with notice of the promise. This notice will typically arise from the fact that the instrument in which the promise was made is on the public record, but it has been held that a uniform scheme of development of parcels of land that were once under common ownership is enough to put the purchaser of one of the parcels on notice to inquire whether a promise to develop in this way was made.

The equitable servitude has been of great importance in land development in Anglo-American jurisdictions. By use of this device land can be developed according to a uniform scheme (residences only, residences of a certain size, even residences with a required style of architecture). The use of the equitable servitude as a device for private land-use control is one of the reasons why public regulation of land use is a relatively recent phenomenon in the United States.

Civil law. The Roman law of real servitudes was not well developed. The rustic praedial servitudes consisted of three types of rights-of-way (iter, via, actus) and one type of water right (aquaeductus); the urban praedial servitudes were principally concerned with light to and support of buildings. A generally restrictive attitude towards servitudes is also manifest in the modern civil law. In French law it is not possible to create a servitude that benefits a person rather than a tenement or piece of land; *i.e.*, a servitude must have both a dominant and servient tenement. There can be no servitude requiring the owner of the servient tenement to do something. Within these limits French law allows a servitude to be created for any purpose. The German law is broader. It recognizes the possibility that servitudes may be created to benefit a person rather than a particular piece of land, although the benefit may last no longer than the lifetime of the beneficiary. As in French law there does not seem to be any way in German law to compel the owner of the servient tenement to do something. Thus, there is no category in civil law corresponding to the Anglo-American affirmative real covenant, and the category of equitable servitudes is unnecessary because the general category of servitudes is broader.

Servitudes in civil law

In French law the methods of creating servitudes are remarkably similar to the methods of creating easements in Anglo-American law. German law makes less use of prescription and implication of servitudes than does Anglo-American law, probably because of its reliance on the *Grundbuch* (see below). Both the French and German systems recognize a right-of-way of necessity (*enclave; Notweg*). The parcels need not originally have been in common ownership, but the landowner seeking a way of necessity must compensate the owner of the servient tenement.

Servitude law is not used in civil-law countries as extensively as it is used in Anglo-American. This is probably because civil-law jurisdictions developed public controls on land use earlier than did most Anglo-American jurisdictions.

PUBLIC REGULATION OF LAND USE

Urban planning was known in the ancient world, and particular regulations of land use designed to ensure the health, safety, or sensibilities of neighbours appear wherever human beings live in reasonably close proximity. The amount, however, of such regulation has increased dramatically in the 20th century. As a result, zoning and planning law has become a topic of general concern to the legal profession. The 20th is probably the first century in which it is possible to outline a body of general principles of zoning and planning law rather than simply listing the types of regulations that one or another jurisdiction has seen fit to enact.

Zoning and planning law

Zoning and planning law is also an area in which the basic distinction between Anglo-American and civil law is not particularly useful. Although the concept of public nuisance does not seem to exist in civil law and the constitutional protection given to "property" in the United States has given rise to a somewhat unusual set of limitations on the power of government to regulate land use, the overall picture of public control of land use in the West is more notable for its similarities than for its differences.

Public nuisance. In Anglo-American law the concept of public nuisance serves as a bridge between the private law of nuisance and the avowedly public law of zoning and eminent domain. The concept of public nuisance is closest to that of private nuisance in situations in which a public officer, acting on behalf of the community, brings suit to abate a nuisance that differs from a private nuisance only in that it affects a large number of people. The concept of public nuisance is further removed from that of private nuisance when legislative bodies declare that certain kinds of land use are public nuisances as a matter of law. Traditional legislatively declared public nuisances include the maintenance of houses of prostitution or illegal gambling establishments and illegal sales or consumption of alcoholic beverages. The direct link between public nuisance and zoning and planning law is provided by the fact that in many Anglo-American jurisdictions violations of zoning law are automatically public nuisances. Thus, constructing a building without obtaining the requisite public approval is automatically a public nuisance, which may be abated by the public prosecutor.

Civil law lacks the concept of public nuisance. Civil law, of course, has a large number of prohibited land uses like those described above, and civil law prohibits the construction of buildings without obtaining the requisite permits. Whether the absence of the concept of public nuisance leads to any real differences in result may be doubted. The concept of public nuisance in Anglo-American law has frequently been criticized as being too broad to afford much predictability of decision.

Direct regulation. *Zoning and planning.* In the 19th century urban areas expanded rapidly throughout the West. Industrialization introduced many new types of land uses, uses that frequently were annoying, dangerous, or injurious to the health of those engaged in more traditional residential, commercial, and agricultural activities. The invention and rapid spread of the automobile created problems of traffic control far exceeding anything that the age of the horse-and-buggy had produced. Fire and police protection in urban areas, the provision of such public services as trash collection, and the provision of water, gas, and electricity were rendered difficult, if not impossible, by the chaotic growth of many areas.

Throughout the West the response to these problems was to regulate development. By and large existing structures and land uses were allowed to remain, but new structures and new land uses were subjected to increasingly stringent regulation. The fact that only new structures and uses are subject to regulation is characteristic of all modern Western forms of land-use control, whether it is deemed constitutionally impermissible to require landowners to change existing uses or whether it is simply politically inexpedient to do so.

Building and zoning codes

In virtually all jurisdictions the key regulatory device is the requirement that new construction or substantial rehabilitation of old structures not be undertaken until official permission is obtained (building permit, *permit de construire, Baugenehmigung*). The landowner seeking the permit will present to the authorities a set of plans for the proposed construction. These plans are examined to determine if they meet two conceptually distinct sets of requirements. The first set of requirements is the building code. This code requires that all buildings or all buildings of a certain type (*e.g.*, multiple residences) conform to regulations concerning the types of materials used, fire safety, and the use of water and electricity within the building. Particularly for buildings designed for human habitation there are normally additional requirements concerning such matters as the amount of space per occupant, lighting, ventilation, plumbing, or electrical service. Some jurisdictions have a housing code in addition to the building code. The housing code frequently operates retroactively; *i.e.*, it sets out minimum requirements for any building in which human beings reside, whether or not it is newly constructed.

The second set of requirements is the zoning code, in a

more restricted sense. The zoning code lays out a series of requirements for construction and land use within particular areas (zones) of the jurisdiction. Zones may be either inclusive or exclusive. If the zones are inclusive, a hierarchy of land uses is created, usually ranging from the least to the most offensive uses. The typical pattern in urban areas begins with the establishment of residential uses and extends to commercial uses and finally to industrial uses. The characteristic of inclusive zoning is that in any given zone the use designated for that zone will be permitted and also any use conceived as being higher—*i.e.*, less offensive. In exclusive zoning, which is less common, only the designated use is permitted in the given zone.

Modern zoning is characterized by a multiplication of districts. Districts will be designated not only for particular types of uses but also for height and density control. The broad types of uses mentioned above may be elaborately subdivided. Residential districts, for example, may be further subdivided into residential districts for single-family detached houses, for single-family row houses, for two-family houses, for more than two-family houses, and for apartment buildings of given types. Height may be limited by stories or by measurement; open space may be required by setback limitations or by limiting the amount of the site that may be covered by the building; density may be controlled by limiting the ratio of floor area in the building to the area of the site.

Planning and process

The establishment of a comprehensive zoning code requires a considerable amount of planning. A full-scale plan, sometimes called a master plan, requires an accurate inventory of the population and of the land-use patterns existing in the area, economic and demographic predictions of what the future is likely to bring, a thorough understanding of the infrastructure that these future changes will require, and considerable imagination in determining what uses should be encouraged and what uses discouraged and where. Such an elaborate plan is normally not legally required before an initial zoning map is drawn. On the other hand, some planning is required, not only as a practical matter but also as a means of fulfilling the universal requirement that the zoning meet some minimum standard of rationality.

The typical statutory scheme vests a given local governmental body with the power to adopt a zoning plan for the region under its control. Public hearings and publication of the proposed plan for comment are frequently required. Property owners aggrieved by the plan are normally given an opportunity to obtain some kind of review of the decisions of the local governmental body. In the Anglo-American jurisdictions this review is normally had in the regular court system, although administrative review may have to be pursued before recourse to the courts is had. In the civil-law jurisdictions review is normally had in separate administrative courts. Both systems tend to give local governmental bodies considerable discretion in making their determinations.

Environmental and historical controls. Both environmental regulation and regulation designed to achieve preservation of historic buildings and districts have greatly increased in the 20th century, particularly since World War II. Broadly speaking, environmental regulations fall into two types: (1) those which limit, frequently by some scientific measure, the amount of a given toxin or pollutant that may be emitted into the air or into the water supply and (2) those that attempt to preserve natural areas in their natural form. Examples of the first type of regulation are the federal clean air and air quality acts in the United States and the Rhine Compact in Europe. Examples of the second type of regulation are the state wetlands preservation acts in the United States and the scenic designation districts in Europe.

Historical preservation regulation normally tightly controls changes in the exterior appearance of buildings. It may leave the building owner with the option of reconfiguring the interior of the building in such a way that he can continue to earn a return on the building, or it may require the preservation of the interior of the building as well. In some areas, if the building cannot be made profitable and still comply with the regulations, the government may be required to pay the building owner compensation or purchase the building from the owner.

Eminent domain. The concept of eminent domain dates back to at least the early 17th century. It states that the sovereign may take private property for public use, but only upon the payment of just compensation. Many instances of the use of the eminent domain power are universal throughout the West and uncontroversial. Governmental bodies everywhere take pieces of land from private owners in order to construct public roads, build government buildings, or install public services, such as electric wires, or water, gas, and sewer pipes.

In the 20th century there has been a considerable increase in the use of the eminent domain power. As noted above, either as a legal or as a political matter, land-use regulation normally operates only prospectively. For this reason major changes in the type of land use existing in a given area or in the quality and quantity of the buildings are most often accomplished by use of the eminent domain power. Urban renewal projects are a familiar example. Here, a governmental body condemns an entire area, frequently one containing a number of substandard buildings and inappropriately mixed land uses, and then razes the area. The governmental body may then either develop the area itself or sell the parcels to private developers on the condition that they develop them according to a plan devised by the governmental body.

Another use of the condemnation power occurs when a governmental body condemns the development rights in a given parcel of property. This may be done because there are doubts about the body's authority to proceed by way of regulation, because there are political objections to its doing so, or because the body wants to achieve greater flexibility. The current land use may continue so long as the owner wants to continue it, but no further development can take place without permission of the governmental body.

Just-compensation law

Whatever interest the governmental body takes, it is required to pay just compensation for it. Just compensation is normally defined as the fair market value of the land or interest taken. While there is considerable variation in just-compensation law and even more variation in what is actually awarded, the fair-market-value standard rarely gives the landowner full compensation for the economic loss that he suffers as a result of the taking. Just compensation rarely includes such items as loss of goodwill, moving costs, or counsel fees. Where the landowner retains land in the area and the value of that land is increased because of the public improvement, the increase in value is frequently deducted from the compensation the landowner receives. Thus, even in a situation where the government's obligation to pay compensation is conceded, the person whose land is condemned pays for the privilege of being a citizen of the community in which the land lies.

Limitations on government action. The notion that some losses by a private owner as a result of government action must be borne by him as part of the cost of living in a community is key to understanding how various jurisdictions determine when a governmental unit must proceed by using the eminent domain power and when it may proceed by way of regulation. Clearly, all government regulations affecting the use of land can have an adverse economic impact on the owner of the land, yet no Western legal system requires that all such economic losses be compensated. Some economic losses must be compensated; some need not be compensated. The question, increasingly debated as regulation of land use becomes more pervasive, is how to draw the line between those that must be compensated and those that need not be.

Every Western jurisdiction requires that, where the government takes property permanently for some public use, some compensation must be paid. But few, if any, Western jurisdictions require that compensation be paid when the government enacts a regulation concerning the prospective use of the land, even if the enactment of that regulation substantially decreases the market value of the land. Where and how the line will be drawn between these extremes varies considerably from jurisdiction to jurisdiction.

The United States probably has the most developed law

on this topic because the enforcement of the provisions of the U.S. Constitution that protect property interests from governmental interference has long been committed to the courts. In the United States two competing and overlapping theories are employed to distinguish "takings," which must be compensated, from "regulations," for which compensation need not be paid: (1) Where a governmental body invades the possessory interest of the landowner, compensation must be paid. There are exceptions to this principle, as, for example, in cases where the invasion of the possessory interest is for a short period and justifiable on grounds of protecting public health or safety or where it is unintentional; but the fact that a governmental body has invaded a landowner's possessory interest is a good predictor that a court will require that compensation be paid. (2) Where the government has not invaded a possessory interest of the landowner but has regulated his use of his property in such a way that no viable use of the property remains, compensation will frequently be required. This principle is considerably more controversial than the first, but it has been followed often enough that it, too, is a good predictor of judicial decisions.

The problem with the second principle is that it is dependent on the particular configuration of the property interests in question. Thus, a regulation that requires that mining operations be conducted in such a way as not to cause subsidence of the surface of the land would not deprive the owner of the entire tract of land of all use because he could continue to use the surface, nor would it deprive the owner of a deep mine of all possible use, since he could conduct his operations in such a way as to avoid the subsidence, but it would deprive the person who owned only a mine close to the surface of all use, since he could not mine without causing subsidence of the surface. (These are basically the facts of *Pennsylvania Coal Co.* v. *Mahon,* 260 U.S. 393 [1922].)

Compensation law in France and Germany

By and large the French legal system requires compensation only in those situations where the government has permanently deprived a landowner of the possession of his property. The concept of "regulatory taking" does not exist in French law. German law, however, because of the constitutional protection given to property since World War II, has developed a considerable jurisprudence on the topic. By and large the German developments have run a course parallel to those in the United States.

Acquisition and transfer of property interests

Conceptually the creation of a property interest de novo and its transfer from one person to another have little in common. The first topic concerns the initial allocation of resources and is closely connected with various theories about the origin of property. The second topic involves the more mundane world of everyday legal transactions. Practically, however, the two topics are closely related. Very few tangible things today do not have an owner. Thus, creation of an original title frequently depends on the extinction of another title, either of another private owner or of the state.

ORIGINAL ACQUISITION

First possession. Speculative jurisprudence, particularly in the 17th and 18th centuries, posited the creation of original titles in a remote period by the physical seizure by individuals of things from a common stock. The theory was based on the fact that most legal systems, including all Western ones, recognize the possibility that an original title can be created by the appropriation of certain kinds of physical things that have no private owner. Wild animals are a notable example. Similarly, tangible personal property lost or abandoned by its owner will frequently become the property of the finder, though regulation may require that in order for the finder to perfect his title he must turn the property over to a public authority for a period of time so as to allow the losing owner an opportunity to reclaim it.

The concept of discovery (meaning discovery by Europeans) played a considerable role in the establishment of European sovereignty over areas of the New World during the age of exploration (1450 to 1650). Survivals of ideas derived from this period may be seen in countries where the discovery of valuable mineral resources, particularly on public land, will give the discoverer some right to the minerals, although that right is frequently less than full ownership of them.

Patent, copyright, and trademark

The concept of discovery also plays a role in the creation of rights in intangible intellectual property. Patents are normally awarded to the person who can demonstrate that he has invented something that has previously not been known. Copyrights are given to the author, the creator of a literary, musical, or artistic work. Trademarks are normally dependent on being the first to use or to register the mark.

Accession. The concepts of accession, specification, and confusion all involve situations in which something new is made out of something old and that which is old belongs to two different people. In most legal systems, the owner of a cow will own the calf without regard to who owns the bull, and the owner of a field will own the plants that grow in it without regard to who owned the seed or who planted it (accession). In many legal systems, if person A makes a silver plate out of person B's silver ingot, the plate will belong to person A (specification), but if person A decorates the plate of person B, even if that greatly increases its value, the plate will belong to person B, although person B may have to compensate person A for his labour. Finally, if one person's wheat becomes commingled with that of another in the same bin in such a way that the two cannot be separated, the ownership of the wheat will probably depend on who had the larger portion to start with, with the owner of what was originally the smaller portion being relegated to a right to compensation (confusion).

Adverse possession, prescription, and expropriation. In the section on ownership and possession (see above), the related concepts of adverse possession and prescription are discussed. A number of possible rules are buried in the two concepts. One might say, for example, that the expiration of the statute of limitations simply bars the action, but it does not bar the right (limitation of actions, strictly speaking). Alternatively, one might say that the passage of the statutory period bars both the action and the right but does not create any new right in the adverse possessor (extinctive prescription). Or one might say that the adverse possessor, or the one who has fulfilled the requirements for prescription, acquires the title of the one whose title is time-barred (acquisitive prescription, strictly speaking). Both Anglo-American and civil law generally take the more extreme position that, once the rights of the original owner have been extinguished, the person who has prescribed or adversely possessed against those rights has a new original title. At a minimum this means that the new owner may prove his title without having to show how the previous owner acquired his title. It may also mean that he is not subject to restrictions that the original owner may have agreed to.

The exercise of the power of eminent domain also normally results in a new title in the sovereign.

Privileges conferred by public authorities. The foregoing discussion makes it clear that the state has the power throughout the West to create new forms of property. The state may be limited by convention or the constitution when it comes to reallocating property already in private hands, but the state is free to transfer to individuals property that it has and also to create new types of rights that will count as property rights.

DERIVATIVE ACQUISITION

Granted that a property right, privilege, or power exists in a private person, it may be asked whether that right, power, or privilege can be transferred to someone else. The general assumption in Western law is that it can be. Freedom of contract and freedom of alienation of property are the twin foundations of a market economy, and despite the extensive regulation and socialization of the market economies of the West, the basic principle has remained unimpaired. Freedom of alienation is less characteristic of socialist economies and legal systems and of non-Western economies and legal systems. Nonetheless, even these sys-

tems allow alienation in a wide variety of circumstances.

The promise to alienate property

Contract and conveyance. Any legal system that distinguishes between property and obligation (as do all Western systems) will distinguish between a promise to alienate property and the alienation itself. The promise may be fully enforceable between the parties; it may even affect the rights of third parties, at least those who know of the promise. But until the property is transferred, the original owner has a real right in the property (good, notionally, against the whole world), and the promisee has simply an enforceable obligation to have the property transferred.

In many transactions the contract and conveyance take place simultaneously so that the distinction between the two makes no practical difference. If person A buys a watch at a jeweler's, pays for it, and walks out of the store with it on his wrist, both a contract of sale and a conveyance of the watch have taken place; there is no need to distinguish between the two. If, however, person A does not pay for the watch but wears it out of the store and then transfers it to some third person, it becomes important to know whether the jeweler still owned the watch when it was transferred (in which case the jeweler may recover it from the third person) or whether person A owned the watch (in which case the third person now owns it, and the jeweler's sole remedy is against person A). Similarly, if person A pays for the watch but leaves it with the jeweler to have a strap put on it, and the jeweler transfers it to some third party before person A comes back to pick it up, it becomes important to know whether the jeweler still owned the watch (in which case the third party now owns it and person A's sole remedy is against the jeweler) or whether person A owned the watch as soon as he paid for it (in which case he may recover the watch from the third party, and his remedy, if any, will be against the jeweler).

In the example given above there are three possible points at which the title to the property could pass: (1) when the contract between the jeweler and person A was formed (normally when they have agreed on a price and a thing to be sold), (2) when person A paid for the watch, or (3) when the jeweler handed over the watch to him. As a general matter, Western law takes the first or the third position and leaves the second possibility to private agreement between the parties. Thus, in the absence of agreement to the contrary, Western law generally provides that transfer of title takes place either when a valid agreement to transfer is made or when the thing is delivered to the conveyee.

Registration and recordation. In the example of the watch, the distinction between contract and conveyance became important as soon as the rights of a third person became involved. But from the point of view of the third party any one of the three suggested rules about conveyance might be unsatisfactory, because it may be difficult for the third party to know whether a contract has been formed, a payment under it has been made, or even whether the property has been delivered to the purchaser as owner, as opposed to as borrower or hirer. To protect third parties in these situations, many legal systems provide for the registration or recordation of transactions, particularly transactions involving items of great value (such as airplanes or boats or cars) or items of great durability (such as land).

Types of registration systems

Registration systems fall into two general types. The first type provides for the registration of title. Under this system transfer of title does not take place unless and until the transfer is registered in the system. This is the system of the German *Grundbuch,* in which titles to land are registered, and of the systems for registration of automobile titles that prevail in the United States. The other type of system is a recording system. Under such a system a transfer is effective even if it is not recorded, but a good-faith purchaser who relies on the record is not protected unless the transaction is recorded. Under this system the previous owner who the record shows is still the owner has the power to convey good title to an innocent third party unless and until the new owner records the transaction. This is the system that prevails in most jurisdictions in the United States for land and under the French system of registration for transfers of land. The English land registration system is more like the German system than it is like the French or the American.

Sales. In Anglo-American law three things must be established about a conveyance before the law applicable to it can be determined: (1) whether it is a sale or a gift, (2) whether it is of personal (movable) or real (immovable) property, and (3) whether it is immediately effective between living parties (inter vivos) or will take effect only upon the death of the conveyor (testamentary). Whereas inter vivos sales and inter vivos gifts of movables are treated quite differently, the conveyancing aspects of inter vivos sales and inter vivos gifts of immovables are quite similar. Testamentary sales of either movables or immovables are rare, and testamentary gifts of movables and of immovables are treated similarly. In civil law the distinction between conveyances of movables and conveyances of immovables is far less important than it is in Anglo-American law, whereas the distinction between sales and gifts of immovables is more important than it is in Anglo-American law.

Movables. In both Anglo-American and civil law the sale of a movable is both a contract and a conveyance. In both Anglo-American and French law the contract also serves to transfer the title to the thing unless the parties agree otherwise. German law, on the other hand, following Roman law, requires that there be a handing over of the thing from the seller to the buyer before title may pass. Indeed, in German law title to the thing will pass even if there is no valid contract of sale, so long as the parties intend to transfer ownership of the thing.

The difference between the Anglo-American and French systems, on the one hand, and the German, on the other, can be exaggerated. The number of situations in which there is intent to transfer ownership of a thing in German law without there being a valid contract of sale (or gift, see below) is small. Further, German law allows the transferor and transferee to agree that the transferor will remain in physical possession of the goods, even though title has passed to the purchaser. Thus, in the example given above where the watch remained with the jeweler to have the strap put on it, all three systems would probably hold that title had passed to the purchaser, but the German system would require evidence that the purchaser and the seller had agreed that the seller retain possession in fact on behalf of the new owner, the purchaser.

Despite the fact that all three systems probably would hold that the purchaser had good title to the watch even though the seller retained physical possession of it, all three systems, somewhat surprisingly, would probably protect the third party to whom the jeweler transferred it. All three systems hold as a basic principle that one cannot transfer more rights in a thing than one has (*nemo dat quod non habet; nemo plus iuris ad alium transferre potest quam ipse habet*), but all three systems recognize numerous exceptions to this principle, particularly in the case of movables. Both the French and German systems recognize that the actual possessor of movable goods (with the notable exception of stolen goods) may give good title, at least to a good-faith purchaser. The Anglo-American system is narrower in this regard, but, at least in the United States, someone who entrusts his goods to a merchant, such as the jeweler in this case, who regularly deals in such goods, is liable to lose his title to the person to whom the merchant sells the goods.

Sale of immovables. Sale of real property in Anglo-American law is radically different from the sale of goods.

The Statute of Frauds

The Statute of Frauds of 1677, which in one form or another is in effect in all Anglo-American jurisdictions, requires that the transfer of most types of interests in land be made by a writing (deed). Contracts for the sale of land also have to be evidenced by a writing, but unless the contract and the transfer are evidenced by the same piece of writing (something that in practice is very rare), the contract will not suffice to transfer the title to real property.

In practice, the sale of real property is always preceded by a contract. The contract will fix the price and other terms of the arrangement and will normally fix a date (the "law day") on which the seller is to appear with a deed to

the property conveying "good and merchantable title" and the buyer is to appear with the purchase price. A contract for the sale of land is specifically enforceable. If either side fails to perform, the other party, if ready, willing, and able to perform, may compel the performance. But the ability to compel the performance is not the same thing as having legal (as opposed to equitable) title to the property. That only happens when the conveyance is made—*i.e.*, when the seller delivers the deed to the buyer. The period intervening between the contract and the conveyance is normally occupied by the buyer's obtaining financing for the purchase and the seller's obtaining evidence, based on the public record or on his own muniments of title, that he has merchantable title to the property.

In French law a contract of sale of an immovable passes title to the immovable. Subsequent registration serves to protect that title against third-party purchasers in good faith from the original vendor. In German law the contract of sale and the transfer are conceptually distinct, but in practice they are frequently merged in the same transaction. The transfer of title is not valid as to third parties, or even between the parties themselves, until the transaction is registered in the *Grundbuch*.

In both the French and German systems the time between the contract and its ultimate consummation is markedly shorter than it is in the Anglo-American system. This may be explained in part by the fact that the public recording and registration systems are more effective (despite the differences in how they operate) and by the fact that in both systems there are fewer possible outstanding interests in land. Another explanation of the differences between Anglo-American and civil-law conveyancing practices would look to the differences in the ways that real estate transactions are financed.

Donative transactions

Gifts. In Anglo-American law a promise to make a gift is not a binding contract, because it lacks the essential element of consideration (the requirement that to be valid a contract must involve a bargained-for exchange). By contrast, in civil law a contract to make a gift is valid if it is accompanied by certain formalities and if it does not violate the expectancies that the close relatives of the donor have in the property. It is not surprising, then, that donative transactions operate in civil law in much the same way as do sale transactions.

Inter vivos. Lacking the contract to make the gift valid, Anglo-American law has long puzzled over the donative conveyance of movables. Traditional doctrine holds that there has to be delivery, a transfer of possession of the thing accompanied by donative intent on the part of the donor, and acceptance by the donee. Acceptance will be presumed, but evidence of both delivery and donative intent has long been thought to be essential. The contortions that this doctrine produces, particularly in situations where the donative intent is clear but the thing in question is awkward or impossible to deliver, have long been noted by courts and commentators alike. Recently, Anglo-American courts seem to be increasingly willing to allow the delivery of a writing embodying a statement of the gift to substitute for the delivery of the thing itself.

Gifts of real property have caused less difficulty in Anglo-American jurisdictions. It is well established that a writing (deed) is necessary for the transfer of title to real estate; it is common for deeds to recite at least nominal consideration, but no preliminary contract is required for title to pass. Recording of the deed is necessary to make it binding as to subsequent good-faith purchasers from (but not donees of) the same donor.

In civil law a promise of a gift is binding if it is notarized and if it does not deprive the donor's expectant heirs of their obligatory share in his estate. In French law the contract alone suffices to transfer the property. In German law, as in the case of sale, there must be transfer of possession or an agreement that the donor retain possession on behalf of the donee if the thing is movable or an entry in the *Grundbuch* if the thing is immovable. Thus, in civil law inter vivos transfers by way of gift parallel those by way of sale, with the important exception that gifts of either movables or immovables may be subject to the overriding interests of the donor's expectant heirs.

Testaments. Western law generally permits a property owner not only to transfer his property while he is alive but also to transfer the property that he owns at his death. This is done by a document called a will or testament. A will is revocable at any time before the testator's death, but if he dies without having changed it, it comes into effect. Thus, the principal characteristic of the will in Western legal systems is its ambulatory nature. It confers no rights on the beneficiaries at the time it is executed but only at the time of the testator's death, and it transfers not the property that the testator owns when he makes the will but rather what he owns at the time of his death.

On this much both the Anglo-American and civil-law systems are in agreement. Beyond this they differ substantially, largely for historical reasons. Decedents' estates are administered quite differently in the two systems, and there are substantial differences in the amount of freedom of disposition that each system gives the deceased. These differences are considered in the next section.

The form of a will

While the form required for a valid will varies from jurisdiction to jurisdiction, a few common principles are observable: in most civil-law jurisdictions and in some Anglo-American jurisdictions a document entirely in the writing of the testator (holograph), signed and dated by the testator, will constitute a valid will. In France and Germany such wills are quite common, perhaps even the norm, and they are normally executed after seeking advice from a notary. In those Anglo-American jurisdictions in which they are valid, their use is far less common than in civil-law countries, and they are almost never recommended by professionals.

Both Anglo-American and civil-law jurisdictions also make use of a formal will, derived from the Roman testament. The characteristic of such a will is that it must be witnessed by a certain number (generally two or three in modern law) of disinterested witnesses. It is normally prepared by a professional, a notary on the Continent or a solicitor or other lawyer in the Anglo-American jurisdictions, and it tends to formality of language.

Many Western jurisdictions will excuse some of the formalities required for will making in certain circumstances. Soldiers' and sailors' wills, for example, are frequently effective with fewer than the usual formalities, and oral wills (nuncupative wills) at least of certain types of property may be valid if made under certain circumstances, such as when the testator is dying. The nuncupative will is related to, though conceptually distinct from, the *causa mortis* gift, a device that exists in most Anglo-American and in some civil-law jurisdictions.

Succession. After a long historical development Western legal systems have come to regard succession to property by relatives of the deceased as normal. Succession by the wider community is regarded as abnormal, something that happens only when no near kin can be found and when the deceased has made no other disposition of his property. Taxation of the passage of property at death, though nearly universal in the West, is not regarded as a form of succession but as a public-law element engrafted onto the system.

Executors and universal succession. As noted above, civil-law systems of succession differ markedly from those in Anglo-American countries. In Anglo-American countries the death of a property holder initiates a period in which his assets are in transition. The deceased's will, if he made one, must be probated (proven) in a court. The court appoints a personal representative of the deceased. In many jurisdictions, if the personal representative has been named in the will, he is called the executor of the will. If no personal representative is named in the will, if the personal representative named in the will fails to qualify or declines to serve, or if there is no will, the personal representative appointed by the court is called the administrator of the estate. The personal representative then marshals the deceased's assets, pays the deceased's debts and the taxes on his estate, and distributes the balance to those named in the will or those entitled to the assets by law.

Title to the deceased's personal property passes to the personal representative, subject to his fiduciary responsibility

to discharge the obligations of the estate and distribute the balance to those entitled to it. At common law and in a number of Anglo-American jurisdictions today, title to real property passes immediately upon death to those who succeed to it, but since the real property is also liable to the payment of the deceased's debts and taxes, the title is normally not marketable until the estate is closed. Thus, as a practical matter, the deceased's estate constitutes an entity held by the personal representative until the distribution is made.

Civil-law systems generally avoid the hiatus period characteristic of Anglo-American estate administration. By and large civil-law systems provide for universal succession, a term derived from Roman law, which means that the heirs succeed immediately to both the assets and the liabilities of the deceased. They are responsible for paying the debts and taxes and any legacies that may lawfully be charged against the estate. Personal representatives are not unknown in civil law, but they generally play a less important role than do those in the Anglo-American systems.

Universal succession

Freedom of testation and legitim. Another major difference between Anglo-American and civil-law systems of succession is that wills, though important, are less important in civil law than they are in the Anglo-American system. In civil law someone who dies leaving a spouse or close kin (descendants or ascendants) may effectively dispose of only a portion of his estate by will. The rest must go to the statutory heirs (*réserve héréditaire, legitim*). Wills remain important in the civil-law systems, however, both because the disposable share of the estate may amount to a large monetary sum and because the statutory share of the heirs tends to be viewed in monetary terms. Thus, the will may direct that certain assets be given to certain members of the family, so long as each member receives the value to which he or she is entitled under the statute.

Anglo-American law affords, at least in theory, greater freedom of testation. In England a deceased may dispose of his entire estate by will to the detriment of his spouse and children, subject however to contravention by a court upon petition of the spouse or children if they are not adequately provided for. In the United States a deceased may generally not disinherit his spouse but may disinherit his children, even if this leaves them without any means of support.

Theoretical possibilities, however, do not determine practical realities. Many Americans, for example, avoid the probate system entirely, either because they make lifetime dispositions of their property (for example, in trust) or because their heirs behave as if universal succession were in fact in place; *i.e.*, the heirs divide the property among themselves and pay the creditors and the tax collector out of their own pockets. Similarly, there seems to be little pressure to change the amount of freedom of testation offered to many Anglo-American testators because that freedom is rarely used to disinherit spouses or children. (Perhaps the most common form of American will is one that gives the surviving spouse everything, usually with the tacit understanding that he or she will give anything left over to the children on his or her death.)

Patterns of intestate succession. Patterns of intestate succession vary markedly from jurisdiction to jurisdiction in the West, although the differences tend to be ones of detail and not of principle. The typical Anglo-American intestacy statute gives the surviving spouse a half or a third of the property, with the remaining half or two-thirds going to the children of the deceased, the children of any deceased child dividing their parent's share among them (representation). In the absence of a surviving spouse, the children (or their representatives) take all. In the absence of children, the surviving spouse takes all or shares his or her portion with the deceased's parents. Beyond that the patterns vary, but almost all provide for succession by the deceased's next of kin, at least so long as he left grandparents or descendants of grandparents. If no one survives in these categories, some modern systems give the property to the state; others continue the search for blood relatives of the deceased.

Civil-law patterns do not vary greatly, though they tend to give less to the surviving spouse because he or she is presumed to have a share of the community property (see above). The French system is notable for the fact that it divides the deceased's property between his maternal and paternal kin, if there are no descendants. The German system is more like the Anglo-American.

Historically in the West illegitimate children were totally excluded from inheritance. Modern Western legal systems have come increasingly to recognize inheritance rights of illegitimates, although not all systems give them equal rights with legitimates.

Inheritance rights of illegitimate children

English law did not recognize adoption until 1926. Modern Anglo-American law has come to recognize adopted children as, in most jurisdictions and for the most part, equal in inheritance rights to natural children. The civil law has had less difficulty recognizing the rights of adopted children because Roman law freely allowed adoption.

Nonconsensual transfers. Western law generally recognizes some situations in which transfers of property are made without the consent of the property holder. By far the most important of these are the situations in which a property holder may be forced to give up his property to someone to whom he owes money. As a general matter, in Western legal systems a creditor may reach any property which the debtor has the power voluntarily to transfer. Thus, even in the absence of a mortgage or a pledge, a person's general creditors may reach and execute upon any property that he could convey voluntarily to them. The property that the general creditor can reach varies considerably from jurisdiction to jurisdiction. Many jurisdictions exempt certain kinds of property from execution by creditors. The "debtor states" of the United States are particularly notable for the amount and character of the property that may be so exempt. The jurisdictions also differ widely in the extent to which they require court supervision of the process of seizing a debtor's property. Some will allow it upon very little showing; others require a quite elaborate process.

As mentioned above, co-owners in many jurisdictions are able to force the judicial sale of the property that they hold, particularly if partition in kind is not possible.

Although acquisitive prescription or expropriation (see above) can be viewed as involuntary transfers, they are technically probably better viewed as the creation of new titles by the state and the extinction of old ones. There are, however, some situations in which this analysis will not suffice. In the Anglo-American system a court of equity has considerable discretion to deny an injunction and remit the complainant to money damages. Where this happens, a property right is in effect forcibly sold by the property holder, who must be satisfied with money damages. The same problem does not seem to have arisen in civil-law systems, but a similar result can occur in the civil law when a landowner claims a way of necessity.

The Western concept of property: assessment

One of the striking results of this survey of property law in the West has been to recognize the extent to which the differences between the civil-law systems and the Anglo-American systems are not of great practical significance. Despite substantial differences in the history of the two systems—differences that are still manifest today in the different vocabulary and different devices that the two systems use to solve legal problems—it was frequently possible to say either that the two systems arrive at the same practical result or that the practical results in the two systems are converging.

The tendency toward convergence

This fact should not be surprising. As said above, property law is intimately connected with the economy of the society, with its family structure, and with its political system. Commentators have long noted that the industrialized, bureaucratic democracies of the West have become increasingly interdependent and increasingly like each other as the 20th century has progressed. To the extent that the differences between the Anglo-American and civil-law systems are based on historical differences that are not reflected in the modern societies, one should expect that these societies would tend toward the same legal solutions despite their differences in conceptual starting points.

This tendency toward convergence has been aided by the fact that most of the changes noted above have come by way of legislation (*e.g.*, landlord and tenant). Legislation in the West has traditionally been able to change the conceptual scheme that history has imposed on the law. Most of these legislative changes have also come about in the area denominated public law (*e.g.*, land-use regulation). Public law operates largely free of the historical differences in conceptualization that characterize the private law of the Anglo-American and civil-law systems. As public law has become increasingly important for property law in the definitional sense, the practical results in the two systems have been able to come closer together.

The last decades of the 20th century have also seen a progressive eroding of both the legal concept and the practical importance of private property in the West. Attacks on the concept of private property itself are more common in the United States than in other countries, both because of the American constitutional tradition of protecting property and also because the United States has made more changes in its law by way of judicial decision. American judges are forced, in a way that legislators are not, to make their decisions in the context of traditional property law.

In the United States, then, examples of direct attacks on the concept of property are easy to find. The "sit-in" cases and the cases involving access to migrant labour camps (referred to above) were expressly framed as conflicts between "property rights" and "civil rights" and left a strong suggestion that a humane society could hardly prefer the former to the latter. In a quite different area of law, many judges and commentators expressly characterized the judicial changes in the law of landlord and tenant as a triumph of contract over property concepts. A number of academic writers have suggested that property rules, with their characteristic injunctive remedy, interfere with the achievement of allocational efficiency and should be replaced by liability rules; damages, they argue, should be awarded for the violation of property rights rather than specific performance.

Property rules versus liability rules

The only recent academic development that suggests anything other than a dim future for the concept of private property in the United States was the attempt to apply the concept of property to a citizen's expectation that he would continue to be the recipient of government benefits. As noted above, the idea found favour with the courts for a while but then seemed to be lost from view.

Most of these developments are evident throughout the Western world, although their relationship to the concept of private property has not always been put as sharply as in the United States. Thus, the problem of access to restaurants or places of employment is in many Western countries subject to regulation similar to that which exists in the United States, but the regulations are perceived as "public law" regulations of business and hence outside the scope of the "private law" of property. Landlord and tenant law has been reformed again and again in the West in the 20th century. The question of the appropriate remedy to defend property rights has appeared wherever thinking about law in the manner of a new subspeciality, that of law and economics, has been felt. The security of job, pension, insurance, and welfare are standard political and legal issues throughout the Western world.

Other developments suggest that the institution, if not the concept, of private property is in trouble throughout the West. A new body of regulatory law designed to protect the environment has greatly restricted the traditional privileges of owners of resources to use them to their advantage. Further, and perhaps in the long run most significantly, family ties are weakening throughout the West. Divorce rates are at an all-time high. Birth rates are declining, particularly among the middle class. Finally, continuing high discount rates, whether accompanied by inflation or not, and heavy taxation on private accumulations of wealth and particularly on its passage from generation to generation have affected the attitudes and practices of many Westerners with regard to the use of traditional forms of property as a source of personal financial security. While the concept of private property itself is not directly involved in these developments, the traditional devices that property law offers for ensuring the financial security of present and future generations are made increasingly irrelevant by these developments.

Some of these modern developments directly involve the agglomerative tendency characteristic of the Western concept of property. One may, for example, consider the developments in landlord-tenant law. The landlord is the owner; all that the lease purports to do, in the Anglo-American system, is to convey possession to the tenant for a term, and, in the civil-law countries, to create in the landlord an obligation to allow the tenant the use of the premises for a particular period. The agglomerative tendency would dictate that the tenant should be entitled to no more than was conveyed to him or to no more than bargained for. To the extent that the modern developments give the tenant more, they undercut the tendency. The agglomerative tendency is also involved in the conflict between property rights and civil rights because the tendency would lead one to define the right to exclude broadly and to require that it be protected even where it is being used discriminatorily. To the extent that modern law has reached a different result, it has, once again, undercut the tendency.

The agglomerative tendency undercut

Some of these modern developments involve tensions inherent in the agglomerative tendency. Perhaps the only reason that increasing environmental regulation is not always perceived as simply a variant on the old land-use problem is that adverse environmental effects are sufficiently diffused that it is difficult to see that it is the property of those who feel the effect that is at stake. The academic controversy over property rules versus liability rules involves the same tension. The economic approach suggests that, once the law has chosen one land use in preference to another, it should compromise with the nonfavoured land use by giving the favoured use less than full remedies.

Neither the agglomerative tendency nor the tensions within it are involved in the changes in the family or in the controversy over the security of employment and government benefits, but the two may be related and their product affects the role, at least of traditional property, in the society. Traditional property operates in a family context and serves to support the role of the family in providing security for its members. If family ties are weakening and if traditional property cannot provide security against exogenous economic forces, people will look elsewhere for their security. They will look, for example, to their employment or to government benefits. As these, in turn, become more the focus of attention and more secure, traditional property and the security that it provides within the family will be further weakened. It is perhaps for this reason that the different compromises the Anglo-American and civil-law systems made historically between the power to convey of present holders and the power to convey of the conveyee have remained more stable in the West than have those about land use. They have remained stable because private settlements are increasingly irrelevant to the financial security of most of the population.

Thus, although nothing concerning the descriptive definition of property used in this article is at stake in these modern controversies and developments, the concept and the institution of property has changed fundamentally in the West in the last generations. Coming as these developments do, at a time of the decay of liberalism, one may fairly ask whether these developments indicate a change not only from the extremes to which the agglomerative tendency was brought during the 19th and early 20th centuries but also whether they indicate a reversal of the fundamental tendency itself. That is an issue that only the next century can resolve.

BIBLIOGRAPHY

General sources: An accessible, comprehensive treatment of the subject with a range as broad as that of this article is to be found in separate chapters of FREDERICK H. LAWSON (ed.), *Property and Trust*, published in fascicles as vol. 6 of a major undertaking of the International Association of Legal Science, *International Encyclopedia of Comparative Law* (1971–). The civil-law systems (with particular focus on Louisiana) are

treated comparatively in A.N. YIANNOPOULOS, *Property: The Law of Things, Real Rights, Real Actions,* 2nd ed. (1980), kept up-to-date by supplements, *Personal Servitudes: Usufruct, Habitation, Rights of Use,* 3rd ed. (1989), and *Predial Servitudes* (1983).

History: The origins of the Western idea of property are examined in BARRY NICHOLAS, *An Introduction to Roman Law* (1962, reprinted 1987); W.W. BUCKLAND, *A Text-Book of Roman Law from Augustus to Justinian,* 3rd ed., rev. by PETER STEIN (1963, reprinted 1975); MAX KASER, *Das römishe Privatrecht,* 2nd rev. ed. (1971–75); FREDERICK POLLOCK and FREDERICK WILLIAM MAITLAND, *The History of English Law Before the Time of Edward I,* 2nd ed., 2 vol. (1898, reissued with a new introduction by S.F.C. MILSOM, 1968); A.W.B. SIMPSON, *A History of the Land Law,* 2nd ed. (1986), also on English law; and on that of the Continent, HELMUT COING, *Europäisches Privatrecht,* 2 vol. (1985–89); and PAOLO GROSSI, *An Alternative to Private Property: Collective Property in the Juridical Consciousness of the Nineteenth Century* (1981; originally published in Italian, 1977). For the United States, see LAWRENCE M. FRIEDMAN, *A History of American Law,* 2nd ed. (1985).

Theory: Surveys of classical and of early modern theories of property law are found in J. ROLAND PENNOCK and JOHN W. CHAPMAN (eds.), *Property* (1980), which includes a look at the future of the Western concept of property in CHARLES DONAHUE, JR., "The Future of the Concept of Property Predicted from Its Past," pp. 28–68; RICHARD TUCK, *Natural Rights Theories: Their Origin and Development* (1979); and RICHARD SCHLATTER, *Private Property: The History of an Idea* (1951, reprinted 1973).

Specific modern legal systems: For American developments, see ROGER A. CUNNINGHAM, WILLIAM B. STOEBUCK, and DALE A. WHITMAN, *The Law of Property* (1984); RAY ANDREWS BROWN, *The Law of Personal Property,* 3rd ed., rev. by WALTER B. RAUSHENBUSH (1975); RICHARD R. POWELL, *The Law of Real Property,* rev. by PATRICK J. ROHAN (1949–), a multivolume classic treatise, still being updated with supplements; and A. JAMES CASNER (ed.), *American Law of Property,* 7 vol. in 8 (1952–54). For England, see FREDERICK H. LAWSON and BERNARD RUDDEN, *The Law of Property,* 2nd ed. (1982); KEVIN GRAY, *Elements of Land Law* (1987); ROBERT MEGARRY and H.W.R. WADE, *The Law of Real Property,* 5th ed. (1984); and J. CROSSLEY VAINES, *Crossley Vaines' Personal Property,* 5th ed., rev. by E.L.G. TYLER and N.E. PALMER (1973). French law is reviewed in MAURICE S. AMOS, *Amos & Walton's Introduction to French Law,* 3rd ed., rev. by FREDERICK H. LAWSON, A.E. ANTON, and L. NEVILLE BROWN (1967); CHRISTIAN LARROUMET, *Droit civil,* vol. 2: *Les Biens, droits réels principaux* (1985); GABRIEL MARTY and PIERRE RAYNAUD, *Les Biens,* 2nd ed. (1980), and *Les Régimes matrimoniaux,* 2nd ed. (1985); and MARCEL PLANIOL and GEORGES RIPERT, *Traité pratique de droit civil français,* 2nd ed., 14 vol. (1952–62).

For Germany, see NORBERT HORN, HEIN KÖTZ, and HANS G. LESER, *German Private and Commercial Law,* trans. from German (1982); E.J. COHN, *Manual of German Law,* 2nd rev. ed., 2 vol. (1968); ERNST WOLF, *Lehrbuch des Sachenrechts,* 2nd ed. (1979); LUDWIG ENNECCERUS, THEODORE KIPP, and MARTIN WOLFF (eds.), *Lehrbuch des bürgerlichen Rechts,* vol. 1 in 2: *Allgemeiner Teil des bürgerlichen Rechts,* 15th ed., rev. by H.C. NIPPERDEY (1959–60), vol. 3: *Sachenrecht,* 10th ed., rev. by MARTIN WOLFF and L. RAISER (1957), and vol. 5: *Erbrecht,* 13th ed., rev. by HELMUT COING (1978); and KURT REBMANN and FRANZ-JÜRGEN SÄCKER (eds.), *Münchener Kommentar zum Bürgerlichen Gesetzbuch,* 2nd ed. (1984–), planned for 7 vol., some multipart, of which 5 had been published by 1989.

Specific studies: Following are works discussing other topics, listed in the order they are treated in the text. The vocabulary of jural relationships is presented in WESLEY N. HOHFELD, *Fundamental Legal Conceptions as Applied in Judicial Reasoning: And Other Legal Essays,* ed. by WALTER WHEELER COOK (1923), available also in later abridged editions. For non-Western systems of property, see chapter 2, "Structural Variations in Property Law," published as a separate fascicle of the incomplete vol. 6 of the *International Encyclopedia of Comparative Law,* covering Islāmic, Hindu, and African law, as well as that of socialist countries; and MAX GLUCKMAN, *The Ideas in Barotse Jurisprudence* (1965, reprinted with a new preface and amendments, 1972).

Evolutionary anthropology and the development of the concept of property are addressed in FRIEDRICH ENGELS, *The Origin of the Family, Private Property, and the State* (1902, originally published in German, 1884), also available in many later editions and translations; HENRY SUMNER MAINE, *Ancient Law: Its Connection with the Early History of Society, and Its Relation to Modern Ideas* (1861), available in many later editions; PETER STEIN, *Legal Evolution: The Story of an Idea* (1980); and ALAN WATSON, *The Evolution of Law* (1985).

S.F.C. MILSOM, *The Legal Framework of English Feudalism* (1976, reprinted 1986), examines the tradition of land tenure; D.R. COQUILLETTE, "Mosses from an Old Manse: Another Look at Some Historic Property Cases About the Environment," *Cornell Law Review* 64:761–821 (June 1979), discusses the nuisance law; WILLIAM B. STOEBUCK, "A General Theory of Eminent Domain," *Washington Law Review* 47:553–608 (August 1972), traces the history of the governmental authority over private property; and DAVID J. SEIPP, "Bracton, the Year Books and the 'Transformation of Elementary Legal Ideas' in the Early Common Law," *Law and History Review* 7:175–217 (Spring 1989), analyzes the concept of property in Bracton. On "possessive individualism," see C.B. MACPHERSON, *The Political Theory of Possessive Individualism: Hobbes to Locke* (1962, reprinted 1985). GEORGE L. HASKINS, "Extending the Grasp of the Dead Hand: Reflections on the Origins of the Rule Against Perpetuities," *University of Pennsylvania Law Review* 126:19–46 (November 1977), explores social conflicts in connection with this rule. See also J.H.C. MORRIS and W. BARTON LEACH, *The Rule Against Perpetuities,* 2nd ed. (1962, reprinted 1986); and RONALD H. MAUDSLEY, *The Modern Law of Perpetuities* (1979). C. REICH, "The New Property," *Yale Law Journal* 73(5):733–787 (April 1964), looks at government-granted rights as "property." Property and personal financial security are the topic of MARY ANN GLENDON, *The New Family and the New Property* (1981), and *The Transformation of Family Law: State, Law, and Family in the United States and Western Europe* (1989), focusing on marital property. For corporate property, see ADOLF A. BERLE and GARDINER C. MEANS, *The Modern Corporation and Private Property,* rev. ed. (1968).

Definitions of ownership are given in FELIX S. COHEN, "Dialogue on Private Property," *Rutgers Law Review* 9(2):357–387 (Winter 1954); and A.M. HONORÉ, "Ownership," ch. 5, pp. 107–147 in A.G. GUEST (ed.), *Oxford Essays in Jurisprudence* (1961). Modern legal relations between landlord and tenant are examined in CHARLES DONAHUE, JR., "Change in the American Law of Landlord and Tenant," *Modern Law Review* 37:242–263 (May 1974). For trusts, see AUSTIN WAKEMAN SCOTT, *The Law of Trusts,* 3rd ed., 6 vol. (1967), with a 4th ed., by AUSTIN WAKEMAN SCOTT and WILLIAM FRANKLIN FRATCHER, appearing in parts since 1987. Civil-law functional equivalents of the trust are discussed in CHRISTIAN DE WULF, *The Trust and Corresponding Institutions in the Civil Law* (1965). FRANCIS ALLEN, "Offenses Against Property," pp. 57–76 in LOUIS B. SCHWARTZ (ed.), *Crime and the American Penal System* (1962), studies the protection of property in criminal law. DONALD G. HAGMAN and JULIAN CONRAD JUERGENSMEYER, *Urban Planning and Land Development Control Law,* 2nd ed. (1986); and J.F. GARNER and N.P. GRAVELLS (eds.), *Planning Law in Western Europe,* 2nd rev. ed. (1986), deal with public control of land use. Constitutional protection of property is the subject of BRUCE A. ACKERMAN, *Private Property and the Constitution* (1977). A comparative treatment of gifts (both inter vivos and testamentary) is offered in JOHN P. DAWSON, *Gifts and Promises: Continental and American Law Compared* (1980). American wills are examined in M.L. FELLOWS *et al.,* "An Empirical Study of the Illinois Statutory Estate Plan," *University of Illinois Law Forum* 1976:714–745 (1976). Property from the perspective of law and economics is presented in G. CALABRESI and A.D. MELAMED, "Property Rules, Liability Rules, and Inalienability: One View of the Cathedral," *Harvard Law Review* 85:1089–1128 (April 1972); and A.M. POLINSKY, "Controlling Externalities and Protecting Entitlements: Property Right, Liability Rule, and Tax-Subsidy Approaches," *Journal of Legal Studies* 8(1):1–48 (January 1979).

(Ch.D.)

Protestantism

Protestantism, beginning in northern Europe in the early 16th century in reaction to medieval Roman Catholic doctrines and practices, became, along with Roman Catholicism and Eastern Orthodoxy, one of three major forces in Christianity. After a series of European religious wars, and especially in the 19th century, it spread rapidly in various forms throughout the world. Wherever Protestantism gained a foothold, it influenced, to a greater or lesser extent, the social, economic, political, and cultural life of the area.

This article treats the history of the Protestant movement and the teachings, practices, and organizational principles both common among and peculiar to the respective Protestant denominations. For further treatment of the life and works of the two principal Reformation leaders, see CALVINISM, CALVIN AND; LUTHER. See also *Micropædia* for biographical treatment of other Reformers (*e.g.*, John Knox; Thomas Müntzer; Huldrych Zwingli).

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 811, 827, and 961, and the *Index.*

The article is divided into the following sections:

History of the Protestant movement 206
  The context of the late medieval church 207
  The continental Reformation: Germany, Switzerland, and France 208
    The role of Luther
    Radical Reformers related to Luther's reform
    Zwingli and his influence
    The role of Calvin
    Calvinism in France
  The Reformation in England and Scotland 213
    Henry VIII and the separation from Rome
    The role of John Knox
    The rise of Puritanism
  The expansion of the Reformation in Europe 217
  Protestant renewal and the rise of the denominations 218
    The setting for renewal
    The rise of Pietism
    Rationalism
    Evangelicalism in England and the colonies
    Legacies of the American and French revolutions
    Movements toward reunion
    The revival of Pietism
    The era of Protestant expansion
    Revivalism in the 19th century
    New issues facing Protestantism in the 19th century
  Protestantism in the 20th century 227
    Mainstream Protestantism
    Conservative and evangelistic forms of Protestantism
    Theological movements within Protestantism
    The ecumenical movement
The Protestant heritage 230
  Teaching, worship, and organization 230
    Common principles and practices of the magisterial Reformers and their successors
    Common principles and practices of the radical Reformers and their successors
  Protestantism's influence in the modern world 236
    Influence on nationalism
    Influence on the arts
    Ecumenical concerns
    Conclusion
The major Protestant denominations 237
  Lutheran churches 237
    History
    Teachings
    Worship and organization
  Reformed and Presbyterian churches 242
    History
    Teachings
    Worship and organization
  Anglican Communion 246
    History
    Teachings
    Worship and organization
  Baptists 249
    History
    Teachings
    Worship and organization
  Congregationalists 252
    History
    Teachings
    Worship and organization
  Friends 255
    History
    Teachings
    Worship and organization
  Methodists 258
    History
    Teachings
    Worship and organization
  Disciples of Christ 260
    History
    Teachings
    Worship and organization
  Unitarians and Universalists 263
    History
    Teachings
    Worship and organization
Bibliography 265

HISTORY OF THE PROTESTANT MOVEMENT

Origin of the term Protestant

Protestantism was given its name at the Diet of Speyer in 1529. At that imperial assembly the Roman Catholic princes of Germany, along with the Holy Roman emperor Charles V, rescinded most of what toleration had been granted to the followers of Martin Luther three years earlier. On April 19, 1529, a protest was read against this decision, on behalf of 14 free cities of Germany and six Lutheran princes, who declared that the decision did not bind them because they were not a party to it, and that if forced to choose between obedience to God and obedience to Caesar they must choose obedience to God. They appealed from the diet to a general council of all Christendom or to a congress of the whole German nation. Those who made this protest became known as Protestants. The name was adopted not by the protesters but by their opponents, and gradually it was applied as a general description to those who adhered to the tenets of the Reformation, especially to those living outside Germany. In Germany the adherents of the Reformation preferred the name evangelicals and in France Huguenots.

The name Protestant was attached not only to the disciples of Luther (*c.* 1483–1546) but also to the Swiss disciples of Huldrych Zwingli (1484–1531) and later of John Calvin (1509–64). The Swiss Reformers and their followers in Holland, England, and Scotland, especially after the 17th century, preferred the name Reformed.

In the 16th century the name Protestant was used primarily in connection with the two great schools of thought that arose in the Reformation, the Lutheran and the Reformed. In England in the early 17th century the word Protestant was used in the sense of "orthodox Protestant," as opposed to those who were regarded by Anglicans as

unorthodox, such as the Baptists or the Quakers. Roman Catholics, however, used it for all who claimed to be Christian but opposed Catholicism (except the Eastern churches). They therefore included under the term Baptists, Quakers, and Catholic-minded Anglicans. Before the year 1700 this broad usage was accepted, though the word was not yet applied to Unitarians. The English Toleration Act of 1689 was entitled "an Act for exempting their Majesties' Protestant subjects dissenting from the Church of England." But the act provided only for the toleration of the opinions known in England as "orthodox dissent" and conceded nothing to Unitarians. Throughout the 18th century the word Protestant was still defined in relation to the historical reference of the 16th-century Reformation. Samuel Johnson's dictionary (1755), which is characteristic of other dictionaries in that age, defines the word thus: "one of those who adhere to them, who, at the beginning of the reformation, protested against the errours of the church of Rome." (W.O.C.)

The context of the late medieval church

The Protestant Reformation occurred against the background of long developments and rich ferment in the Roman Catholic Church and the world of the late Middle Ages. For two reasons it has been difficult to gain perspective on those times. Catholic historians had an interest in showing how much reform was occurring before and apart from the radical disrupters of the 16th century, the Protestant reformers. Protestant historians, on the other hand, portrayed the late medieval church in the most negative terms to show the necessity of the Reformation, which consequently came to look like a complete break with a corrupt past.

The other reason for difficulty in understanding stems from the fact that the 15th-century agents of change were not "Pre-Reformers"; they neither anticipated Protestantism nor acquired their importance only from the subsequent Reformation. The events of that period were also not "Pre-Reformation" happenings but had an identity and meaning of their own.

There has always been agreement on the fact that there were reform developments and ferment in the 15th-century church all the way from Spain and Italy northward through Germany, France, and England. Some of these were directed against abuses by the papacy, the clergy, and monks and nuns. The pious, for example, abhorred Innocent VIII (1484–92), who performed marriage ceremonies for his own illegitimate children in the Vatican, and Alexander VI (1492–1503), who was and was seen to be depraved. The public was also increasingly aware of and angered by luxurious papal projects, for which funds were exacted.

The distaste for the papacy increased at a time of rising



Religious situation in Europe in 1546.



The suffering of Christ (left) contrasted with the worldly glory of the pope. Woodcuts from *Passional Christi und Antichristi*, from the studio of Lucas Cranach, 1521.
By courtesy of the trustees of the British Museum; photographs, John R. Freeman & Co. Ltd.

nationalist spirits. The popes, who had long intervened in the politics of Germany, France, and England, faced setbacks when the monarchies in each country acquired new power. The sovereigns found a need to assert this power against the papacy and, in most cases, against local clerical representatives of the church.

At this time of rising national consciousness there appeared a generation of theologians who remained entirely within the context of medieval Roman Catholicism but who engaged in fundamental criticisms of it. Thus William of Ockham (d. 1349?) spoke up as a reformer within the Franciscan order. He wished to return this religious order to the ideal of poverty, which it had in large part abandoned. As part of his reform he maintained that Pope John XXII was heretical. Ockham saw the papacy and empire as independent but related governments or realms. When the church was in danger of heresy, lay people—princes and commoners alike—must come to its rescue. This meant, in the present case, reform.

Ockham's reform efforts

In England, John Wycliffe engaged in similar struggles, which weakened papal power and the hold of the medieval church. Wycliffe also traded on national consciousness, which he directed toward reform of the church. His instrument was the moral law of the Bible. Wycliffe gave impetus to its translation, and in 1380 he helped make it available to rulers and ruled alike, though he always granted uncommon spiritual authority to the king.

In Bohemia, Jan Hus, who became rector of the University of Prague, used that school as his base to criticize a luxury-minded clergy. He also exploited national feelings and came to argue that the pope had no right to use the temporal sword. Hus's bold accusations led to his death by burning at the Council of Constance in 1415.

Alongside a piety that combined moral revulsion with national feelings, Christian humanism was a further sign of stirring in the late medieval church. In Italy, Lorenzo Valla (1407–57) used his sophisticated techniques of historical inquiry to expose a number of forgeries that had given the papacy many of its powers and much of its domain. In Germany, Johannes Reuchlin (1455–1522) studied Greek and Hebrew, the biblical languages, and fought for the rights of scholars to question traditional claims of the church. In Holland, Desiderius Erasmus (1466/69–1536), who remained a Roman Catholic, used his vast learning and his satiric pen to question the practices of the church.

Still another factor that disturbed a complacent late medieval church was a flowering of mysticism in the spirit of Meister Eckehart (d. 1327/28) or Johann Tauler (d. 1361). These people of profound devotion gained followers who sought and claimed to have a direct access to God, bypassing many of the church's rites and practices. Reformers like Martin Luther were to speak well of some of these devotionalists and to translate their writings.

While the Reformers attacked people in high places, they also regarded the Catholicism of ordinary people as being in need of reform. Devotion to the Virgin Mary had come to look superstitious to them as well as to occur at the expense of devotion to Christ. Such practices as pilgrims visiting shrines or parishioners regarding relics of saints with awe seemed to perpetuate a kind of paganism under a Christian veneer. The pestilences and plagues of the 14th century had bred an inordinate fear of death, which led to the exploitation of simple people by a church that was, in effect, offering salvation for sale. By the turn of the 16th century much of Europe was ripe for reforms that Catholicism could neither open itself up to nor contain.

(M.E.M.)

The continental Reformation: Germany, Switzerland, and France

THE ROLE OF LUTHER

Luther said that what differentiated him from previous reformers was that they attacked the life, he the doctrine of the church. Whereas they denounced the sins of churchmen, he was disillusioned by the whole scholastic scheme of redemption. The assumption was that man could erase his sins one by one through confession and absolution in the sacrament of penance. Luther discovered that he could not remember or even recognize all of his sins, and the attempt to dispose of them one by one was like trying to cure smallpox by picking off the scabs. Indeed, he believed that the whole man was sick. The church, however, held that the individual was not too sick to make up for bad deeds by some good deeds. God gave to all a measure of grace. If human beings lay hold of it and did the best they could, God would reward them with a further gift of grace with which they could perform deeds of genuine merit, which would give them credit before God. Human beings might even die with more than enough credits for salvation. These extra credits constituted a treasury of the merits of the saints, from which the pope could make transfers to those whose accounts were in arrears. The transfer was called an indulgence and for this, in Luther's day, the grateful recipient made a contribution to the church.

The indulgence system and its effects

This arrangement proved to be a popular way of raising money particularly because, unlike tithes, it was voluntary and could provoke no resentment. By this means crusades, cathedrals, hospitals, and even bridges were financed. At first the indulgence, according to the Germanic law of commutation of a physical punishment to a fine, applied only to penalties imposed by the church on earth. Then it was extended to penalties imposed by God in purgatory. In Luther's day immediate release from purgatory was being offered, and the remission not only of penalties but even of sins was assured. Thus the indulgence encroached upon the sacrament of penance.

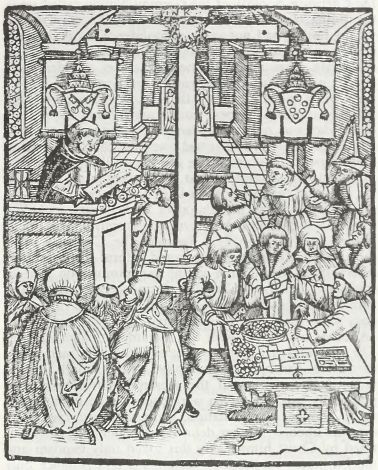
Luther was desperately in earnest about his standing before God and Christ. The woodcuts of Christ the Judge on a rainbow consigning the damned to hell filled him with terror. He believed the monastic life to be the way par excellence to acquire those extra merits that would more than balance his account. He became a monk and subjected himself to rigorous asceticism, but he could never reach the assurance that a sinful pygmy like himself could ever stand before the inexorable justice and majesty of God. Continual recourse to the confessional simply convinced him of the fundamental sickness of the whole man. He began then to question the goodness of a God who would make human beings so weak and then damn them for what they could not help. Relief came through the study of the Psalms. Luther found the 22nd Psalm particularly revealing because it contains the words quoted by Christ upon the cross, "My God, my God, why hast thou forsaken me?" Evidently then, Christ, being without sin, so identified himself with sinful humanity as to feel himself estranged from God. Christ the Judge seated upon the rainbow had become Christ the Derelict upon the cross, and here the wrath and the mercy of God could find their point of meeting so that God was able to forgive those utterly devoid of merit. He could justify the unjust, and this required of man only that he accept the gift of God in faith. This was the doctrine of justification by faith, which became the watchword of the Reformation.

What this insight meant for the doctrine of indulgences is at once apparent. The great offense was not the financial aspect but rather the very notion that human beings dared to engage in bookkeeping with God. Luther by now had become a professor at the University of Wittenberg and also a pastor. His parishioners were obtaining the indulgences issued by Albert, the new archbishop of Mainz, half of the proceeds to be retained by him as reimbursement for his installation fee as archbishop, the other half to go to the pope for the building of the Basilica of St. Peter's at Rome. For this indulgence Albert made unprecedented claims. If the indulgence were on behalf of the donor himself, he would receive preferential treatment in case of future sin, if for someone else already in purgatory, he need not be contrite for his own sin. Remission was promised not only of penalties but also of sins, and the vendor of the indulgences offered immediate release from purgatory.

Ninety-five Theses

Against these instructions Luther launched his Ninety-five Theses on All Saints' Day of the year 1517. In the theses he presented three main points. The first concerned financial abuses; for example, if the pope realized the poverty of the German people, he would rather that St. Peter's lay in ashes than that it should be built out of the blood and hide of his sheep. The second focused attention on doctrinal abuses; for example, the pope had no jurisdiction over purgatory and if he did, he should empty the place free of charge. The third attacked religious abuses; for example, the treasury of the merits of the saints was denied by implication in the assertion that the treasury of the church was the gospel. This was the crucial point. When the papacy pronounced Luther's position heretical, he countered by denying the infallibility of popes and for good measure of councils also. Scripture was declared to be the only basis of authority.

Luther found support in many quarters. Already a widespread liberal Catholic evangelical reform sought to correct the moral abuses such as clerical concubinage, financial extortion, and pluralism (*i.e.*, the holding of several benefices by one man) and ridiculed the popular superstitions associated with the cult of the saints and their relics, religious pilgrimages, and the like. This movement had representatives in all lands, notably John Colet in England, Jacques Lefèvre in France, Francisco Jiménez de Cisneros in Spain, Juan de Valdés in Naples, and, above all, Erasmus of Rotterdam. Erasmus found nothing amiss in Luther's theses except that he had been too tart as to purgatory, and when the cry of heresy was raised against Luther, he wrote to the elector Frederick III the Wise, Luther's prince, telling him that as a Christian ruler he was obligated to see to it that his subject should have a fair hearing.



The sale of indulgences in church; woodcut from the title page of Luther's pamphlet "On Aplas von Rom," published anonymously in Augsburg, 1525.

By courtesy of the trustees of the British Museum; photograph, John R. Freeman & Co. Ltd.

Another party that rallied to Luther was that of the German nationalists led by Ulrich von Hutten, who aspired to convert the Holy Roman Empire into a German national state. This program would entail the suppression of the whole system of prince-bishops and could never be achieved without a war with the papacy. Luther was hailed because of his attack on the papacy, though he would not condone the program of violence.

Reaction of Pope Leo X

Yet despite the support from these parties, Luther would have been speedily crushed had Pope Leo X taken seriously the religious side of his office. The secularization of the papacy saved Luther, and he destroyed the secularization of the papacy. At the moment when Luther appeared to be foredoomed, an election for the office of Holy Roman emperor was pending. It was elective and any European prince was eligible, including Henry VIII of England, Francis I of France, Charles I of Spain. The Pope wished none of them because the position entailed control over Germany, and the augmentation of power to one of the three would destroy the balance of power. His preference was for a minor prince, and none fitted the role better than Luther's protector, Frederick the Wise of Saxony. In consequence the Pope dallied in the case of Luther and even after Charles was elected, the Pope was willing to play Frederick against him. Not until June 1520, nearly three years after the Ninety-five Theses, was Luther summoned to submit within 60 days. The time was reckoned from the date of the actual delivery of the bull to the person named. So great was the obstruction to Rome on the part even of German bishops that the bull was not handed to Luther until October 10.

He employed the summer of 1520 to bring out some of the great manifestos of the Reformation. The *Address to*

the Christian Nobility of the German Nation called upon the ruling class in Germany, including the emperor, in whom Luther had not yet lost confidence, to reform the church in externals by returning to apostolic poverty and simplicity. This appeal to the civil power to reform the church was a return to the earlier practice of the Middle Ages when emperors more than once had deposed and replaced unworthy popes. Luther argued that the papacy of his day was only 400 years old, meaning that it was the Gregorian reform that had given the church its lead in matters political, encroaching thereby on the sphere of the magistrate on the ground that the lowliest priest did more for mankind than the loftiest king. Luther countered with the doctrine of the priesthood of all believers, including Christian magistrates. Any layman was spiritually a priest, though not vocationally a parson. The Christian ruler, then, being himself a priest, could reform the church in externals, as the church might excommunicate him in spirituals. The liberal Catholic reformers could sympathize with this program except for the identification of the papacy with Antichrist. This savoured of the medieval sects.

The sacraments as defined in *The Babylonian Captivity*

Another tract dealt with the sacraments. The title was *The Babylonian Captivity,* meaning that the sacraments themselves had been taken captive by the church. Luther reduced the number of the sacraments from seven to practically two. The seven were baptism, the Eucharist or mass, penance, confirmation, ordination, marriage, and extreme unction. Luther defined a sacrament as rite instituted by Christ himself. By this token only baptism and the Eucharist were strictly sacraments and penance only as confession. Extreme unction, that is anointing with oil those on the verge of death, was dropped entirely. Confirmation went out for a time but was later restored. Ordination continued as a rite of the church. Penance included contrition, confession, and satisfaction. Luther felt that none could be sure of genuine contrition, none could make satisfaction. Confession was wholesome but should be voluntary and could be made to any fellow Christian. Marriage was not a Christian sacrament, because it was not instituted by Christ but by God in the garden of Eden, and valid not only for Christians but also for Turks and Jews. Baptism was to be administered but once only and to babies on the ground of their dormant faith.

This left the mass, and at this point Luther gave the greatest offense. The wine, he asserted, should be given to the laity as well as the bread, as in the Hussite practice. No masses should be said for the dead by the priest alone without communicants, because the Eucharist involved fellowship not only with Christ but also with believers. The most drastic change was that Luther denied the doctrine of transubstantiation, according to which, at the pronouncement of the words of institution, the elements of bread and wine, though retaining their accidents of colour, shape, and taste, nevertheless lost their substance, which was replaced by the substance of the body of Christ as God. This Luther denied, saying that no change was wrought by the words of Christ.

The doctrine of Real Presence

Luther, nevertheless, believed that the body of Christ was physically present upon the altar because Christ said, "This is my body." Therefore, in some inexplicable manner, his body must be "with, in, and under" the elements. But if no change was wrought, how did his body come to be on the altar? Because his body was everywhere. But if everywhere, why especially there? Because in view of human limitations God had decreed two modes of self-disclosure, the preaching of the Word and the administration of the sacrament. There the eyes of the believer were opened. This view undercut sacerdotalism, since the words of the priest did not bring the body of Christ to the altar. The undercutting of sacerdotalism destroyed the hierarchical structure of society culminating in the papacy.

But what was to be done with Luther? On December 10, instead of submitting, he defiantly burned the papal bull together with a copy of the canon law. The normal course would then have been to excommunicate him outright, but Frederick the Wise insisted that he be given a fair hearing. The natural body to pass judgment would have been a council of the church. But the popes were the greatest obstructionists when it came to calling a council



Luther and Hus distributing the sacramental bread and wine to the Elector of Saxony and his family. Woodcut by an unknown artist.

By courtesy of the Lutherhalle, Wittenberg

because they feared the revival of conciliarism, which in the previous century bade fair to convert the church into a constitutional monarchy. There would have been no Council of Trent save for Luther. Only after another 20 years, when the spread of his teaching left no other expedients, was a council convened. Consequently, his hearing had to be before a secular tribunal, the Diet of the empire meeting at Worms in the winter and spring of 1521. Since this was a secular tribunal the attempt was made to prove that he was not simply a heretic but also a rebel whose views were more subversive of the civil than of the ecclesiastical order, because he was undermining the very principle of authority. Luther was brought before the Diet and given an opportunity to repudiate his books. Had he disclaimed the one on the sacraments, the other points might have been negotiated. He acknowledged them all. Would he then disclaim some of their teaching? Who was he to reject the teaching of the ages? Let him give an answer without horns, to which he replied: "I will answer without horns and without teeth. Unless I am convicted by Scripture and plain reason—I do not accept the authority of popes and councils, for they have contradicted each other—my conscience is captive to the Word of God, I cannot and I will not recant anything, for to go against conscience is neither right nor safe. God help me. Amen." The Emperor then placed Luther under the imperial ban. The bull of excommunication by the church was formally released only later. Frederick the Wise at this point intervened and wafted Luther away to a place of hiding.

Diet of Worms

Luther was concealed for a year at the castle of the Wartburg. During this enforced withdrawal he made perhaps his greatest contribution in that he translated the whole of the New Testament from the Greek text of Erasmus into an idiomatic, pungent, powerful German. In many respects his German helped to create the idiomatic. Nothing did so much to win popular adherence to his teaching as the dissemination of this translation.

But some were not so convinced. Many of the liberal Catholic reformers, like Erasmus, recoiled from Luther's paradoxes, from his confidence that his interpretation of Scripture was correct, from his acceptance of the doctrine of predestination, which makes of God a tyrant when he elects some and damns others regardless of their behaviour. The German national movement collapsed. Then in Luther's own circle variant forms of Protestantism arose, which in the aggregate are variously described as the left wing of the Reformation or as the radical Reformation. The terminology does not matter so much as the recognition that no neat classification is possible.

RADICAL REFORMERS RELATED TO LUTHER'S REFORM

Karlstadt and Thomas Müntzer

Two figures emerging in Luther's circle are significant by way of anticipation. One was Karlstadt (*c.* 1477/81–1541), who drew the radical inference from the dualism of flesh and spirit that art and music should be abolished as external aids to religion and the Presence of Christ's body on the altar should be interpreted in a spiritual sense. His program issued in iconoclastic riots. He extended Luther's doctrine of the priesthood of all believers to mean that all laymen were pastors. If one person was assigned the tasks of a parson, he should dress like others and, like others, should work with his hands. The clergy not only might but must marry. The sabbath should be strictly observed. This program anticipated the Puritan movement. It entailed a blending of spiritualism and legalism. The sensory aids to religion were to be discarded by those advanced in the spiritual life and then snatched away by laws from those still weak.

A much more disquieting figure than Karlstadt was Thomas Müntzer (*c.* 1490–1525), a man of learning and a creative firebrand, who may be regarded not as the progenitor but as the first formulator of the concept of the Protestant Holy Commonwealth. He believed that the elect, those predestined by God for salvation, could be sufficiently identified to compose a definite group. Luther denied the possibility of distinguishing the elect from the nonelect. Müntzer's test was the new birth in the spirit. The test was not for him an absolute mark, and he recognized that among the wheat there might be some weeds, yet he accepted it as an adequate test for the formation of a community bound together by a covenant. The mission of this group was to set up the Kingdom of God on Earth, the Holy Commonwealth, by wiping out the ungodly. In the attempt they would have to endure suffering, and here Müntzer drew from German mysticism the theme of walking in Christ's steps toward the cross. But the trial would end in triumph, for the Lord Jesus would speedily come to vindicate his saints and erect his Kingdom. There are obviously incompatibles here, the way of suffering and the infliction of suffering, the feverish activity of man to achieve that which will be established by God. But logical incompatibles fuse at high emotional temperatures. Müntzer appealed to the Saxon princes to implement his program, but they banished him. He found a hearing among the revolting peasants and led them at the Battle of Frankenhausen, where they were butchered and he captured and beheaded. Luther execrated his memory because he seized the sword in defense of the gospel. The Marxists have exalted him as the prophet of social revolution because he was the only one of the Reformers who had a deep feeling for the sufferings of the socially oppressed. In grasping the sword he did not essentially differ from Huldrych Zwingli, Gaspard de Coligny, or Oliver Cromwell.

ZWINGLI AND HIS INFLUENCE

Huldrych Zwingli (1484–1531), the great figure in Swiss Protestantism, was in fact if anything more committed to military action than Müntzer because he fell as a combatant with sword and helmet on the field of battle. He became a Reformer independently of Luther, with whom he was entirely in accord as to justification by faith and predestination. At certain points Zwingli drew from Erasmus and Karlstadt, notably with respect to the disparagement of the sensory aids to religion. Zwingli, though an accomplished musician, considered that the function of music was to put the babies to sleep rather than to worship God. The organ was dismantled and the images removed from the cathedral at Zürich. The Lord's Supper was understood by Zwingli in his most extreme period simply as a memorial of Christ's death and, on the part of the recipient, as a public declaration of faith with more significance for the members of the congregation who saw him take his stand than for his own spiritual life. Zwingli could the more readily retain the baptism of infants because it was simply a recognition that the child belongs to the people of God as the child in the Old Testament belonged by circumcision to Israel. The analogy with Judaism applied at many points, for Zwingli regarded the Christian congregation as the new Israel of God, an elect people, reasonably identifiable, not as with Müntzer by the new birth but by adherence to the faith. This company could be called theocratic in the sense that it was under the rule of God, whom church and state should alike serve in close collaboration. The identification of the whole populace of Zürich with this elect people was the more tenable because those not in accord with the ideal were disposed to leave. Zwingli approved of even an aggressive war to forestall interference from the Roman Catholic cantons. In the second war of Kappel he fell in 1531.

Zwingli's theocratic ideals

In Zwingli's circle arose the group who formed the mainstay of the radical Reformation. They shared with Zwingli, and with all the reformers to a degree, the desire to restore the church to the primitive pattern, but they were more drastic in their restitution. Manifestly the early church had not been allied with the state. Luther, Zwingli, and other Reformers saw no sense in forcing the church back into the period when the state was hostile and the Christians were persecuted. After the state became Christian, there could very well be a close alliance, as indeed there had been in ancient Israel.

The Anabaptists. The radicals restricted their biblicism to the New Testament and espoused three tenets therefrom that have come to be axiomatic in the United States: the separation of church and state, the voluntary church, and religious liberty. They were called Anabaptists on the ground that, having rejected infant baptism, they rebaptized adults previously baptized. But they called themselves simply Baptists, denying that they repeated baptism since the dipping of babies was no baptism at all. Baptism, they held, did not itself regenerate but was only the outward sign of an inner experience, the rebirth in the spirit, of which only an adult was capable. The Anabaptists, so-called, also believed in the possibility of a Christian society whose members were marked both by the conversion experience and also by a highly disciplined deportment. In obedience to the New Testament they repudiated swearing oaths and recourse to violence, whether in war or at the hands of the magistrate. The saints should withdraw from the wicked world.

This whole program obviously had political and social aspects and was a threat to that society or any other, for no society, save that of a small sect, has ever renounced the use of the sword. The Anabaptists were marked for extermination by Catholics and Protestants alike. One of their first leaders, Felix Manz, was drowned in Zürich in 1527. The Diet of Speyer in 1529, at which the Lutherans protested, subjected the Anabaptists to the penalty of death with the concurrence of the Lutherans. Persecution in the first decade eliminated the leaders, most of them educated and moderate men. Less temperate spirits came to the fore, sustaining their courage by setting dates for the speedy coming of the Lord. One band, composed mainly of Anabaptists, took over the town of Münster in Westphalia in 1534 and, contrary to the tenets of their fellows, seized the sword and, in accord with Old Testament practice, restored polygamy. The town was captured by Catholics and Lutherans conjoined and the leaders were executed. Persecution everywhere intensified.

Other groups. In Holland Menno Simons (*c.* 1496–1561), the founder of the Mennonites, repudiated violence, polygamy, and the setting of dates for the coming of the Lord and returned to the teaching of the early founders. The Mennonites survived partly by reason of accommodation to military service in Holland, partly by migration first to eastern Europe and then to the Americas. Another group, named Hutterites from Jakob Hutter (died 1536), was allowed to form communal colonies in Moravia on the estates of tolerant feudal nobles who were willing to drop the demand for military service in return for excellent craftsmanship in field and shop. Because of subsequent persecution these groups also migrated to the New World. The Swiss branch, which survives in the United States, is called the Amish. The entire pattern of ideas has reappeared in various combinations in subsequent history, not only among the Church of the Brethren and the Quakers but among all of the free churches disclaiming a state connection.

Mennonites and Hutterites

THE ROLE OF CALVIN

Another form of Protestantism was Calvinism, named for John Calvin (1509–64), a Frenchman educated in humanist and legal studies, who in consequence of a conversion to the Protestant reform had to flee France. In Basel, at the age of 27, he brought out the first edition of his *Institutes of the Christian Religion,* which in successive expansions became for centuries the manual of Protestant theology. Calvin was in basic agreement with Luther as to justification by faith and the sole authority of Scripture. On the sacrament of the Lord's Supper he took a mediating position between the radical Swiss and the Lutheran view. Thus he believed that the body of Christ was not everywhere present, but that his spirit was universal and there was a genuine communion with the risen Lord. Calvin took a middle view likewise with respect to music and art. He favoured congregational singing of the Psalms, and this became a characteristic mark of the Huguenots in France and the Presbyterians in Scotland and the New World. As to art, he rejected the images of saints and the crucifix (that is, the body of Christ upon the cross), but allowed a plain cross. These modifications do not refute the generalization that Calvinism was alien to art and music in the service of religion but not in the secular sphere.

As over against Luther, there was a shift of emphasis in Calvin, whose *Institutes* did not begin with justification by faith but with the knowledge of God. Luther found refuge from the terror of God's dispensations in the mercy of Christ. Calvin could the more calmly contemplate the frightfulness of God's judgments because they would not descend upon the elect. Luther, as noted, saw no way of knowing who were the elect. He could not be sure of himself and throughout his life had a continual struggle for faith and assurance. Calvin had certain approximate and attainable tests. He did not require the experience of the new birth, which is so inward and intangible, though to be sure later Calvinism moved away from him on this point and agonized over the marks of election. For Calvin there were three tests: the profession of faith, as with Zwingli; a rigorously disciplined Christian deportment, as with the Anabaptists; and a love of the sacraments, which meant the Lord's Supper since infant baptism was not to be repeated. If a person could meet these three tests let him assume his election and stop worrying.

If one could achieve such assurance, what an enormous release of energy to be directed to the glory of God and the erection on Earth of some semblance of a holy commonwealth! The term became common in New England. Calvin's own statement was that "the Church reformed is the kingdom of God." Calvin saw more of a possibility of its realization through the efforts of the elect because he muted the expectation of the imminent return of the Lord. The service of the Kingdom did not require a particular vocation. Any worthy occupation is a divine calling demanding unremitting zeal. Luther had emphasized the secular callings as over against the monastic, which in the Middle Ages alone had been called a vocation. With Calvin the point was not so much that one should accept one's lot and rejoice in the assigned task, however menial, as that the work would contribute to the larger realization of the Christian society.

Geneva as the centre of Calvinistic reform

Calvin had a concrete opportunity for the realization of his ideal, albeit at first only on a small scale. The city of Geneva had recently thrown off the authority of the bishop and of the duke of Savoy and had not yet joined the Protestant Swiss Confederation, though aided in the fight for liberation by the Protestant city of Bern. Through the Bernese, Protestant preachers began to evangelize Geneva. The city was threatened by civil war. The bellicose preacher Guillaume Farel, unable himself to contain the violence he had helped to unleash, laid hold of Calvin merely passing through the city and impressed him into the unwelcome task of leadership. After turbulent years, a banishment and a recall, he was able for the last two decades of his life to direct the city that John Knox considered "the most godly since the days of the apostles." There was actually scarcely a feature of Thomas More's *Utopia* that Geneva did not seek to realize.

The program, despite all the turbulence, was the more attainable because of a selective process with respect to the population. At the outset all the Catholics who would not submit to the new regime had to leave. Among those who remained, excommunication from the church, if not removed within six months, meant banishment from the city. Control over excommunication, after a long struggle, came to be entirely in the hands of the church. The state, having long suffered from the abuse of excommunication for political purposes, was loath to concede to the church exclusive control. Abortive attempts to achieve independence had been made by the Protestant churches at Basel and Strassburg. Calvin succeeded, with the result that one who was not in the graces of the church could not for long be a member of the community. A further factor ensuring a select constituency was the influx of 6,000 refugees from France, Italy, Spain, and, for a time, from England into a city of 13,000. Thus in Geneva, church, state, and community came to be one. The ministers and the magistrates with differentiated functions were alike the servants of God in the erection of this new Israel; and the comparison with ancient Israel was the more striking and the inner cohesion the more intensified because Geneva also was begirt by foes, the duke of Savoy and the duke of Alba, like the old Canaanites and Philistines.

CALVINISM IN FRANCE

The situation in France with respect to the Reformation was not altogether dissimilar to that in Germany because, although the decentralization of government was not as great, some of the French provinces enjoyed a considerable autonomy, particularly in the south, and it was in the Midi and French Navarre that the Protestant movement had its initial strength. Then, too, noble houses were continually conspiring to manipulate or eviscerate the monarchy. The religious issues came to be intertwined with the political ambitions. The ruling houses, first the Valois from Francis I through Henry III and then the Bourbon, beginning with Henry IV, sought to secure the stability of the land and the throne by quelling religious strife either by the extermination or toleration of minorities.

The ground was better prepared for the reform of the church in France than in Germany because of the efforts of liberal Catholics such as the scholar Jacques Lefèvre d'Étaples and the bishop of Meaux, Guillaume Briçonnet. King Francis I and his sister Margaret of Angoulême not infrequently intervened to save humanist reformers from the menaces of the obscurantists, and Margaret's daughter, Jeanne d'Albret, the queen of Navarre, a feudatory of France, provided an asylum for the persecuted in her domain, though she did not herself espouse the Huguenot cause until 1560. When Lutheran teaching first began to infiltrate France, Francis I, who would not abet heresy, fluctuated in his policy of repression, depending on whether he desired a political alliance with the pope, the Turk, or the German Lutherans. The year 1534 precipitated a crisis when placards were posted in Paris savagely attacking the mass. Severe repression followed. Bishop Briçonnet made his submission. Farel fled to Geneva, Lefèvre to Strassburg, Calvin to Basel. Under Henry II, the son of Francis, repression was intensified, particularly when in 1559 France and Spain made peace and thus each was free to devote attention to the suppression of heresy at home. The persecution of the Huguenots, as the Protestants came to be called in France, would have been intense save for the death of the King in a tournament.

The Huguenots and the French religious wars

At this point the rivalry of the noble houses injected itself more overtly into the religious struggle. The crown, with its alternating policy of eradication or recognition, was flanked by two extreme houses for whom the religious issue was of intense concern. The House of Guise was so Catholic as to be willing to call in Spanish aid, and the family of Admiral Coligny so Huguenot as to be willing to court help from England and even from Germany. Under Francis II the Guises were in the ascendant because the queen, later queen of Scots, was of that house. Some of the Huguenots, foreseeing the suppression in store, hatched the Conspiracy of Amboise, an attempted assassination of the leaders of the Guise party and transferral of power to the House of Bourbon.

This was plainly rebellion and acutely raised a problem with which the Protestants had long been wrestling. The Lutherans had had to face it earlier when the Diet of Augsburg in 1530 gave them a year in which to submit on pain of war. The Lutheran princes then had formed the Schmalkaldic League to resist arms with arms. Luther was loath to condone any use of the sword in defense of the Gospel and absolutely forbade any recourse to violence on the part of a private citizen against the magistrates. This was his reason for disapproval of the Peasants' War. But now the jurists pointed out to Luther that the emperor was an elected ruler and that if he transgressed against the true religion he might be brought to book by the electors, who also were magistrates. Thus arose the doctrine of the right of resistance of the lower magistrate against the higher. The concept lost its pertinence in Germany after the Peace of Augsburg in 1555, which granted toleration to the Lutherans in the territories where they were predominant. Minorities in Lutheran and Catholic lands were granted the right of migration without loss of goods.

But the Calvinists were not included in the peace, and the problem of armed resistance again became acute in France. Calvin would not condone the Conspiracy of Amboise because it was not led by a lower magistrate. The term was now applied to the princes of the blood in line for succession to the throne. This meant the House of Bourbon. The Conspiracy of Amboise failed. Francis II died, and was succeeded by his brother, the young Charles IX. The queen mother, Catherine de Médicis, took the lead and sought to avert religious war by granting the Huguenots limited toleration in restricted areas in the edict of 1562. When François, duc de Guise, discovered the Huguenots worshiping outside the prescribed limits, as he claimed, he opened fire. The Massacre of Vassy set off the wars. The Huguenots now were led by a prince of the blood, Louis I, 1st prince de Condé, of the House of Bourbon. Calvin approved. There followed three inconclusive wars. Condé was killed in the first and François, duc de Guise, was assassinated. His son, now Henri, duc de Guise, believed in the complicity of Coligny, the new leader of the Huguenots. At the end of 10 years of indecisive conflict, Catherine made another effort at a settlement to be cemented by the marriage of Henry of Navarre, a Bourbon, the son of Jeanne d'Albret and the hope of the Huguenots, and her own daughter Margaret (Marguerite de Valois), a Catholic. The leaders of all parties came to Paris for the wedding. The Duke of Guise made an attempt on the life of Coligny, which failed. Then the Guise, with the connivance of Catherine and her son Charles, who panicked, tried to wipe out all of the leaders of the Huguenot party in the Massacre of St. Bartholomew's Day in August 1572. Other massacres followed in the provinces.

The St. Bartholomew's Day Massacre and the Edict of Nantes

Charles IX was succeeded by his brother, Henry III, two years later (1574). Such was the revulsion against the massacre that the King could rule only by forming an alliance with the Huguenot Henry of Navarre. A fanatical Catholic was thereby so outraged that he assassinated the King. Both sides had abandoned the fiction of the inferior magistrate and had gone in unabashedly for popular revolution. Henry of Navarre then became Henry IV, but he was unable to take Paris and rule France so long as he was a Protestant. In order to pacify the land he made his submission to Rome and promulgated an edict of toleration for the Huguenots, the Edict of Nantes, in 1598. It gave them liberty of worship again in limited areas but full rights of participation in public life. The edict remained in force until the revocation in 1685.

The Reformation in England and Scotland

HENRY VIII AND THE SEPARATION FROM ROME

In the meantime the Reformation had taken hold in England. The beginning there was political rather than religious, a quarrel between the king and the pope of the sort that had occurred in the Middle Ages without resulting in a permanent schism, and might not have in this instance save for the total European situation. The dispute had its root in the assumption that the king was a national stallion expected to provide an heir to the throne. England did not have the Salic law, which in France forbade female succession, but England had just emerged from the Wars of the Roses and the fear was not unwarranted that the struggle might be resumed if there were not a male succession. Catherine of Aragon, the queen of Henry VIII, had borne him numerous children of whom only one survived, the princess Mary, and more were not to be expected. The ordinary procedure in such a case was to discover some flaw in the marriage that would allow an annulment or, in the terminology of that day, a divorce. In this instance the flaw was not difficult to find, because Catherine had been married to Henry's brother Arthur, and the law of England, following the prohibition in the book of Leviticus, forbade the marriage of a man with his deceased brother's widow. At the time of the marriage the pope had given a dispensation to cover this infraction of the rule. The question now was whether the pope had the authority to dispense from the divine law. Catherine said there had been no need for a dispensation because

England's problem of succession

By courtesy of the Musee Cantonal des Beaux Arts, Lausanne; photograph, Andre Held



"The Massacre of St. Bartholomew's Day," oil painting by François Dubois (1529–84). The Huguenot leader, Gaspard de Coligny, is shown twice, hanging from a window of his house and lying beheaded in front of the house with the Duc de Guise standing over him. In the Musée Cantonal des Beaux-Arts, Lausanne, Switz.

her marriage to Arthur had not been consummated and there had been no impediment to her marriage to Henry. The knot would have been cut by some casuistry had Catherine not been the aunt of Emperor Charles V, who was not prepared to see her cast aside in favour of another wife, and who controlled the pope. Clement VII, wishing neither to provoke the emperor nor to alienate the king, dallied so long that Henry took the matter into his own hands, repudiated papal authority, and in 1534 set up the Anglican Church with the king as the supreme head. The spiritual head was the archbishop of Canterbury, now Thomas Cranmer, who married Henry to Anne Boleyn. She bore the princess Elizabeth. By still another wife Henry did have a son who succeeded as Edward VI.

Although the basic concern of Henry was political, the alterations in the structure of the church gave scope for a reformation religious in character. Part of the impulse came from the survivals of Lollardy, part from the Lutheran movement on the Continent, and even more from the Christian humanism represented by Erasmus. The major changes under Henry were the suppression of the monasteries, the introduction of the Bible in the vernacular in the parish churches, and permission to the clergy to marry, though this was later revoked. The resistance to Henry's program was not formidable and the executions resulting were not numerous. Henry was impartial in burning some Lutherans who would not submit to his later reactionary legislation and toward some Catholics who would not accept the royal supremacy over the church, notably John Fisher and Thomas More.

On his ascension to the throne in 1547, young Edward VI was hailed by Cranmer and other Protestants as England's Josiah, the young 7th-century-BC king of Judah who enforced the Deuteronomic reform. Edward, it was held, would rid the land of idolatry so that England might be blessed. Protestantism advanced rapidly during his reign through the systematic reformation of doctrine, worship, and discipline—the three external marks of the true church. A reformed confession of faith and a prayer book were adopted, but the reformation of the ecclesiastical laws that would have defined the basis of discipline was blocked in Parliament by the most powerful of the English nobility.

The external marks of the church

The death of Edward and England's return to Roman Catholicism in 1553 under Queen Mary was interpreted by Protestants as a judgment by God upon a nation that had not taken the Reformation seriously enough. Many, including Cranmer, died as martyrs to the Protestant cause. Others fled to the European continent. Those in exile experimented with more radical forms of worship and discipline. Leading clergymen published material justifying rebellion against an idolatrous ruler. Many saw in Geneva, which was a haven for English exiles, a working model of a disciplined church. Exiles produced two large volumes of incalculable consequence for English religious thought. John Foxe's *Actes and Monuments,* popularly known as *The Book of Martyrs,* and the Geneva Bible were the most popular books in England for many years after they were published. They provided a view of England as an elect nation chosen by God to bring the power of the Antichrist (understood to be the pope) to an end. An England obedient to God would receive his favour. Otherwise, it would experience his plagues.

Elizabeth I, beginning her rule in 1558, was hailed as the glorious Deborah (12th-century-BC Israelite leader), the "restorer of Israel." She did not restore it far enough for English Protestants, however. Two statutes promulgated in her first year—the Act of Supremacy, stating that the queen was "supreme governor" of the Church of England, and the Act of Uniformity, ensuring that English worship should follow *The Book of Common Prayer*—defined the nature of the English religious establishment. In 1563 the primary church legislative body, the Convocations of Canterbury and York, defined standard doctrine in the Thirty-nine Articles, but attempts in the Convocation to reform the prayerbook further and to produce a reformed discipline failed. Defeated there, the reformers came to rely more on Parliament, where they could always depend on strong support.

THE ROLE OF JOHN KNOX

In Scotland the Reformation is associated with the name of John Knox, who declared that one celebration of the mass is worse than a cup of poison. He faced the very real threat that Mary, Queen of Scots, would do for Scotland what Mary Tudor had done for England. Therefore Knox defied her to her face in matters of religion and, though a commoner, addressed her as if he were all Scotland. He very nearly was, because in the period prior to 1560 many an obscure evangelist had converted the lowlands largely to the religion of John Calvin. The church had been given a Presbyterian structure, culminating in a General Assembly, which had actually as great and perhaps a greater influence than the Parliament. Because of her follies, and very probably her crimes (complicity in the murder of her husband), Mary had to seek asylum in England. There she became the focus of plots on the life of Elizabeth until Parliament decreed her execution. Presbyterianism came to be established in Scotland, and this very fact alone made possible the union of Scotland with England. Union of Protestant England with a Catholic Scotland would have been unthinkable.

Knox is frequently reproached for his intolerance in regarding one celebration of the mass as worse than a cup of poison, but one must remember that the year 1560 marked the peak of polarization between the confessions. Similar intolerance had been mounting at Rome. Paul III, after an abortive attempt at reform, had introduced the Roman Inquisition in 1542. His successor, Paul IV, placed everything that Erasmus had ever written on the Index. The Council of Trent began its sittings in 1545, introducing rigidity in dogma and austerity in morals. The Protestant views of justification by faith alone, the Lord's Supper, and the propriety of clerical marriage were sharply rejected. All deviation within the Catholic fold was rigidly suppressed. When Carranza, the archbishop of Toledo, returned to Spain in 1559, after assisting Mary in the restoration of Catholicism in England, he arrived in time for the last great auto-da-fé of the Lutherans. Himself under suspicion for ideas no more heretical than those of Erasmus, he was incarcerated for 17 years in the prison of the Inquisition. The liberal cardinal Giovanni Morone was imprisoned during the pontificate by Paul IV, and under Pius V, Pietro Carnesecchi, an Erasmian and one-time secretary of Clement VII, was burned in Rome. John Knox and Pope Pius V represent the acme of divergence between the confessions. (R.H.B./J.C.S./Ed.)

Council of Trent

THE RISE OF PURITANISM

Origins. Despite Elizabeth I's conservatism the Protestant reformers in England began to see their programs and ideas take hold more firmly during her reign. The movement known as Puritanism was part of this growing Protestant influence in English society in the late 16th and early 17th centuries.

Puritanism first emerged as a distinct movement in a controversy over clerical vestments and liturgical practices. Immediately following the Elizabethan Settlement, a practical latitude existed for Protestant clergy to wear what they chose while leading worship. Many preachers took this opportunity to do away with the formal attire as well as other practices traditionally associated with the Roman Catholic mass. But in 1564 Queen Elizabeth demanded that Matthew Parker, the archbishop of Canterbury, enforce uniformity in the liturgy. He did so somewhat reluctantly with the publication of his *Advertisements* in 1566. Those who refused to wear the now prescribed garb came to be considered collectively, and with scorn, as "Puritans" or "precisians" for their unwillingness to submit in these seemingly minor points to the supremacy of the queen.

Aside from vestments and liturgy the form of church government was a second controversial issue among Elizabethan English Protestants. In 1570 Thomas Cartwright (1535–1603) delivered a series of lectures at Cambridge University proposing that presbyterian government, or government by local councils of clergy and laity, might be an improvement over the current system of archbishops, bishops, and appointments. Cartwright was dismissed for his opinions and fled to Geneva. Two years later John

Thomas Cartwright

Field and Thomas Wilcox anonymously published an *Admonition to the Parliament,* which pushed Cartwright's ideas even further. In reply John Whitgift, vice-chancellor at Cambridge, maintained that the government of the church should be suited to the government of the state and that episcopal government best suited monarchy. In this dispute most Puritans shied away from extremes and supported some form of episcopacy, but a small number went beyond even Cartwright and Field in seeking to effect immediately a "reformation without tarrying for any." These Separatists broke with the established parish system to set up voluntary congregations that covenanted with God and with themselves, chose ministers by common consent, and put into practice the Puritan marks of the true church. Robert Browne (d. 1633) was an early advocate of the Separatist mentality.

The leaders of the Puritan movement, however, including Cartwright (who had returned to England in 1585) and Field, repudiated the Separatists and sought to set up "presbyterianism in episcopacy," or a "church within the church." This compromise between presbyterianism and episcopacy was preferred by the most prominent Puritans, and they began to institute such a system by means of informal public meetings of clergy and laity to expound and discuss the Bible. These meetings were called "prophesyings," and they were favoured for their educational value to the rural population by Edmund Grindal, who had succeeded Parker as archbishop of Canterbury in 1576. But the prophesyings were also the occasions for local Puritan clergy, laity, and gentry to mobilize, and they were viewed by Elizabeth, in the context of the more radical groups, as a political threat. An increasingly clear alliance between Puritans and certain factions within Parliament did not allay Elizabeth's fears.

Edmund Grindal

Thus, the Queen ordered Grindal to suppress the prophesyings. When he refused, Elizabeth effectively suspended him from the exercise of his office. This suspension further alienated Puritans. Meetings continued, often in a modified form, called *classis* or conferences, which were loosely coordinated by John Field in London. Following Grindal's death in 1583, John Whitgift, Cartwright's old opponent, advanced to Canterbury. Whitgift had no hesitance in closing down the prophesyings, but he proceeded with caution in formal prosecution of Puritans. Extended ecclesiastical hearings by the Court of High Commission, under the leadership of John Aylmer, and civil proceedings by the Star Chamber were accompanied by the imprisonment of only a few of the most prominent Puritans.

Whitgift's policy, along with the death of Field and other Puritan leaders between 1588 and 1590, effectively ended any grand plan for a continuing reformation of the English Church under Elizabeth. The generally moderate Elizabethan Puritan movement was over, and the forces of reform dispersed into various parties and programs ranging from nonseparating congregationalism (as advocated by William Ames) to open subversion of the established hierarchy as in the anonymous Marprelate Tracts (1588–89). Despite failure to promote reform in matters of church structure, the Puritan spirit continued to spread throughout the society. Protestants with Puritan sympathies controlled colleges and professorships at Oxford and Cambridge, had the ears of many leaders in the House of Commons, and worked tirelessly as preachers and pastors to continue the preaching of Protestantism in its distinctively "hot" Puritan form to the laity. (M.E.M.)

Puritanism under the Stuarts (1603–49). *Events under James I.* Puritan hopes were raised when James VI of Scotland succeeded Elizabeth as James I of England in 1603. James was known to be Calvinist in theology, and he had once signed the Negative Confession of 1581 favouring the Puritan position. In 1603 the Millenary Petition (with a claimed thousand signatures) presented Puritan grievances to the King, and in 1604 the Hampton Court Conference was held to deal with them. The petitioners were sadly in error in their estimate of the King, who had learned by personal experience to resent Presbyterian clericalism. At Hampton Court he coined the phrase, "no bishop, no king." Outmaneuvered in the conference, the Puritans were made to appear petty in their requests.

The Hampton Court Conference

As a seal upon the Hampton Court Conference James appointed Richard Bancroft to be Whitgift's successor as archbishop of Canterbury and encouraged the Convocation of 1604 to draw up the *Constitutions and Canons* against Nonconformists. Conformity in ecclesiastical matters became a pattern in areas where forms of nonconformity had survived under Elizabeth. Though a number of the clergy were deprived of their positions, others took evasive action and got by with minimal conformity. Members of Parliament supported them in their position by arguing that since the canons had not been ratified by Parliament they did not have the force of law.

Puritans remained under pressure, but men of Puritan sympathies still came close to the seat of power in James's reign. The enforced reading from pulpits of James's *Book of Sports,* dealing with recreations permissible on Sundays, in 1618, however, was a further affront to those who espoused strict observance of the sabbath, making compromise more difficult.

Increasing numbers of Separatist groups could not accept compromise, and in 1607 a congregation from Scrooby, Eng., fled to Holland and then migrated on the *Mayflower* to establish the Plymouth Colony on the shore of Cape Cod Bay in 1620.

Events under Charles I. Despite the presence of controversy, Puritan and non-Puritan Protestants under Elizabeth and James had been united by adherence to a broadly Calvinistic theology of grace. Much of Whitgift's restraint in handling Puritans, for instance, can be traced to the prevailing Calvinist consensus he shared with the Nonconformists. Even as late as 1618 the English delegation to the Synod of Dort supported the strongly Calvinistic decisions of that body. Under Charles I, however, this consensus broke down, driving yet another rift into the Church of England. Anti-Puritanism in matters of liturgy and organization became linked with anti-Calvinism in theology.

The leaders of the anti-Puritan and anti-Calvinist party, notably Richard Montagu, whose *New Gagg for an Old Goose* (1624) first linked Calvinism with the abusive term "Puritan," drew upon the development of Arminianism in Holland. Arminians stressed God's universal offer of salvation to mankind in contrast to the Calvinistic doctrine according to which God predestined a few to salvation, with the rest of humanity reprobated or damned. Early English Arminians added to this an increased reverence for the sacraments and liturgical ceremony. Richard Neile, bishop of Durham, was the first significant patron of Arminians among the hierarchy, but by the time William Laud was appointed bishop of London in 1628, he was the acknowledged leader of the anti-Puritan party. London was regarded as the stronghold of Puritanism, and a policy of thorough anti-Puritanism was begun there. Men who were not Separatists found their positions increasingly difficult to maintain.

William Laud

Laud, who became archbishop of Canterbury in 1633, was clearly a favourite of Charles. He oversaw the advance of Arminians to influential positions in the church and subtly promoted the propagation of Arminian theology. His fortunes began to turn, however, when he attempted to introduce into the Church of Scotland a liturgy comparable to the Anglican *Book of Common Prayer.* When "Laud's Liturgy" was introduced at the Church of St. Giles at Edinburgh, a riot broke out leading to a popular uprising that restored Presbyterianism in Scotland.

Charles sought to put down the Scots, but his armies were no match for the Scottish forces. In 1640 he was faced with an army of occupation in northern England demanding money as a part of its settlement. Short of funds, Charles was forced to call Parliament, without which he had been trying to rule since 1629.

Religion played perhaps the key role in the parliamentary elections, and Calvinists came to dominate the Commons. Puritans, who had been increasingly alienated from the ecclesiastical and civil hierarchy since the mid-1620s, suddenly saw an opportunity to return the Church of England to its original doctrinal system and to carry out reforms that had been held in check since the Elizabethan Settlement. Arminianism in theology, liturgy, and government was linked in the popular mind with Catholicism,

as fears of a Spanish conspiracy to undermine Protestant England became widespread. The first act of the Long Parliament, as it came to be called (1640–53), was to set aside Nov. 17, 1640, as a day of fasting and humiliation. Cornelius Burges and Stephen Marshall were appointed to preach that day to members of Parliament. Their sermons urged the nation to renew its covenant with God in order to bring about true religion through the maintenance of "an able, godly, faithful, zealous, profitable, preaching ministry in every parish church and chapel throughout England and Wales" and through the establishment of a civil magistracy that would be "ever at hand to back such a ministry."

Hundreds of similar sermons were preached on monthly fast days and on other occasions before Parliament during the next few years, urging the people to adopt true doctrine, pure worship, and the maintenance of discipline as a means to claim God's blessing so that England might become "our Jerusalem, a praise in the midst of the earth."

Civil war. In the course of his reign it had become apparent that Charles himself was the patron of Arminians and their attempt to redefine the doctrine of the Church of England. Arminians in turn favoured Charles's causes against Puritans and Parliament. This alliance held despite increasing pressure on Charles to cooperate with Parliament on economic and military matters. The resulting civil war between the forces of the King and the troops of Parliament was hardly just a religious struggle between Arminians and Calvinists, but conflict over religion played an undeniably large role in bringing about the Puritan Revolution. As Protestantism split, so did English society.

The Westminster Assembly

Fighting broke out in 1642, and after the first battles members of Parliament called together a committee of over a hundred clergymen from all over England to advise them on "the good government of the Church." This body, the Westminster Assembly of Divines, convened on July 1, 1643, and continued daily meetings for more than five years.

A majority of the Puritan clergy of England probably would still have opted for a modified episcopal church government. Parliament, however, needed Scotland's military help. It adopted the Solemn League and Covenant, which committed the Westminster Assembly to develop a church polity close to Scotland's presbyterian form. A small, determined Assembly group of "Dissenting Brethren" held out for the freedom of the congregation, or "Independency," as opposed to the power of presbytery. Others, called Erastians, wanted to limit the offenses under the power of church discipline. Because both groups had support in Parliament, the reform of church government and discipline was frustrated.

Dissent within the assembly was negligible compared with dissent outside it. Pamphlets by John Milton, Roger Williams, and others schooled in Puritanism pleaded for greater freedom of the press and of religion. Such dissent

The New Model Army

was supported in the New Model Army, a Parliamentarian army of 22,000 men organized and disciplined under Sir Thomas Fairfax (1612–71) as commander in chief and Oliver Cromwell (1599–1658), and the real power in England was passing to the military leaders who had defeated all Royalist forces. Late in 1648 the victors feared that the Westminster Assembly and Parliament would reach a compromise with the defeated Charles that would destroy their gains for Puritanism. In December 1648 Parliament was purged of members unsatisfactory to the Army, and in January 1649 King Charles was tried and executed.

The age of Cromwell (1649–60). Both Parliament and the assembly continued to sit on a "rump" basis (containing only a remnant after the purges), and Oliver Cromwell emerged as England's Lord Protector. Cromwell was a typical Puritan in that he saw the judgment and mercy of God in events. Military successes to him were definite signs of the blessing of God upon his work.

The Independent clergyman John Owen guided the religious settlement under Cromwell. He maintained that the "reformation of England shall be more glorious than of any Nation in the world, being carried on, neither by might nor power, but only by the spirit of the Lord of Hosts." Error was a problem for both Cromwell and Owen, but, as Owen expressed it, it was better for 500 errors to be scattered among individuals than for one error to have power and jurisdiction over all others.

Such was the basis for a pluralistic religious settlement in England under the Commonwealth in which parish churches were led by men of Presbyterian, Independent, Baptist, or other opinions. Jews were permitted to live in England. But it was unacceptable for such groups as Roman Catholics or Unitarians to hold religious views publicly. Cromwell was personally willing to tolerate *The Book of Common Prayer,* but his Parliament was not. Voluntary associations of churches were formed, such as the Worcestershire Association, to keep up a semblance of church order among churches and pastors of differing persuasions.

Radical reform groups

In the upheaval brought on by the wars radical groups appeared that both challenged and advanced the Puritan vision of the New Jerusalem. The Levellers (a republican and democratic political party) in the New Model Army in 1647 and 1648 interpreted the liberty that comes from the free grace of God offered to all men in Christ as having direct implications for political democracy. The Diggers (agrarian communists) in 1649 planted crops on common land, first at St. George's Hill near Kingston and later at Cobham Manor, also near Kingston, to encourage God to bring soon the day when all men would live in an unstructured community of love with a communal economy. The Fifth Monarchy Men (an extreme Puritan millennialist sect) in 1649 presented their message of no compromise with the old political structures and advocated a new structure, composed of saints joined together in congregations with ascending representative assemblies, to bring all men under the kingship of Jesus Christ. As distinct units these groups were short-lived. A more enduring group was founded by George Fox (1624–91) as the Society of Friends, or Quakers, which pushed the Puritan logic disallowing any remnants of popery to its ultimate limit with a program of no ministers, no sacraments, and no liturgy. Puritanism had never been a monolithic movement, and accession to power had brought the factions to bear. The limits of the Puritan spirit of reform showed clearly in the widespread persecution of the Quakers.

The Restoration (1660–85). After the death of Cromwell chaos threatened, and in the interest of order even some Puritans supported the restoration of Charles II. They hoped for a modified episcopal government, such as had been suggested in 1641 by the archbishop of Armagh, James Ussher (1581–1656). Such a proposal was satisfactory to many Episcopals, Presbyterians, and Independents. When some veterans of the Westminster Assembly went to Holland in 1660 to meet with Charles before he returned, the King made it clear that there would be modifications to satisfy "tender consciences."

These Puritans were outmaneuvered in their attempt to obtain a comprehensive church, however, by those who favoured the strict episcopal pattern. A new Act of Uniformity was passed on May 19, 1662, by the Cavalier Parliament. The act required reordination of many pastors, gave unconditional consent to *The Book of Common Prayer,* advocated the taking of the oath of canonical obedience, and renounced the Solemn League and Covenant. Between 1660 and when the act was enforced on Aug. 24, 1662, almost 2,000 Puritan ministers were ejected from their positions.

As a result of the Act of Uniformity, English Puritanism entered the period of the Great Persecution. The Conventicle Act of 1664 punished any person over 16 years of age for attending a religious meeting not conducted according to *The Book of Common Prayer.* The Five Mile Act of 1665 prohibited any ejected minister from living within five miles of a corporate town or any place where he had formerly served. Still, some Puritans did not give up the idea of comprehension (inclusiveness of various persuasions). There were conferences with sympathetic bishops and brief periods of indulgence for Puritans to preach, but fines and jailings set the tone. Puritanism became a form of Nonconformist Protestantism.

During the short reign of Charles's Roman Catholic brother, James II (1685–88), fear of Roman Catholic

The Glorious Revolution

tyranny united politically both establishment and Nonconformist Protestants. This new unity brought about the "Glorious Revolution" (1688), establishing William and Mary on the throne. The last attempt at comprehension failed to receive approval by either Parliament or the Convocation under the new rulers. In 1689 England's religious solution was defined by an Act of Toleration that continued the established church as episcopal but also made it possible for dissenting groups to have licensed chapels. The Puritan goal to further reform the nation as a whole was transmuted into the more individualistic spiritual concerns of Pietism or else the more secular concerns of the Age of Reason.

Puritanism in the English colonies. *Virginia.* A decade before the landing of the *Mayflower* (1620) in Massachusetts a strong Puritan influence was planted in Virginia. Leaders of the Virginia Company who settled Jamestown in 1607 saw themselves in a covenant relation to God, and they carefully read the message of their successes and failures. A typical Puritan vision was held by the Virginia settler Sir Thomas Dale. His strict application of severe laws disciplining the Jamestown community in 1611 probably saved the colony from extinction, but he also earned a reputation as a tyrant. Dale thought of himself as a labourer in the vineyard of the Lord, as a member of Israel building up a "heavenly New Jerusalem." Like Oliver Cromwell later, whom he resembled, Dale interpreted his military success as a direct sign of God's lending "a helping hand."

Puritan clergymen saw excellent opportunity for their cause in Virginia. The Reverend Alexander Whitaker, the "apostle of Virginia," wrote to his London Puritan cousin in 1614, "But I much more muse, that so few of our English ministers, that were so hot against the surplice and subscription, come hither where neither is spoken of." The church in Virginia, however, became more directly aligned with the English establishment when the settlements were made into a royal colony in 1624.

The Puritan experiment in America

Massachusetts Bay. In New England, however, the Puritans had their greatest opportunity. Between 1628 and 1640 the Massachusetts Bay Colony was developed as a covenant community. Governor John Winthrop stated the case concisely in his lay sermon on board the *Arbella* before the colonists landed,

> Thus stands the cause between God and us; we are entered into covenant with Him for this work; we have taken out a commission; the Lord hath given us leave to draw our own articles . . . Now if the Lord shall be pleased to hear us and bring us in peace to the place we desire, then hath He ratified this covenant and sealed our Commission, [and] will expect a strict performance of the articles contained in it.

Lack of performance of the articles, in this view, would bring down the wrath of God.

The pattern for church organization in the colony was determined by John Cotton, who pursued "that very Middle-way" between English Separatism and the presbyterian form of government. Unlike the Separatists he held the Church of England to be a true church, though blemished; and unlike the Presbyterians he held that there should be no ecclesiastical authority between the congregation and the Lordship of Christ. Cotton proposed that the church maintain its purity by permitting only those who could make a "declaration of their experience of a work of grace" to be members. Cotton's plan ensured that church government should be in the hands of the elect, the chosen of God.

Taking their cue from Thomas Cartwright, the Puritans of the Bay Colony fashioned the civil commonwealth according to the framework of the church. Only the elect could vote and rule in the commonwealth. The church was not itself to govern, but it was the means through which were prepared "instruments both to rule and to choose rulers." Biblical law was the primary law for the ordering of both church and state.

The colony prospered; thus it seemed evident that God was blessing Puritan performance. As a result the leadership could not take kindly to those who were publicly critical of their basic program. Hence Roger Williams in 1635 and Anne Hutchinson in 1638 were banished from the colony in spite of their ability to declare experience of the work of grace.

More troublesome than these dissenters were persons such as Mary Dyer. She and other Quakers who returned again and again after being punished and banished were finally hanged. It was difficult for the state to keep the church pure.

The Cambridge Platform and the Half-Way Covenant

In order to head off a possible new form of church government dictated from England at the time of the Westminster Assembly, churches from the four Puritan colonies of Massachusetts Bay, Plymouth, Connecticut, and New Haven met in a voluntary synod in 1648. They adopted the Cambridge Platform, in which the congregational form of church government was worked out in detail. The standard for church membership came under question when it was found that numbers of second-generation residents could not testify to the experience of grace in their lives. This resulted in the Half-Way Covenant of 1657 and 1662 that permitted baptized, moral, and orthodox persons to share in the privileges of church membership except for partaking of communion.

Late in the 17th century it was apparent to all that the ideal commonwealth was not being maintained. Ministers pointed to wars with the Indians and other problems as signs of God's judgment. Visitation by demonic powers in the form of witches was believable to people expecting the wrath of God. The Salem witchcraft trials and hangings took place in 1692 at a period of declining confidence in the old ideal.

Other colonies. Massachusetts Bay, Plymouth, Connecticut, and New Haven were variations on the main theme of realizing the Holy Commonwealth in America. Roger Williams and the other founders of Rhode Island must also be regarded as Puritans with the "one principle, that every one should have liberty to worship God according to the light of their consciences."

William Penn's "holy experiment" in Pennsylvania represented another Puritan variation, only this time under Quaker norms. When Penn came into the ownership of this vast tract of land, he saw it as a mandate from God to form an ideal commonwealth. In New Jersey, Puritans from the New Haven colony who were dissatisfied with the Half-Way Convenant sought to reestablish the pristine Puritan community at Newark. Maryland, which had been established under Roman Catholic auspices, soon had a strong Puritan majority among its settlers.

There was no colony in which the Puritan influence was not strong in one form or another. One estimate is that 85 percent of the churches in the original 13 colonies were Puritan in spirit.

The expansion of the Reformation in Europe

Lutheranism in northern Europe

By the middle of the 16th century Lutheranism was dominant in northern Europe. Württemberg, after the restoration of Duke Ulrich, adopted the reform in 1534. The outstanding Reformer was Johannes Brenz and the great centre Tübingen. Brandenburg, with Berlin as its capital, embraced the reform in 1539. In that same year ducal Saxony, until then vehemently Catholic, changed sides. Elisabeth of Braunschweig, also in that year, became a convert, but only after long turbulence did her faith prevail in the land. Very significant for the north as a whole was the stand taken by Albert of Prussia, who was a member of the Polish Diet and whose wife was Danish. He secularized the Teutonic Knights and in 1525 acknowledged himself a Lutheran. In the Scandinavian lands Denmark toyed with Lutheranism as early as the 1520s, but not until 1539 was the Danish Church established on a national basis with the king as the head and the clergy as leaders in matters of faith. Norway followed Denmark. The Diet of Västerås officially declared what had for some time been true, namely, that Sweden was an evangelical state. The outstanding Swedish Reformers were the brothers Olaus and Laurentius Petri. Finland, under Swedish rule, followed suit. The Reformer there was Mikael Agricola, called "the father of written Finnish." The Baltic states of Livonia and Estonia were officially Lutheran in 1554. Subsequently ravished by the Russians,

portions of these lands united with Sweden, Denmark, and Poland. Lutheranism survived. Toward the east, Austria under the Habsburgs could enjoy no state support for the evangelical movement, which nevertheless gained adherents. In Moravia, as noted, the Hutterites established their colonies under tolerant magnates.

Eastern Europe was a seedbed for even more radical varieties of Protestantism, because kings were weak, nobles strong, and cities few, and because religious pluralism had long existed. Poland acquired a large German Lutheran population when the Danzig area came under Polish control, and a large contingent of the Bohemian Brethren migrated to Poland when the Habsburg ruler attempted their extermination. Several of the Polish noblemen adopted their pacifism and would wear only swords made of wood. To Poland also flocked the Italian anti-Trinitarians, having been granted an asylum, perhaps merely because they were Italian, by the Italian queen of Poland, Bona Sforza. Named Socinians from their leader, Faustus Socinus, they flourished until dissipated by the Counter-Reformation. Much more extensive was the Calvinist influx not only into Poland but into the whole of eastern Europe. This variety of Protestantism appealed to those of non-German stock because it was not German and no longer markedly French, as well as because of its revolutionary temper and republican sentiments. The Compact of Warsaw in 1573 called the *Pax Dissidentium* ("The Peace of Those Who Differ") granted toleration to Roman Catholics, Lutherans, Calvinists, and Bohemian Brethren, but not to the Socinians.

In Hungary, the Turkish victory at the Battle of Mohács in 1526 brought about a division of the land into three sections, the northwest ruled by the Habsburg Ferdinand, the eastern province of Transylvania under Zápolya, and the area of Buda under the Turk. Even before this date Lutheranism had made inroads not only in the German but also in the Magyar sections. Subsequently Calvinism made even greater gains. The anti-Trinitarians found a permanent locus in Transylvania. The weakness of the government and the diversity of religion in this whole area made for a large degree of toleration.

The Reformation gained no lasting hold in Spain and Italy. In Spain the main reason for this must be found in the conflicts of the previous century when the Christians were striving to achieve political, cultural, and religious unification by converting or expelling the unbelievers, the Jews and the Moors. The Inquisition was introduced in 1482 to root out all remnants of Jewish practices among the Marranos, the Jewish converts to Christianity. The non-Christian Jews were expelled in 1492. Then Granada fell and the same process was applied to the Moriscos, the Moorish converts, and the unconverted Moors, after a century, also were expelled. Because the process had thus far been successful, the pressures were relaxed, and Spain enjoyed a decade of Erasmian liberalism in the 1520s. But with the infiltration of Lutheranism the machinery of repression again was brought into force.

The Inquisition

In Italy sectarian and heretical movements had proliferated in the late Middle Ages. But one by one they had been crushed, and the Italians may well have felt that such rebellions were futile. Furthermore, the friars preached moral rather than doctrinal reform as Luther had done. Another consideration was that the new monastic orders, the Capuchins, Theatines, and Jesuits, gained papal favour and became a mighty force in counteracting Protestant infiltration, which nevertheless did take place. Venice was a centre, with its branch house of the Lutheran banking family of Fugger, and so was Lucca. At Naples the Spanish mystic Valdés, though not a Protestant, expounded a piety of the type of the liberal Catholic reform, and some of his followers were attracted to the movements coming from beyond the Alps. Calvinism gained a hold. But the Roman Inquisition, as above noted, was established in 1542, and those with Protestant leanings either made cloisters of their own hearts, or went to the stake, or crossed the mountains into permanent exile. The most radical theological views of the Reformation were those propounded by the Spanish and Italian anti-Trinitarians.

(R.H.B./J.C.S./M.E.M.)

Protestant renewal and the rise of the denominations

THE SETTING FOR RENEWAL

Survival of a mystical tradition. The Thirty Years' War (1618–48) must be seen as one of the circumstances out of which the desire for spiritual renewal emerged. Although modern historical research has modified the exaggerated contemporary accounts of the war's effects, it is unquestioned that distress was widespread and profound. In some places the economy was reduced to barter, schools were closed, churches were burned, the sick and needy were forgotten. Not unexpectedly spiritual and moral deterioration accompanied the physical destruction. Drunkenness, sexual license, thievery, and greed were the despair of faithful pastors and earnest laymen.

During the war some notable signs of renewal began to appear. There reemerged, for example, an interest in the earlier devotional literature, some of which reflected the pious mysticism associated with such names as Johannes Tauler (*c.* 1300–61), Thomas à Kempis (*c.* 1380–1471), and other German, Dutch, and even Spanish authors. The mystical tradition had lived on into the Reformation century and found representatives in Kaspar Schwenckfeld (1489–1561), Valentin Weigel (1533–88), and Jakob Böhme (1575–1624). Although both Lutherans and Calvinists opposed these mystics, many of their religious and theological ideas were subsequently absorbed by orthodox theologians.

Renewed interest in the mystical tradition

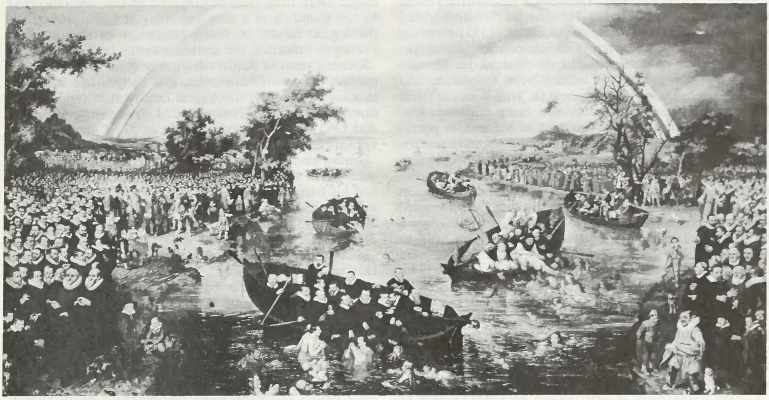
Catholic recovery of Protestant territories. After the Peace of Westphalia in 1648 that ended the war, Catholicism regained some territories from Lutheran Protestantism: first, because the rise of toleration was somewhat more rapid in Protestant countries than in Catholic lands and, second, because Louis XIV identified French power with universal French acceptance of the Roman Catholic faith. In 1685 he revoked the Edict of Nantes and expelled thousands of Huguenots, who fled to England, Holland, or Germany, much to the advantage of those countries. Several of the French refugees became prominent in English religious life, and in Prussia groups of them founded flourishing congregations known as the French Reformed. In 1702 a determined group of Huguenots in the mountains of the Cévennes in France, known as the Camisards, rose in rebellion but was suppressed by military power two years later. There was a further small outbreak of war in 1709. For a time the few surviving Huguenot congregations met only in secret. They were led by Antoine Court (1695–1760), who secured ordination from Zürich and founded (1730) a college at Lausanne to train pastors. French Protestants barely held out until the French Revolution, after which they had a revival.

Survival of French Protestantism

France gained Alsace in 1648. This enabled Catholics to increase rapidly, and Protestants decreased in strength. Strassburg, once one of the leading cities of the Protestant Reformation, returned its cathedral to the Catholics (1681) and became a town with a large Catholic population. Louis XIV ruled the Palatinate for nine years and allowed the French Catholics to share the churches with the Protestants; though he was compelled to surrender the country at the Treaty of Rijswijk (1697) to the Holy Roman Empire, a clause (the *Simultaneum*) of the treaty (added at the last moment and not recognized by the Protestants) preserved certain legal rights and endowments of Catholics in Protestant churches. As a result of France's greater power Protestant authority in the Rhineland between Switzerland and the Netherlands diminished.

Another shock to Protestantism was the conversion of Augustus II, elector of Saxony, to Roman Catholicism in 1697. It appeared as though Protestantism was not even safe in its original home. The conversion involved political motives; Augustus was a candidate for the throne of Poland and was loyal to his new allegiance, assisting the Roman Catholic Church in Poland and also, somewhat, in Saxony; but such assistance had no effect on the Lutheranism of Saxony.

Protestant scholasticism. The second half of the 17th century was at once the high age of Protestant systematic orthodoxy and the age when the first signs of its dis-



"Fishing for Souls," by Adriaen Pieterszoon van de Venne, 1614, depicting Catholics and Protestants competing for converts in the Netherlands. Protestant James I of England and Prince Maurice of Nassau are shown standing on the left bank and the Catholic archduke Albert of Austria and his wife, Isabella, on the right bank. In the Rijksmuseum, Amsterdam.
By courtesy of the Rijksmuseum, Amsterdam

solution appeared. The axioms of the Reformation were worked out in a great and systematic body of doctrine.

The theologians defended and the pastors taught Luther's or Calvin's dogmatic systems—relying also upon authoritative sources such as the Formula of Concord (1577) in Lutheranism or the conclusions of the Synod of Dort (1618) in Calvinism—which were extended and made into a tradition. Even when the system was not of the ordinary Protestant tradition, it was generally worked out in many volumes, based upon coherent axioms, defended against all assailants, appealing always to reason and to biblical authority and seldom to feeling or conscience. This age has sometimes been known as the age of Protestant scholasticism. But that pejorative term came from a posterity that would no longer accept the axioms on which the systems were founded. These were the last scriptural theologians before the period of the Enlightenment, when the understanding of Scripture was altered. The old axioms were changed by Pietism, science, and philosophy.

THE RISE OF PIETISM

Influences from English Puritanism reached the Continent through the translation of works by Richard Baxter (1615–91), Lewis Bayly (1565–1631), and John Bunyan (1628–88). Most frequently read were Baxter's *A Call to the Unconverted,* Bayly's *The Practice of Piety,* and Bunyan's *Pilgrim's Progress.*

Dutch Pietism, influenced by the Englishman William Ames (1576–1633) whose *Medulla Sacrae Theologiae* (1623) and *De Conscientia* (1630) were basic textbooks for "federal theology" and Puritan casuistry in England and New England, was represented by Willem Teellinck, Johannes Cocceius, Gisbert Voetius, and Jodoeus van Lodensteyn. Impulses from these men became a part of the reform movement that had already appeared in German Lutheran circles and was to be known as "Reform Orthodoxy." Older historians of Pietism, notably Albrecht Ritschl, paid little or no attention to this reform phenomenon within Lutheranism. Ritschl saw Pietism as an alien mysticism uncongenial to the spirit of both Luther and the 17th-century theologians. More recent scholars (E. Benz, M. Schmidt, H. Leube, F.W. Kantzenbach) have exposed the Ritschlian prejudice and deepened the understanding of the role played by such representatives of "Reform Orthodoxy" as Johann Arndt (1555–1621) and Johann Dannhauer (1603–66). The "pectoral [heart] theology" of these orthodox Lutherans found its highest expression and widest audience in the writings of Arndt, who, rather than Philipp Jakob Spener, can be called the "father of Pietism." His chief work, *Four Books on True Christianity* (1606–10), was soon being read in countless homes. Although Arndt developed devotionally the *unio mystica* (mystical union), a 17th-century Lutheran doctrinal addition to the *ordo salutis* (order of salvation), the central Arndtian theme was not that of mystical union. Rather, he stressed repentance, regeneration, and the new life, and this was the very essence of Pietism.

Alongside the orthodox piety of the 17th century one of the most significant contributions to spiritual renewal was the rich treasures of Lutheran hymnody. Examples from this classical period of church song are the works of Philipp Nicolai (1556–1608; "Wake, Awake" and "How Brightly Beams the Morning Star!"), Paul Gerhardt (1607–76; "O Sacred Head Now Wounded," "O How Shall I Receive Thee," "Put Thou Thy Trust in God"); and Martin Rinkart (1586–1649; "Now Thank We All Our God").

Pietism in the 17th century. The various streams of concern for renewal converged in the life and work of Spener (1635–1705). In 1666, after having earned his theological doctorate at Strasbourg, he was called to be Senior of the clergy in Frankfurt am Main, where he was soon distressed by the conspicuous worldliness of the city. His sermons urged repentance and renewal, and each Sunday afternoon he held catechism classes for both children and adults. This led to efforts to revitalize the rite of confirmation, which, since the days of Martin Bucer, had been practiced in the Church of Hesse (Frankfurt).

Philipp Jakob Spener and the *collegia pietatis*

The origin of the so-called *collegia pietatis* (assembly of piety) has been traced to a sermon of 1669, in which Spener exhorted the laity to come together on Sunday afternoon not to drink, play cards, or gamble, as was the custom among Frankfurt's smart set, but to review the morning's sermon and to engage in devotional reading and conversation "about the divine mysteries." The next year, at the request of a few parishioners, such meetings were held each Sunday and Wednesday at Spener's home. Although some of the Frankfurt ministers, over whom Spener was superintendent, took a dim view of the *collegia pietatis,* the practice flourished and in time became a distinguishing feature of the movement. Those who attended the conventicles were soon called Pietists.

In a relatively short time, Spener became a household name in Germany. Through his writings and extensive correspondence, especially with men in high places, Spener came to be called "the spiritual counselor of all Germany." Most significant was the publication in 1675 of his *Pia*

Desideria (*Pious Desires*). The book's first part reviewed the low estate of the church. He charged civil authorities, who since before the Peace of Augsburg (1555) were the de jure heads of the church, with irresponsible caesaropapism (doctrine of state control over church). He likewise flayed the clergy, many of whom were scandalous and self-seeking, often confusing assent to "true doctrine" with faith. The laymen, too, he claimed, were not blameless. Drunkenness must not be excused as a German peccadillo; prostitution, adultery, fornication, homosexuality, thievery, and assault must be rooted out lest people lose God's promised salvation. The second part of the work reminded readers of the possibility of better conditions in the church: ". . . we can have no doubt that God promised His church here on earth a better state than this." When the full number of heathen (Gentiles) had been brought in, God would even convert the Jews. But the fulfillment of these hopes was not to be achieved by sitting with folded hands. Part three, therefore, set forth a six-point reform program:

The six-point reform program

1. The Word of God—the whole Bible, not merely the pericopes (biblical texts used in a set sequence in worship services)—must be made known widely through public and private reading, group study (conventicles under the guidance of pastors), and family devotions.

2. There should be a reactivation of Luther's idea of the priesthood of believers, which included not only the "rights of the laity" but also responsibility toward one's fellow men.

3. People should be taught that Christianity consists not only in knowing God's will but also in doing it, especially by implementing the command to love one's neighbour.

4. Religious controversies with unbelievers and heretics unfortunately may be necessary. If they cannot be avoided, they should be entered prayerfully and with love for those in error.

5. Theological education must be reformed. Professors must see that future pastors are not only theologically learned but spiritually committed.

6. Finally, preaching should have edification and the cultivation of inner piety as its goal.

Initially the *Pia Desideria* was received with enthusiasm and given wide acclaim. Some clergymen, however, felt threatened by the implications of the reform program's emphasis on the laity. Professors resented Spener's criticism of scholastic theology and advocacy of curricular reform. Spener's response was to emphasize more and more the *collegia pietatis*. Contrary to Spener's wishes the conventicles in time became divisive and abrasively donatistic (referring to the Early Church heresy that held that priests must be morally righteous or the sacraments would not be valid), tending to develop into "little churches within the church" (*ecclesiolae in ecclesia*). In an attempt to stem separatism and other questionable attitudes, Spener wrote tracts that expounded the doctrines of the spiritual priesthood (1677) and ecclesiology (1684). In the latter he argued that despite the faults of the church its teachings were not false and separation from services and sacraments was wrong.

Spener's influence had spread widely by 1686. In many circles, not least among the nobility, he was praised and imitated. In other quarters his emphases produced vigorous and, in many instances, unjust criticism. Weary of opposition and controversies, Spener accepted a call to be the court chaplain in Dresden, where he was soon disillusioned by the unresponsiveness and vulgarity of the court and the hostility of the pastors in this stronghold of orthodoxy. Two items of special significance from the Dresden period should be noted: (1) There he wrote his *Impediments to Theological Study* (1690), which was hardly calculated to win friends at the famous Saxon University of Leipzig; (2) there, too, he made the acquaintance of a young instructor, August Hermann Francke (1663–1727), who was to become in a sense Spener's successor and the second great leader of Pietism.

Court chaplain in Dresden

By 1691 Spener welcomed a call from the elector of Brandenburg, who soon brought in other Pietists, opened his domain to persecuted French Huguenots, and made Berlin a strong spiritual centre, thus taking religious leadership away from rival Saxony. All of this was enhanced by the founding of a new university at Halle (1694), the theological faculty of which became, with Spener's and Francke's influence, the academic centre of Pietism.

Spener's years in Berlin were not without bitterness. The conflict between Orthodoxists and Pietists had mounted to a high pitch. The theological faculty at Wittenberg, for example, charged Spener with 284 deviations and prayed that God would save "our Lutheran Zion" from the ravages of pietistic heresies.

During his last years Spener collected and edited several volumes of his papers (*Theologische Bedencken*), continued his friendship with and support of Francke at Halle, and, significantly, served as a sponsor at the baptism of Nikolaus von Zinzendorf, who was to lead evangelical Pietism in a new direction. Spener died on Feb. 5, 1705.

Meanwhile, Francke became the central figure of Pietism. During his student years at Leipzig he had been engaged in group Bible study, being one of the organizers of a *collegium philobiblicum* (assembly of Bible lovers), dedicated largely to the scholarly rather than devotional approach to the Scriptures. A religious experience in 1687 led Francke to make conversion—characterized by a severe penitential struggle and commitment to holy living—the norm for distinguishing the true Christians from unbelievers. Francke's Pietism, going beyond the spirit of Spener, came to stress a legalistic and ascetic way of life. Under Francke's leadership (he became professor in 1698) Halle became famous not only for its university but for the many "Halle institutions" that sprang up: an orphan asylum with affiliated schools, a publishing house and Bible institute, a Collegium Orientale Theologicum (Oriental College of Theology) for linguistic training of missionaries, and an infirmary that the medical faculty welcomed as compensation for the university's lack of a clinic. All of this gave to Halle and Franckean Pietism an energetic and activist character.

Pietism in the 18th century. *Central Europe and England.* One of Francke's institutions in Halle was the *paedagogium* (1698), which was intended for the education of boys whose well-to-do parents lived at a distance. Nikolaus Ludwig, Graf von Zinzendorf (1700–60) attended the Halle boarding school from 1710 to 1716. Having been drawn earlier to Spener, his godfather, Zinzendorf was now greatly stimulated by Francke. As a 14-year-old lad he organized the "Order of the Grain of Mustard Seed," whose youthful members pledged themselves to reach out in ever-expanding love to "the whole human race."

By 1721 Zinzendorf had settled down on his estate (Berthelsdorf) near the Bohemian border, where he brought believers together in a nonseparatist *ecclesiola in ecclesia*, which denied the Halle Pietists' demand for penitential remorse as a mark of "heart religion." Zinzendorf formulated the slogan that came to play such a great role in the history of revivals: "Come as you are. It is only necessary to believe in the atonement of Christ."

A small band of Moravian exiles took refuge on his estate in 1722. Looking upon this event as an opportunity to realize his cherished project of "the Mustard Seed," he gave up his position in the Saxon civil service and welcomed other Moravian refugees. They, like Zinzendorf, had been primarily influenced by Pietism and had only a hazy idea that their ancestors were Hussites. Zinzendorf soon organized the colony, now called Herrnhut, into the community of the Bohemian Brethren. They were not to separate from the Lutheran Church of Saxony. They would attend services in the village church at Berthelsdorf and call upon the local pastor for ministerial acts; but they were to look upon themselves as "the salt" of the earth, an *ecclesiola* from which "heart religion" would be disseminated throughout Christendom. Under Zinzendorf's "superintendency" the Herrnhut Brethren became more and more a distinct church, the reborn Moravian Church, or Unitas Fratrum. Although Zinzendorf received a license as a minister in 1734 and three years later was consecrated bishop, he left Herrnhut under pressure in 1736, traveling in western Germany, England, and America. The chief centres of his missionary work in Pennsylvania were Germantown and Bethlehem. He returned to Herrnhut in

Herrnhut

1749 and presided over the Church of the Brethren until his death (1760).

The influence of the Moravians on the Evangelical Awakening in England was significant. By 1775 there were 15 Moravian congregations in England, and it was in one of these that John Wesley, founder of Methodism, had his famous "Aldersgate Street Experience" (1738) as he was listening to a Moravian preacher reading Luther's *Preface to the Romans:*

> while he was describing the change which God works in the heart through faith in Christ, I felt my heart strangely warmed. I felt that I did trust in Christ . . .; and an assurance was given me that he had taken away *my* sins."

He allied himself with the Moravian society in Fetter Lane, London, and the same year journeyed to Hernnhut to learn at first hand about the people to whom he owed so much. Although Wesley later parted from the Moravians, his initial experience of saving grace in the company of the Brethren shaped the wide-reaching evangelical movement that associated the names of the two Wesleys (John and Charles) and George Whitefield.

Germany. A slightly different type of Pietism appeared in Württemberg, where Spener had established relations with Swabian churchmen. Avoiding the extremes of Franckean Pietism, it accepted conventicles but opposed all temptation to separatism and sought an evangelical, as opposed to legalistic, sanctification of life in the congregations. Interested in academic theology and a scholarly study of the Scriptures, the leader of Württemberg Pietism, Johann Albrecht Bengel (1687–1752), was a pioneer in textual criticism and biblical, in contradistinction to systematic, theology. His *Gnomon Novi Testamenti* (1752; "Interpretation of the New Testament") was widely distributed in the Lutheran and English world; a fresh approach to the Bible by its emphasis on *Heilsgeschichte* (the history of salvation).

Radical Pietism

Radical Pietism always lurked beneath the surface of Evangelical Pietism. The appeal to mystical and emotional experiences and the depiction of the church as "unholy Babylon" were common characteristics of Radical Pietism. Difficult to trace historically because of a tendency to flare up spontaneously, it can nevertheless be divided into two main forms. The first was a fanatic sectarianism in which ecstatic and visionary elements were dominant. A favourite doctrine was chiliasm (referring to the thousand-year reign of Christ at the end of history), in which the *apocatastasis* (the eventual salvation of all men) played a large role. Somewhat different but still under the first rubric were the "inspired congregations," whose inspiration was expressed in convulsive physical phenomena accompanied with glossolalia, "speaking in tongues." The second main form was "separatistic" or "nonchurch" and emphasized the "inner light." Because the "inner light" and human reason were often identified, the advocates of "Spiritual Pietism" tended to move toward Rationalism (see below). Chief among these men were Gottfried Arnold (1666–1714), Johann Konrad Dippel (1673–1734), and Gerhard Tersteegen (1697–1769).

18th-century Pietism in Scandinavia, Russia, and America. *Denmark/Norway.* As in Germany, the age of orthodoxy in the Dano-Norwegian kingdom had its deeply spiritual side, which came to expression in men like Bishop Jens Dinesen Jersin (died 1632) and Holger Rosenkrantz (died 1642), both of whom taught the necessity of pious living. Also, as in Germany, the "reform orthodoxy" was evidenced in hymns, especially those of Thomas Kingo (1634–1703). Pietism, as such, arrived in Copenhagen at the turn of the century and was welcomed, strangely enough, by the unpietistic king Frederick IV. It was during his reign (1699–1730) that the royal chaplain, the German R.J. Lütkens, was able to give status to pietistic pastors and to win the King for the cause of missions in India. The King initiated a search for missionaries and, finding none in his domain, he turned to Germany, where Lütken's contacts brought about the connection with two young Halle-trained Pietists, Bartholomäus Ziegenbalg (1683–1719) and Heinrich Plütschau (1678–1747). Ordained at Copenhagen in 1705, these men became the founders of the famous Tamil mission at Tranquebar, India, and stimulated foreign mission interest among the Halle Pietists. To this period belongs the Christian work among the semipagan Lapps in northern Norway carried on by the Norwegian Pietist Thomas von Westen. Another Norwegian, Hans Egede, became the pioneer missionary in Greenland. King Christian VI, known as the "Pietist on the throne," gave support to numerous pietistic causes: an orphan home and schools modeled after Halle, a missionary institute, and even conventicles (the 1741 decree permitted them only under pastoral leadership). The name of Erik Pontoppidan, court preacher at Copenhagen and later bishop of Bergen in Norway, was to have enduring significance largely because of his excellent exposition of Luther's catechism, entitled *Truth unto Godliness.* Virtually a national reader for many generations, especially in Norway, this "layman's dogmatics" combined Law and Gospel, orthodoxy and Pietism, in such a manner that its power persisted into 20th-century American Lutheranism.

Pietistic missionary activity

Sweden, Finland, and Russia. Original royal opposition to Pietism in Sweden was softened only after Francke personally visited King Charles XII on his Russian campaign. Meanwhile, Swedish students at Halle returned to their homeland imbued with Francke's ideas and practices. Following the defeat of Charles XII at Poltava in Russia (1709), thousands of Swedish prisoners of war were quartered in Siberia. Many sought comfort in religion under the leadership of a Swedish Pietist, J. Cederhielm. Correspondence with Halle and the writings of Francke and Arndt produced a strong pietistic movement in the prison camps from which only 5,000 of the original 30,000 captives were able to return to Sweden by 1724. The zealous returnees carried their pietistic convictions back to Swedish parishes. In a short time both church and government looked upon Pietism as a threat to national unity. The result was the Conventicle Act of 1726, which retarded Pietism and held Swedish church life to conventional forms for the next century. Finns as well as Swedes had followed Charles XII to defeat. Those who returned from Russia were the apostles of a religious awakening. For a time the literature of Pietism was influential, but due to the Conventicle Act of 1726 (Finland was partially a Swedish domain), its role was somewhat limited.

Pietism in Siberian prison camps

Meanwhile, Pietism came to the Russian-occupied Baltic states, where it experienced greater freedom than under the Swedes. From the foreign quarter of Moscow, inhabited mainly by German Lutherans, the work of Francke reached Peter the Great and some of his government ministers.

America. In 1703 three pastors from New Sweden on the Delaware River ordained Justus Falckner, a Halle-educated Pietist, for service among the Dutch Lutherans in New York. Most of the Dutch Lutherans were of Pietist orientation, as were the many Germans from the Rhineland and Southern German valleys. These "Palatines," who settled in New York and Pennsylvania, and the famous refugee Salzburgers, who settled in Georgia, came via London where the Pietist court chaplain M. Ziegenhagen assisted them on their way to America. Accompanying the Salzburgers were two Francke-selected pastors, J.M. Boltzius and I.C. Gronau, who naturally shaped the spiritual life of the Georgia settlement. Zinzendorf's visit to America (1741–42) led to a clash between his type of Pietism and that of Halle, represented by Henry Melchior Mühlenberg (1711–87). The victory belonged to Mühlenberg, who became the organizing genius and spiritual leader, later called "The Patriarch of American Lutheranism."

RATIONALISM

The first signs of a Rationalist movement, which was to have as powerful an influence on Protestantism as the Pietists had had, may be traced back to those few who at the end of the 16th century attacked Calvinism on grounds of reason. In Leyden, the Netherlands, Jacobus Arminius (1560–1609) reacted against Calvinist doctrines of predestination (God's foreordaining men to heaven or hell). Though anyone not a Calvinist after a time came to be called Arminian, there were groups so designated in Holland and England that had members who were more marked by their use of reason in theology than by their

opposition to Calvin. In England the enemies of such liberal theologians gave them the name Latitudinarians. The so-called Latitudinarians sought to maintain church unity based upon a few fundamental articles of faith and otherwise to allow for a wide diversity of doctrine, polity, and ways of worship. Their best representatives were the Cambridge Platonists—philosophical theologians at Cambridge (*c.* 1640–80)—who claimed that reason is the reflection of the divine mind in the soul.

During the 17th century philosophy, hitherto considered a handmaid to theology, was expanded beyond the limits of Aristotelian philosophy and the Bible and—partly due to natural science and partly due to the reflections of thinkers from Francis Bacon (1561–1626) and René Descartes (1596–1650) onward—developed its independence. The successes of science, especially to be noted in the work of Sir Isaac Newton (1642–1727), persuaded many men of the power of reason and, by 1680, of the necessity that all things be tested by reason, including even those realms of the conscience or spirit that hitherto had been thought inaccessible to reason. The signs of the age of Rationalism were the rapid decline of belief in witchcraft; the slow and painful rise of a belief in toleration; a more widespread symbolic comprehension of conceptions like heaven and hell; and the recognition of the small size of the planet Earth within the universe. On the Continent Benedict de Spinoza (1632–77) and G.W. Leibniz (1646–1716), and in England John Locke (1632–1704), were regarded as the philosophers of the age. Among the German theologians Christian Wolff (1679–1754) of Halle approached theology almost as if it were a form of mathematics, seeking for a truth that would be incontrovertible among all reasonable men. Under prompting from Pietists of Halle, he was expelled from Prussia in 1723. But before Wolff's death Rationalist theologians had displaced the Pietists in control of Halle University and had made it the centre of Rationalist theology among Protestants.

Effects of Rationalism

In England the same trend among the disciples of John Locke issued in the Deists (especially John Toland, 1670–1722) for whom Christianity was never mysterious and was understood only as a republication of the natural religion of the human race. Like Wolff and his disciples the English Deists had no permanent influence on the history of Protestantism, except by forcing the theologians to answer them and thereby to treat the philosophy of religion with seriousness. The most important of all the answers to the Deists lay in the work of Bishop Joseph Butler (1692–1752), whose sermons and *Analogy of Religion* formed the most cogent defense of the basis of Christian philosophy known in that age.

Rationalist theology, contemporaneous though certainly not in harmony with Pietism and evangelicalism, began to modify or even destroy the traditional orthodoxies—*i.e.*, Lutheran or Calvinist—of the later Reformation. The Rationalist theologians insisted that goodness in God could not be different in kind from goodness in men and therefore that God cannot do what in a man would be immoral. Though for the most part they accepted the miracles of the New Testament—until toward the end of the 18th century—the Rationalists were critical of miracles outside the New Testament, since they suspected everything that did not fit their mechanistic view of the universe.

EVANGELICALISM IN ENGLAND AND THE COLONIES

Methodism. Similar to the Pietists in Germany was the evangelical, or Methodist (named from the use of methodical study and devotion), movement in England led by John Wesley. While a fellow of Lincoln College, Oxford, Wesley gathered a group of earnest students of the Bible about him, made a missionary expedition to Georgia, and became a friend of the Moravians. Like the Pietists he laid much emphasis upon the necessity of conversion and devoted the remainder of his life to evangelistic preaching in England. He did not intend any separation, but the parish system of the Church of England as then organized was incapable of adjustment to his plan of free evangelism and lay preachers. In 1744 Wesley held the first conference of his preachers; soon this became an annual conference, the governing body of the Methodist societies, and was given a legal constitution in 1784. The Methodist movement had remarkable success, especially where the Church of England was failing—in the industrial parishes, in the deep countryside, in little hamlets, and in hilly country, such as Wales, Cumberland, Yorkshire, and Cornwall. In 1768 Methodist emigrants in the American colonies opened a chapel in New York, and thereafter the movement spread rapidly in the United States. It also succeeded in French-speaking cantons of Switzerland.

The role of John Wesley

The Methodist movement seized upon the elements of feeling and conscience that Protestant orthodoxy had tended to neglect. It gave a renewed and devotional impetus to the doctrines of grace and justification and to the tradition of moral earnestness, which had once appeared in Puritanism but which had temporarily faded during the reaction against Puritanism in the middle and late 17th century. In England it slowly began to strengthen the tradition of free churchmanship over against the tradition of the established church, though for a century or more many English Methodists believed themselves to be much nearer the Anglican Church from which they had issued than any other body of English Protestants. It enabled hymns—hitherto confined (except for metrical Psalms) to the Lutheran churches—slowly to be accepted in other Protestants bodies, such as the Church of England, the Congregationalists, and the Baptists. The evangelical movement of the 18th century produced several of the most eminent of Christian hymn writers, especially Philip Doddridge (1702–51) and Charles Wesley (1707–88).

Though John Wesley himself had not been Calvinist, in Wales the Methodists retained both the name and the theology of Calvinistic Methodists. In the United States Methodism made even more rapid progress.

The Great Awakening. Churches in the 13 colonies of the American states practiced the Congregational or Baptist church polity on a scale not known in Europe. The small Anabaptist groups had required evidence of faith, and this sometimes meant public testimony to the experience of conversion. In the larger congregations of America a similar testimony—because it was given to a wider circle—became more evident, more solemn, and at times more emotional. The pastors of the Calvinistic tradition of New England, trying to escape from the religion of forms and to seek the religion of the heart, gave unusual stress to the necessity for an immediate experience of salvation. Pastors found that under certain conditions a wave of emotion could sweep through an entire congregation and believed that they could here observe conversion and its subsequent issue in a better life. The movement owed something to the German Pietist T.J. Frelinghuysen (1691–*c.* 1748) and something to John Wesley's colleague George Whitefield (1714–70). The chief mind at the beginning of the Great Awakening, however, was that of an intellectual mystic rather than of a conventional Calvinist preacher. Jonathan Edwards (1703–58) was the Congregational pastor at Northampton in Massachusetts, where the conversions began in 1734–35. In the middle years of the 18th century waves of revivals and conversions spread through the colonies. Though the revivals were led by Congregationalists and Presbyterians, many small, independent, Bible-centred groups, which often professed allegiance to Baptist teaching, came into being because of the revivals. As Wesley in England and Zinzendorf in Germany had been forced to carry their new methods outside the established churches of their lands, so too were the American revivalistic leaders.

Jonathan Edwards

The movement was not native to America. But the conditions of the American frontier gave this kind of evangelicalism a new vigour, and from America it permanently influenced the future development of Protestantism. In the towns and new cities with moving populations, Protestantism found methods that became a feature of evangelical endeavours to reach the unregenerate or the unchurched crowds of the coming industrial cities.

LEGACIES OF THE AMERICAN AND FRENCH REVOLUTIONS

The American Revolution and the French Revolution changed the history of Western society and within it the history of the Protestant movement. The American Con-

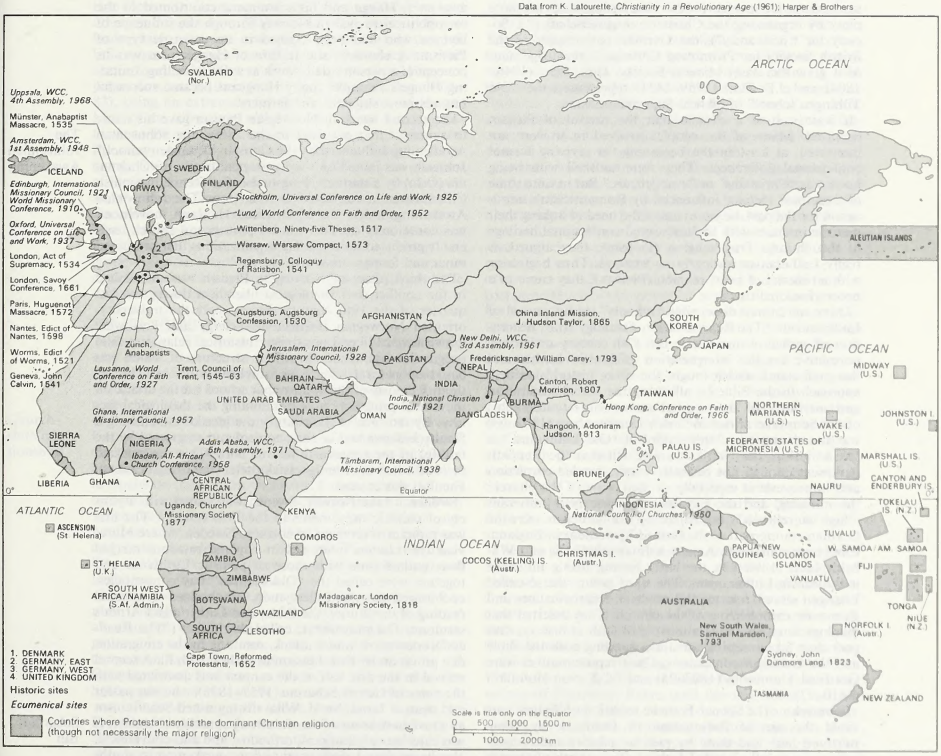
stitution, with its inferred separation of state and churches, owed something to the spirit of free churchmanship that had been inherited from colonial days, something to the religious mixture of immigrants continually arriving from Europe, something to the reaction against the "Church and King" alliance that prevailed in Britain, and something to the secular spirit of the Enlightenment. With the French Revolution and Napoleon, the idea of the secular state became an ideal for many European liberals, especially among the anticlericals in Roman Catholic countries. The American pattern was probably more influential than the Napoleonic in Protestant Europe. The Protestant states of Germany, Scandinavia, the Netherlands, Switzerland, England, and Scotland, which were all accustomed to established Protestant churches, for a time met no strong demand anywhere for disestablishment. In all those countries the members of the free, or dissenting, churches were able to secure complete toleration and civil rights during the 19th century, but in no Protestant country was the formal link between state and an established church totally broken during the 19th century, except in Ireland (1871) and in Wales (1914–19), where the Church of England was a minority. At least as an outward and historical form, however, established churches remained in England, Scotland, and all the Scandinavian countries.

MOVEMENTS TOWARD REUNION

Early in the 19th century the greatest acts leading to reunion since the Reformation were initiated. During the later 17th century the states of Europe—especially as they allowed more than one denomination—moved slowly toward toleration for all men as long as they were good citizens. The Christian leaders, especially of the new Rational, or Latitudinarian, school, sought to show that the doctrines that divided Protestants from each other (if not Protestants from Catholics) mattered less than the truths upon which they agreed. Among the Lutheran and Reformed, the German theologian George Calixtus (1586–1656) already had sought to prove their essential unity by showing that the doctrines that divided them were not essential to faith. A Scotsman, John Durie (1596–1680), traveled from England to eastern Germany and from Sweden to Switzerland on practical endeavours to persuade churchmen to unite. In 1631 the Huguenot Synod of Charenton (France) agreed to accept Lutherans who married Reformed or were godparents, without compelling them to abandon their special beliefs, on the ground that there was a sufficient agreement in the essential gospel between the Lutheran and Reformed. Lutherans (except for Calixtus and his school) could not take this view. Neither Calixtus nor Durie had much influence. Leibniz and the French Roman Catholic bishop Jacques Bénigne Bossuet (1627–1704) corresponded about the possibility of union between Catholics and Protestants, but in vain. In Prussia, with a mainly Lutheran population and a dynasty of Reformed princes, the policy of reconciliation became more effective. In 1708 King Frederick I built a "union-church" in Berlin, with the Lutheran Catechism and the Heidelberg

Pre-19th-century efforts at reunion

Data from K. Latourette, *Christianity in a Revolutionary Age* (1961); Harper & Brothers



Significant sites, missionary expansion, important events (with associated persons), and relative numerical strength of Protestantism in the Eastern Hemisphere.

Catechism side by side on the altar. In 1817, in a Prussia stimulated by the national revival that followed the fall of Napoleon (1815), King Frederick William III (1770–1840) used the third centenary of the Reformation to unite the Lutheran and Reformed of Prussia by royal decree (the Prussian Union), and despite resistance the union was slowly accepted by the majority of Prussian congregations. Other, though not all, German states succeeded in uniting their Protestant communities about the same time. Many of the more conservative Lutherans, rejecting the Prussian Union, emigrated to the United States.

The Prussian Union

THE REVIVAL OF PIETISM

Germany. Before and for some time after 1815 an awakening occurred in Germany as a reaction against the Enlightenment. In philosophy, literature, and music it found expression in German Idealism and Romanticism. In the congregations the reaction took the form of Pietism. Pietistic orthodoxism and biblicism continued to live on among "the quiet in the land." Some solitary thinkers with pronounced religious interests sought to preserve and awaken genuine Christianity and to point out the banality of the Enlightenment. Among these was Johann Georg Hamann (1730–88), a theologian given to brilliant paradoxical thought, who understood Luther's *theologia crucis* (theology of the cross) better than any other 18th-century person. Matthias Claudius (1740–1815) was another representative of the antirationalist mood of the dawn of the 19th century. Johann Friedrich Oberlin (1740–1826) mixed his biblicistic piety with a concern for social missions. J.A. Urlsperger (1728–1806) sought to promote piety by organizing the Christentumsgesellschaft ("A Society for Christianity"), the German counterpart of the British Society for Promoting Christian Knowledge. Out of it grew the Basel Mission Society. G.C. Storr (1746–1804) and J.F. Flatt (1759–1821) represented the "Old Tübingen school" of biblical Supernaturalism.

It was in such a climate that the revival of Pietism occurred. Many of the people involved in it were not interested, at least in the beginning, in reviving former confessional differences. They were satisfied with being known as "Christians" or "evangelicals." But in time some of these new Pietists, influenced by Romanticism's admiration for the past, began to assert the need of linking their pietistic interests with the traditional confessional heritage of the church. True religion (Pietism), they argued, is really Lutheranism properly understood. Thus beginning with a renewal of heart religion (Pietism), they came to a neoconfessionalism.

There were three discernible "schools" in this revival of Lutheranism. "The Repristination Theology" (*i.e.*, restoration of earlier norms) made the 17th-century orthodoxy normative for the interpretation of Martin Luther and the confessions, and it fought the rising historical-critical approach to the Bible by affirming the verbal inspiration and inerrancy of the original manuscripts (autographs) of the Scriptures. Ernst Wilhelm Hengstenberg (1802–69) was the champion of this school of "Old Lutherans." A second group, the Neo-Lutherans, felt that the Repristinationists, though not basically wrong, needed correction and improvement especially in their view of the church, the ministry, and the sacraments. These Neo-Lutherans ("high churchmen"), influenced by Romanticism, were the German counterpart of the Oxford Movement in England. Chief exponents were August Vilmar (1800–68) and Wilhelm Löhe (1808–72), the latter having strong influence in American Lutheranism. The third group, the so-called Erlangen school, rejected Rationalism, Repristination, and Romantic catholicizing of the church. They asserted that theology must see the relationship of faith to history, thus providing a new setting for understanding both the Bible and the Lutheran confessions. Chief representatives were Gottfried Thomasius (1802–75) and J.C.K. von Hofmann (1810–77).

Denmark. The Spener-Francke tradition of Pietism survived the age of Rationalism in Denmark by being nurtured here and there by pietistic pastors and congregations, especially in rural Jutland. The rebirth of Danish spiritual life and the conquest of Rationalism in the first half of the 19th century, however, came not from Pietism but from the religious and cultural impact of N.F.S. Grundtvig (1783–1872) and Søren Kierkegaard (1813–55). Both of these men were profoundly religious and at times may have sounded like Pietists, but neither had any essential sympathy for Pietism. Grundtvig was in fact definitely opposed to Pietism, while Kierkegaard, though stressing "the individual" and his existential involvement in the truth, found little time for Pietism as such. The actual renaissance of Pietism in Denmark was associated with the Inner Mission Society (established in the 1850s) and its leader Vilhelm Beck (1829–1901), who, deeply influenced by Kierkegaard's *Øieblikket* ("The Present Moment"), brought some of his emphases into the church that Kierkegaard so bitterly criticized.

Grundtvig and Kierkegaard

Norway. Nineteenth-century Pietism in Norway may be seen in three episodes: (1) the life and work of Hans Nielsen Hauge (1771–1824), (2) the pietistic confessionalism of Gisle Johnson (1822–94), and (3) the conflict over liberal theology (*c.* 1875–1908). Hauge was a layman theologically untrained but at home in the Bible, Luther's catechism, and the works of Arndt, Pontoppidan, and Kingo. Converted in 1796, his sense of mission eventually produced a national revival. Hauge and Haugeanism, though sharply critical of the established church, became an example of conventicle Christianity within the framework of the state church. Arrested no fewer than 10 times for violation of the long-neglected Conventicle Act of 1741, Hauge's final imprisonment lasted from 1804 to 1811. Although he thought of himself solely as a religious awakener, Hauge and his movement contributed to the sociopolitical revival in Norway through the influence of laymen who had been trained in an activistic type of Pietism. A characteristic feature of Haugeanism was its concept of a person's daily work as a divine calling. Imitating Hauge's example, many Haugeans became successful businessmen, shippers, and farmers.

The second figure in Norwegian Pietism gave his name to a revival that occurred in the 1850s, the Johnsonian Awakening. Influenced by the German "Erlangen school," Johnson was joined on the theological faculty in Christiania (Oslo) by a staunch Hengstenbergian Repristinationist, C.P. Caspari, a converted German Jew. The Johnsonian Awakening, unlike the lay-oriented Haugean movement, was consciously directed toward pastors and church leaders. It produced powerful lay organizations that promoted inner and foreign missions.

The Johnsonian Awakening

The third phase of Norwegian Pietism was manifested in the conflict over theological liberalism during the last quarter of the 19th century. Increasingly the university-oriented Norwegian intellectuals—clergy and lay—were drawn toward liberal positivism, historical relativism, and progressive optimism, the whole structure of which was based on natural science and biblical criticism. The orthodox Pietists of the Johnsonian school led the attack on the liberal professors now dominating the theological faculty. By the turn of the century the idea of establishing a faculty independent of state control and supported by the faithful in the congregations was born. This was realized in 1908 when the Menighetsfakultetet (the Congregational Faculty) was created.

Sweden. Like Norway, Sweden was visited with a variety of pietistic movements in the 19th century. The first was militant revivalism in northern Sweden, where Moravian Herrnhuters interested in Lapland missions merged their enthusiasms with those of pietistic Lutherans and together were called the "Old Pietists." Lay conventicles, encouraged by some clergymen, emphasized devotional reading of the liturgy, the Bible, and Luther's and Arndt's sermons. The movement, called the Läsare ("The Readers"), soon came under attack, resulting in the emigration of a group under Erik Jansson to Bishop Hill, Ill. A second revival in the first half of the century was associated with the name of Henrik Schartau (1757–1825), who was pastor and dean at Lund, Swed. What distinguished Schartauism as a revival movement was its strong churchly character. It was catechetical, liturgical, orthodox, and anti-conventicle. Yet its profound piety produced an awakening in southwest Sweden, the results of which were still noticeable in

Schartauism

the 20th century. The third revival occurred toward the middle of the century under the leadership of Carl Olof Rosenius (1816–68), a lay preacher strongly influenced by George Scott, an English Methodist evangelist. Rosenian Pietism, or the "New Evangelism," as it was called, made much of "objective justification," appealing to sinners to "Come as you are." Again, as in Denmark and Norway, a healthy inner mission society was one of the fruits of revival, the National Missionary Society. Following the death of Rosenius, leadership came into the hands of Paul Peter Waldenström (1838–1917), whose subjective views of the atonement led to the formation of the Swedish Mission Covenant Church (1878).

Generally speaking, the two Swedish universities, Lund and Uppsala, represented high and low churchism respectively. The latter viewpoint influenced Parliament to allow the Church of Sweden its own Convocation (1865) with lay representation.

North America. The great 19th-century German and Scandinavian immigration began in 1839–40. The first Germans to arrive were "Old Lutherans" from Prussia whose original pietistic impulses had given way to a high-church confessionalism of the Hengstenbergian and "New Lutheranism" line. Colonies of about 1,000 "Old Lutherans" under J.A.A. Grabau settled in the vicinity of Buffalo and others in and around Milwaukee. They were the forerunners of the Buffalo Synod (1845). Saxon immigrants, under Martin Stephan and Carl F.W. Walther likewise arrived in 1839 and settled near St. Louis to become by 1847 the Missouri Synod. Stephan had practiced conventicle Pietism in Germany and had influenced Walther and others in this direction. Walther and other Missouri Synod leaders later moved to a staunch confessionalism that left little room for conventional Pietism. The Norwegians, who also arrived in 1839, were almost entirely of the Haugean persuasion, one of their first leaders, Elling Eielsen (1804–83), being an extremely legalistic lay follower of Hauge. Most of the subsequent waves of immigrants were sympathetic to Pietism, the laity inclining toward Haugeanism, the clergy towards Johnsonianism. The Danish immigrants, fewer in number, eventually split over the question of Pietism. The anti-Pietists, or Grundvigians, were known as "the Happy Danes," while the pietistic, inner-mission disciples of Beck were denominated "the Sad Danes." The Swedish-Americans reflected "Läsare" and Rosenian Pietism initially, but after the Evangelical Lutheran Augustana Synod was formed in 1860 it soon began to evidence a churchly type of Pietism that perhaps could be traced to Schartauism.

THE ERA OF PROTESTANT EXPANSION

Toleration. The great Protestant advance depended in part on the existence of the secular state and toleration. As late as 1715 the Austrian government had denied all protection of the law to the numerous Hungarian Protestants. But after the French Revolution the few survivals of this old church–state unity were rapidly whittled away. Even in countries in which one church was established, all churches were given some form of protection; Protestant groups could spread, though slowly and under difficulty, in Spain or Italy. Even in tsarist Russia, which did not recognize toleration, Baptists obtained a foothold from which they were to build the second largest Christian denomination of Soviet Russia. Wherever western European and American ideas were influential, Protestant evangelists could work fairly freely, especially in the colonial territories of Africa and India.

Church–state relationships

Though the secular state thus helped Protestant (and Roman Catholic) expansion and variety, it also confronted all churches with urgent new problems. The American pattern, in which the state must have no constitutional connection with religion, stemmed as much from the old Congregational tradition as from the ideas of the Enlightenment and was never antireligious in intention. It was influential among the older churches of Europe. In Protestant countries where state and church had been in alliance since the Reformation the effect was twofold: the state became more neutral in its attitude toward the leading denominations of its territory; and the state church pressed harder toward independence from all forms of state control. Lutheran Germany produced a strong movement toward independence in the mid-19th century. In Scotland the evangelical movement demanded independence from the state in the appointment of ministers to parishes, and when this was refused by the courts and by the government, nearly half the Church of Scotland (1843) under the leadership of Thomas Chalmers (1780–1847) left the established church to found the Free Church of Scotland. The two churches continued side by side (until their eventual reunion in 1929). In Switzerland a Reformed theologian, Alexandre-Rodolphe Vinet (1797–1847), pressed for the separation of church and state and in 1845 founded the Free Church.

In England the move toward independence in a state church was a feature of the Oxford Movement, founded by John Henry Newman (1801–90) in 1833. Here the movement took a course unique in Protestantism. It asserted independence by emphasizing all the Catholic elements within the traditional heritage of Protestantism and so created a school of thought that, though remaining within a Protestant Church, came close to repudiating the Protestant tradition as it was then commonly understood in Europe and America. Newman himself became a Roman Catholic in 1845 and was made a cardinal in 1879. Under the leadership of the survivors the Oxford Movement brought about a transformation in the worship, organization, and teaching of the Church of England within the traditional polity of an established and Protestant church. The remarkable sign of this change was the revival from 1840 on of nunneries and from 1860 on of monasteries.

The Oxford Movement

In German Lutheranism, under the influence of Pietism, Theodor Fliedner (1800–64) established in 1836 a "motherhouse" for deaconesses that became a model for the many successor diaconate orders in Germany, Scandinavia, and the United States. These were the first such to appear in Protestant communities since the dissolution of monastic communities during the Reformation. In the mid-20th century France produced a celebrated community at Taizé devoted to ecumenical prayer and study.

On the whole the trend was always, though slowly, toward a free church in a free state. A few powerful conservative theorists, especially Friedrich Julius Stahl (1802–61) among German Lutherans, strenuously defended one version or another of the old link between throne and altar and the necessity for a single privileged church if revolution or rationalism were to be avoided. These theorists were usually viewed, however, as survivals from a past age. Much more powerful and contemporary were the theorists who, in resisting the trend toward denominationalism and pluralism, saw the church as the religious side of the nation and therefore wanted to broaden its doctrines and liberalize its polity. In England Frederick Denison Maurice defended the established church upon these liberal lines; and in Denmark, more easily because the population was so largely Lutheran, N.F.S. Grundtvig shrank from every form of denomination or confessionalism and wanted to make Christianity the spiritual aspect of Danish national life. Grundtvig's movement had extraordinary success; but Denmark, and to a lesser extent Sweden and Norway, were exceptions to the trend. The older Protestant churches steadily moved farther away from the state and unsteadily but gradually secured more autonomy in their organization.

The rise of American Protestant influence in the world. Since the 16th century the two centres of Protestant political power had been Germany and England. With German unity effected under Prussia and the rise to world power of Britain, the political force of Protestantism was stronger during the 19th century than at any time since the Reformation. But about 1860 it began to be clear that a third force was emerging in the United States. After 1820 American frontier conditions helped to extend the variety of Protestant forces, and denominations such as the Disciples of Christ, formed in 1832 from revivalist groups, arose. These Protestant denominations in time extended their influence beyond America. Many of the immigrants to America were Catholic, and in time the largest single denomination in the United States was to be

the Roman Catholic. But the tone of American leadership and culture remained Anglo-Saxon, liberal, and Protestant. Many Germans and Scandinavians, usually of the Lutheran persuasion, emigrated to America, and American Lutheranism expanded until it became a centre of Lutheran life and thought of a weight equal to the original homes of Lutheranism in Germany and Scandinavia. Because the Lutheran leadership came largely from European pietistic groups, the American Lutheran churches tended to be more conservative in theology and discipline than the churches in Germany. The element of revivalism in American Christianity continued throughout the 19th century and helped the concept of a personal Christian faith to penetrate deeply into the American way of life.

The spread of missions. With the background of European strength in Germany and Britain, with the rising strength of the United States, and with the longest period of peace that Europe had ever known, the Protestant churches entered their greatest period of expansion. Confronted at home by the new cities, they developed social services on a scale hitherto unknown, such as in hospitals, orphanages, temperance work, care of the old, extension of education to the young and to working adults, Sunday schools, boys' and men's clubs in city slums, and the countless organizations demanded by the new city life of the 19th century. Abroad they carried Protestantism effectively into all those parts of Africa that were not under French or Portuguese influence, so that in southern Africa the Bantu became largely a federation of Protestant peoples. In India British and American missionaries steadily increased the strength of the newer Indian Christian churches. In China Christianity had been hitherto confined to the seaports and the survivors of Roman Catholic missions in the 17th century; but now a variety of evangelical groups, mostly financed from England or America and led by the China Inland Mission (founded 1865), created congregations deep in the interior of China. Japan had been closed to Christianity since 1630, and after its reopening in 1859 American and British missionaries created Japanese Christian churches. American missionaries developed Protestant congregations in the countries of South and Central America. All of the main Protestant denominations—Lutherans, Presbyterians, Anglicans, Congregationalists, Baptists, Methodists—developed into worldwide bodies, and all suffered strain in adjusting their organizations to meet these extraordinary new needs.

Social service and missionary activities

REVIVALISM IN THE 19TH CENTURY

One of the most prominent features of Protestantism in the 19th century was the development of revivalist methods to meet the needs of an industrial and urban society. Although many urban poor seldom went to church, they listened to evangelical preachers in halls or theatres, or on street corners. Methodists and Baptists, familiar with revivalistic methods, made many strides forward, especially in the United States. Their efforts were not confined to reaching the working class. The English Baptist Charles H. Spurgeon (1834–92) secured a large audience in London and helped to make the ministry of Protestant dissent very powerful. His mission was for the most part to the educated rather than to the urban poor. For the lowest end of the social scale, a former Methodist preacher, William Booth (1829–1912), and his wife, Catherine, created in east London the agency of evangelism that was known from 1878 as the Salvation Army. They directed their mission to the men on the street corners, using brass bands and even dancing to attract attention. They differed from the Methodist revivalist tradition, from which they had sprung, by their belief in the necessity of a strong central government under a "general" appointed for life, and by abandoning the use of sacraments. At first they met much hostility and even persecution, but by the end of the 19th century the Salvation Army had securely established its place in British life and had become a worldwide organization.

Revivalism in England and Scandinavia

In Sweden a Methodist preacher influenced Karl Olof Rosenius (1816–68), who introduced revivalism into Swedish Lutheranism. He and some disciples also were influenced by the movement that stemmed from Zinzendorf. Though there were links with Pietism, the new movement was quite unlike the little groups of Pietism. The Pietists wanted to gather men to salvation out of the world, whereas the Bornholmers (as they later came to be called in Denmark because of a famous episode in evangelism on the island of Bornholm) wanted to declare salvation to the world. The movement had effects in Norway and Denmark and in the Lutheran Church—the Missouri Synod in the United States—but never became as separate as the Salvation Army.

In the United States the development of revivalism was particularly marked in the expansion of the moving frontier. The memory of the Great Awakening (*c.* 1725–50) was always powerful, and in halls of cities as well as in the camps of the west, revivalistic preaching methods were effective. Protestantism was exceptionally strong because, in many cases, immigrant groups found in religion that link with their historic past that secular society could not for the time give them. Famous evangelists appeared to meet the need of the cities, especially Charles Grandison Finney (1792–1875) and Dwight Lyman Moody (1837–99).

Revivalism in the United States

Thus, some of the evangelistic power in Protestantism of the 19th century was drawn away from the traditional churches of the Reformation—Lutheran, Calvinist, and Anglican—and tended to create new forms of church life and new organizations. These almost always used lay preachers, were far more concerned with bringing the individual to conversion and little concerned with church order, and were sometimes content if they could draw a soul to Christ without worrying if it were drawn into a historical Christian community as understood since the Reformation. Consequently they developed a tendency, not common before the Pietist movement, to identify Protestantism with individualism in religion. Because the evangelistic endeavours subsequently produced separate organizations, the separate denominations and the varieties of Christianity that still called themselves—and with justice—Protestant were rapidly increased.

The secular state allowed or even stimulated the Protestant churches to establish further and powerful varieties of religious groups. Among radical Protestants several important groups or new churches emerged, and several of them were apocalyptic, owing their origin to expectations of the Second Coming of Christ. In Britain appeared the Plymouth Brethren, founded in 1827 by John Nelson Darby (1800–82), who separated themselves from the world in preparation for the imminent coming of the Lord. The Catholic Apostolic Church, formed in 1832 largely by the Scotsman Edward Irving, likewise prepared for an imminent coming. Apocalyptic groups and sects were successfully established in the United States, probably because of the absence in new areas of any settled or habitual church polity. The Seventh-day Adventists were founded by William Miller (1782–1849) of New York, again with an expectation of an immediate end of the world. Though not self-proclaimed Protestants, the Mormons (Church of Jesus Christ of the Latter-day Saints), founded by Joseph Smith (1805–44), came out of a parallel waiting upon the end. Another set of groups arose from the revival of faith healing, the most important being the Christian Scientists, founded in 1879 by Mary Baker Eddy (1821–1910), who set up her first church in Boston.

NEW ISSUES FACING PROTESTANTISM IN THE 19TH CENTURY

Churches and social change. Attacks on the churches during the 19th century (and after) were twofold: social and intellectual. Rapidly growing cities and industry created a proletariat estranged from religious life. Many of the political leaders, especially in Europe, claimed that the churches were bulwarks of that order of society which must be overthrown if justice was to be secured for the working class. Some of the earlier forms of socialism were atheistic or at least deistic and suspected free churches as fiercely as they suspected an alliance between altar and throne. Social and economic thinkers, such as Karl Marx (1818–83), argued that religion was the opium of the people, that it bade human beings to be content with their lot when they ought to be discontented.

In response to such views, in nearly every European

country, Catholic or Protestant, there came into existence groups of "Christian Socialists," who believed (at least) in the doctrine that workers had a right to social and economic justice and that a Christian ought in conscience to work toward those political conditions that would achieve more social justice for them. Except for these basic views the Christian Socialists varied greatly in their outlook and ideas, whether political or theological. Adolf Stöcker (1835–1909), a court preacher in Berlin, was an anti-Semitic radical politician; and Charles Kingsley (1819–75), a clergyman novelist in England, was a warmhearted conservative who deeply sympathized with and understood the working class. The most profound of all the Christian Socialists was Frederick Denison Maurice (1805–72), a theologian of King's College in London until he was ejected in 1853, then a London pastor, and finally a professor of moral philosophy at Cambridge.

Christian Socialists and the Social Gospel

But in England and America the radical wing of Protestants—especially Baptists and primitive Methodists—did as much for the workers' religion as the intellectual leadership of a few Anglican theologians. In some cases the endeavours made Socialist parties possible for the Christian voter; in others they persuaded Christian voters or politicians—without actually voting for a Socialist party—to adopt policies that led toward a welfare state. Nevertheless, they made Christians more conscious of a social responsibility. In America the Social Gospel excited much influence in the churches at the end of the 19th century, and its most influential leader was a Baptist, Walter Rauschenbusch (1861–1918). Whereas in Catholic countries political parties arose that especially appealed to Christian voters and often used the word Christian in their name, in all the Protestant countries all political parties needed to appeal to Christian voters, and few avowedly secular parties had political success.

Biblical criticism. Besides political, social, and economic criticism, Protestantism was encountering an intellectual onslaught on Christianity. There were thinkers who declared that the advance of science and of history proved the Bible, and therefore Christianity, untrue. The question of biblical criticism was first posed in the German universities; *i.e.,* whether a man might be a Christian and even a good Christian though he held some parts of the Bible to be not true. This became the great question for Protestantism, if not for all Christendom, in the 19th century. On the one hand Protestantism stood by the Bible and declared that the truth of God came from the Bible. On the other it rested in part on a fundamental conviction of the liberty of the human spirit as it encountered the Bible. Protestantism was thus seldom friendly to the tactic of meeting argument merely by excommunication or by the blunt exercise of church authority. The theological faculties of German universities, being state faculties and not church institutions, suffered much internal stress, but they arrived at last at the conviction that reasoned criticism—even when it produced conclusions opposed to traditional Christian thinking—should be met rather by refutation than by way of authority. Thus German Protestantism showed at length an elasticity, or openmindedness, in the face of new knowledge, which was as influential in the development of the Christian churches as the original insights of the Reformation. Owing in part to this German example, the Protestant churches of the main tradition—Lutheran, Reformed, Anglican, Congregational, Methodist, and many Baptist communities—adjusted themselves relatively easily (from the intellectual point of view) to the advances of science, to the idea of evolution, and to progress in anthropology and comparative religion.

Influence of philosophical thought on biblical criticism

In such a flux of ideas, with the Protestant tradition seemingly under attack from Protestants, there was naturally a wide variety of approaches, both in philosophy and history. There was an opinion, represented by the German philosopher G.W.F. Hegel (1770–1831), that Christianity should be restated as a form of Idealistic philosophy. This view was influential for a time in Germany and afterward among Oxford philosophers of later Victorian England. Such restatements were subjected to destructive attacks, of which the most powerful were published by the Danish philosopher Søren Kierkegaard, chiefly because such reasoned philosophy failed altogether to account for the depths and tragedies of human existence. An earlier opinion sought to base the justification of Christian faith in the religious feelings commonly found in humanity. A German philosopher, F.D.E. Schleiermacher (1768–1834), sought to infer the Christian and biblical system of thought from an examination of human religious experience. Schleiermacher's attempt had much influence on Protestant thought. Throughout the 19th century the appeal to religious experience was fundamental to liberal Protestant thinking, especially in the attempt to meet the views of modern science. Probably the most important of the successors of Schleiermacher was Albrecht Ritschl, who wholly rejected the ideas of Hegel and the philosophers; he distinguished himself sharply from Schleiermacher by repudiating general religious experience and by resting all his thought upon the special moral impact made by the New Testament on the Christian community. Between 1870 and 1918 the Ritschlian school was one of the leading theological schools of thought within the Protestant churches.

Meanwhile, scholars made long strides in the study and exposition of the Bible. Freed from the necessity of defending every one of its details as historical truth, professors at Protestant universities were able to put the books of the Bible into a historical setting. This made an important difference in the study of the New Testament but was a revolution so far as the Old Testament was concerned, where the entire earlier accepted chronology was changed. German Rationalist or Hegelian historians were the first to study the problems with freedom. Ferdinand Christian Baur (1792–1860) of the University of Tübingen applied the methods of Hegelian philosophy to the documents of the New Testament, which he conceived to be products of the clash between the Jewish Christians led by Peter and the Gentile Christians led by Paul. This theory, known as the Tübingen theory, soon receded in influence; but in aid of this theory Baur expounded the texts with such ability as to make his study a landmark in the study of the Bible. Among a large number of excellent biblical students, Joseph Barber Lightfoot (1828–89) of Cambridge finally demolished the Tübingen theory by showing the 1st-century origin of most of the New Testament texts; and Adolf von Harnack (1851–1930) of Berlin by the end of the century summarized the results of a century that was revolutionary in the area of biblical study.

Historical criticism of the Bible

Protestantism in the 20th century

MAINSTREAM PROTESTANTISM

The war of 1914–18 broke Europe's waning self-confidence in the merits of its own civilization. Since it was fought between Christian nations, it weakened worldwide Christianity. The seizure of power by a formally atheist government in Russia in 1917 brought a new negative pressure into the world of Christendom and sharpened the social and working class conflicts of western Europe and America. During the following 40 years the Protestant churches suffered inestimable losses.

Germany under Adolf Hitler (in power 1933–45) professed to save Europe from the threat of Bolshevism; and the Nazi rule was at first welcomed by many German churchmen. Disillusionment was not slow to follow. From September 1933 there already existed a partial schism between churchmen willing to cooperate with the government in church matters—especially over the Aryan clause that demanded that no Jew should hold office in the church—and those, led by Martin Niemöller, who were not willing to cooperate in church matters. With the support of the state-aided Lutheran churches in the south (Bavaria and Württemberg), Niemöller's group was able to form the Confessing (or Confessional) Church, and the schism was made manifest when the Confessing Church held the Synod of Barmen in May–June 1934. For a time the Confessing Church was strong throughout Germany; but when the German government provided a less doctrinaire government under the minister of church affairs Hanns Kerrl, the Confessing Church was itself divided—

The effect of Nazi control in Germany

into those who were willing to cooperate and Niemöller's men, who were not willing to cooperate because it was a church government imposed by the Nazi government. At the Synod of Bad Oeynhausen (February 1936) the Confessing Church broke up and was never again so strong. In the later stages, especially during World War II when the extreme Nazis secured complete control of Hitler's government, the churches came under increasing pressure and toward the end were struggling in some areas to survive. Bishop Theophil Wurm of Württemberg was a leader in protesting to the government against its inhumane activities, and Pastor Heinrich Grüber, until his arrest, ran the Büro Grüber, which sought to evacuate and protect Jews. Some church leaders, notably the theologian Dietrich Bonhoeffer, paid with their lives for their associations with resistance to the Nazi government.

The results of World War II

The end of the war saw Russian armies in control of eastern Europe and Germany divided. All the churches in the area came under pressure. Most Germans were evacuated or deported from the three Baltic states of Lithuania, Estonia, and Latvia. Although Lutheran communities remained there, they were subjected to persecution, especially under the rule of Stalin. The Lutherans in Transylvania (Romania) and the Reformed in Hungary came under less severe pressure but were much diminished in numbers. The Protestants of Czechoslovakia, led by the theologian Joseph Hromadka, succeeded in maintaining more dialogue with Marxist thinkers than did Protestants elsewhere in Europe. From the viewpoint of Protestant strength, the greatest losses were suffered through the division of Germany. The settlement between the victorious powers gave large areas of former German-speaking (and largely Lutheran) areas to Poland, and many (approximately 8,000,000) Germans were expelled; most went to western Germany. The Soviet occupation zone of Germany in 1945 included Wittenberg and most of the original Protestant homeland. East Germany (the German Democratic Republic) became the sole country in which a Marxist government ruled a largely (70 percent) Protestant population. For a time the Lutheran churches were the chief link between East and West Germany, and the annual meeting, or *Kirchentag,* the single expression of a lost German unity. But the building of the Berlin Wall in 1961 stopped this communication and isolated the East German churches. Despite governmental pressure, especially in relation to money, education, and church building and in the national (and anti-Christian) form of youth dedication, the East German Protestants worked courageously and flourished. The 450th anniversary of the Reformation on Oct. 31, 1967, showed how strong a hold the Protestant churches still had over the affections of a large number of people.

In Russia, a deeply Orthodox state before the Revolution of 1917, the 40 years after the Revolution witnessed a growth in the Baptist community. The flexibility and simplicity of Baptist organization made it in some respects more suitable to activity under difficult legal conditions. In the years after Stalin's death in 1953 there was evidence of rapid advance; but after 1960 the Baptist communities, like the Orthodox, again came under pressure, which at times was severe.

The effect of the decline of the British Empire

The material losses that Great Britain suffered in World War II and the end of the British Empire in the years after 1947 had serious effects on the Protestant churches in former British territories. The home country could no longer provide money and human resources to the overseas churches on the same scale, and in a few areas church government was handed over to leaders who were not ready to take over church leadership. But in other areas the change of status for Britain hastened the process of change in leadership that had been proceeding slowly; and some of the failing resources were supplemented from elsewhere, especially from the United States, Canada, and Australia. Thus the so-called younger churches came to be a new fact of world Christianity, led by men who no longer saw the history of Christianity solely through European eyes and had an impatience partly derived from a different attitude to the Christian past. This was to be of primary importance in the ecumenical movement. Meanwhile, the secularizing trend of a technological age assailed the old European churches and had an even greater effect upon the areas where the younger churches ministered.

The growth of mainline Protestantism in sub-Saharan Africa, as of Lutheranism in South West Africa/Namibia or Anglicanism in South Africa—as well as of the Pentecostal and Evangelical churches and sects in South America and Asia—helped compensate for losses in Europe and North America. Because of conversions and population growth, the Protestant church actually increased in size as it changed its scope and ethos.

There were also surprising survivals and reappearances of Protestantism in areas of the world where its demise had been foreseen. Thus, in 1948–49 the Communist seizure of power in China effectively ended Protestant missions there. By 1951 there were hardly any European missionaries in the country, and the Chinese churches had to stand without outside aid. They came under severe pressure, especially during the so-called Cultural Revolution in the 1960s and '70s. They could no longer evangelize and sought barely to survive. The partial reopening of China to the West and the cautious measures granting more freedom of religion and speech beginning in the late 1970s and the 1980s led to new contacts between Chinese Protestants and Westerners. It was estimated that several million Protestants and other Christians had endured the suppression and persecution of the two previous decades, and, however uncertain their futures remained, they represented a vital group of churches.

CONSERVATIVE AND EVANGELISTIC FORMS OF PROTESTANTISM

The most important movements in 20th-century Protestantism took root in soil that most call conservative, and some of their founding had a reactionary character. At the same time not all members of these movements wished to be typed as conservative. Their forward-looking and exuberant expressions of faith displayed more radical outlooks. The three main movements are usually called Pentecostalism, Fundamentalism, and Evangelicalism. The first has been of immeasurable importance in the spread of Protestantism beyond its historic European home.

Pentecostalism. Pentecostalism emerged out of Wesleyan Holiness movements at the turn of the century in the United States. In 1901 in Topeka, Kan., and in 1906 in Los Angeles, there were particularly notable manifestations of various phenomena that characterize the movement. Central to these is glossolalia, "speaking in tongues." This is a form of unrepressed speech whose agents "yield" themselves to the Lord. Normally the syllables they speak or sing are unintelligible, though some claim that they speak in recognizable foreign tongues as the disciples of Jesus did at the first Pentecost, from which the movement derives its name. Pentecostalists believe that they must experience a "second baptism," beyond water baptism, in which the Holy Spirit comes to them. They not only speak in tongues but also interpret them; they prophesy; many engage in healings, claiming that miraculous healings did not cease after the apostolic period, as many other Christians claim they did.

Speaking in tongues

The Pentecostal movement in the United States was often Southern, associated with the "Bible Belt," and developed among the rural poor whites or urban blacks. After the mid-20th century, through fast-growing denominations like the Assemblies of God, Pentecostalism emerged as one of the most visible forms of Protestantism and became increasingly acceptable to the middle classes. After 1960 the movement spread into mainstream churches like the Episcopal, Lutheran, and Presbyterian, where participants often called it a "charismatic" movement.

Pentecostalism had its greatest success in the Caribbean, Latin America, and sub-Saharan Africa. There many prophetic movements erupted, in which Christians adopted emotional forms of worship and healing. Pentecostalism in these parts of the world was often the religion of the poor, bringing hope to people in nations that were emerging from colonialism. The Pentecostalists, building on the work done by missionaries a century earlier, were not often anti-American or anti-European, as some liberation

movements were. In fact, they often accented "otherworldliness" and avoided politics or identified with conservative and even repressive regimes.

Fundamentalism. The second major movement, Fundamentalism, combined late 19th-century premillennialism with more or less rationalistic defenses of biblical inerrancy. It took its name from a sequence of tracts called *The Fundamentals* that were issued between 1910 and 1915 in the United States, and the movement became institutionalized in 1919 and 1920, as Fundamentalism became a formal and militant party in denominational conflict in the United States.

The most obvious causes for the rise of Fundamentalism were the spread of Darwinian evolutionary theory and its acceptance in the more liberal parts of the Protestant churches as well as the higher criticism of the Bible. Fundamentalists in the United States felt that these two movements were subverting seminaries, bureaus, mission boards, and pulpits in the northern branches of denominations like the Baptists and Presbyterians. The Scopes trial in 1925, in which the Fundamentalist champion William Jennings Bryan fought against the teaching of evolution in schools and defended the Genesis record as being scientific, coincided with the climactic denominational battles in those two churches.

The Scopes trial

The Fundamentalists tended to lose the political battles but survived with their own network of Bible colleges, radio programs, and publishing ventures. In the early 1940s they regrouped into several competitive Fundamentalist organizations that steadily gained followers, visibility, morale, and assertiveness. They prospered most when they moved from a generally passive political posture to open participation, particularly in support of Ronald Reagan's successful presidential bids in 1980 and 1984.

Groups like the Moral Majority, founded by Fundamentalist evangelist Jerry Falwell, demonstrated how effective the television ministry of the movement could be. The Fundamentalists concentrated political energies on opposition to abortion, support of an amendment that would permit prayer in public schools, and identification with the causes of Israel and a strong military defense budget.

Evangelicalism. The third movement is Evangelicalism. Focused for decades in the ministry of figures like evangelist Billy Graham and journals like *Christianity Today*, this conservative and evangelistic group tended to agree with Fundamentalism on cardinal doctrines: the virgin birth, substitutionary atonement, and physical resurrection of Jesus. Most Evangelicals insisted on some version of biblical inerrancy, but gradually more and more scholars of the movement questioned whether that was the best way to assert faith in biblical authority. Nor did all agree with those Fundamentalists who stressed premillennialism.

Evangelicals, however, were more moderate than Fundamentalists; they agreed with the older-style Fundamentalists in substance but differed in style. They found Fundamentalists to be too negative about culture, too withdrawn into sects, too rude and blustery and judgmental. When Evangelicals formed the National Association of Evangelicals in 1942, they were attacked from the Fundamentalist right much as they attacked the mainstream moderates and liberals. Most of them preferred to see themselves not as well-mannered Fundamentalists but rather as perpetuators of the 19th-century Protestant mainstream.

To that end the Evangelicals increasingly reentered the world of cultural, social, and political engagement. Rather than build Bible schools, they concentrated on liberal arts colleges. Some Evangelicals even engaged in radical social programs and criticized conservative Protestantism's overidentification with militarism and unfettered capitalism. They also acquired considerable if slightly unpredictable political power in the United States and elsewhere.

Evangelicals also tended to be ecumenical; Billy Graham welcomed Catholic and mainstream Protestant leaders on his platforms, and he prayed with many kinds of Christians whom Fundamentalists would shun. Whereas Fundamentalists and Pentecostalists had counterparts in the Third World, Evangelicals tended to form international movements and hold conferences designed to bring Christians of many nations together.

While Fundamentalists usually split off into churches of their own, millions of Evangelicals remained connected to mainstream denominations and increasingly moved fully into the mainstream. But they always endeavoured to keep alive their doctrinal distinctiveness and their passion for witnessing to Christ.

THEOLOGICAL MOVEMENTS WITHIN PROTESTANTISM

Meanwhile, a certain reaction could be observed in the Protestant tradition of theology. This was partly due to a general doubt about European liberalism after World War I and particularly due, in its further development, to a reaction against attempts by the Nazis to use liberal theology for some of their views of society.

In both the 19th and 20th centuries liberal theology met much criticism on the ground that it narrowed Christianity to the limits of what men believed themselves to be experiencing or turned what was objective truth into subjective feeling. Though himself no conservative, Kierkegaard was the most extreme of these critics. All the conservative theologians—including the earliest members of the Oxford Movement in England, the evangelical tradition generally, and those many who stood by the inerrant word of the Bible and in the 20th century came to be called by the name Fundamentalist—opposed the liberals on the same grounds. But in the 20th century there was a reaction even within the liberal camp. Beginning in 1918 a reaction against all theologies emphasizing religious experience was led by Karl Barth of Basel and Emil Brunner of Zürich. This theological movement, called Neoorthodoxy, widely influenced Protestant thinking in Europe and America. Barth and his disciples regarded their work as a reassertion of the true sovereignty of Scripture and as a return to the authentic principles of the Reformation. In America Reinhold Niebuhr was almost as influential in reacting against liberal Christian philosophies as they applied to society and to man. Yet that the questions the older theologians had sought to meet still remained was shown by the influence exerted by the German theologian Rudolf Bultmann of Marburg, who sought to "demythologize" the New Testament by discovering its core truths and thus allowing its significance for faith to be more fully disclosed. Refugees from Nazi Germany, such as Paul Tillich, interpreted European developments to Americans.

The rise Neo-orthodoxy

The Neoorthodox synthesis did not outlast the generation of the giants who gave voice to it, and Protestant theology after the mid-1960s was in disarray. Europe lost its hegemony, though certain theologians, among them Jürgen Moltmann, began to take elements of Neoorthodoxy and combine them into variously described movements, such as "theology of hope," "political theology," "theology of revolution," or Protestant versions of "liberation theology." Espoused in the Third World by theologians who stressed witness to the fact that God sides with the oppressed and the poor or in the United States by feminists or black theologians who developed new interpretations of biblical and traditional texts, these theologies called into question what seemed to be the patriarchalism, elitism, and racism of much earlier academic theology.

Numerous movements adopting liberation theologies coexisted. In general they shared a tendency to particularize Protestant thought. One approach was to make much of cultural contexts. Thus, there was African or Asian, feminist or black theology. In all these cases interpretations were perceived as coloured by the "pre-understanding" people or groups brought to the reading of the texts. Another approach was to focus on "narrative theology" or "story theology" in an effort to move from abstract theology to concrete understandings centring on people. Finally, thanks to the rise of Pentecostalism and Fundamentalism, there developed across the Protestant spectrum fresh attention to the doctrine of the Holy Spirit and to eschatology, the teaching about "the last things."

THE ECUMENICAL MOVEMENT

The ecumenical movement was in origin exclusively Protestant (though Eastern Orthodox leaders soon took part) and was at first largely dominated by Protestant thinking. Its origins lay principally in (1) the new speed

of transport across the world and the movement of populations that mixed the denominations as never before; (2) the world reach of traditional denominations; (3) the variety of religion within the United States and the problems that such a variety created; and (4) the younger churches of Africa and Asia and their contempt for barriers raised by events of European history for which they felt no special concern. There was always a strong link with the missions, and an American Methodist missionary leader, John R. Mott, whose travels did as much as anything to transform the various ecumenical endeavours into a single organization, represented in his own person the harmony of missionary zeal with desire for Christian unity. A conference at Edinburgh in 1910, which marks the beginning of the movement proper, was a World Missionary Conference. From it sprang conferences on life and work (led by the Swedish Lutheran archbishop Nathan Söderblom), dealing with practical problems, as well as conferences on faith and order, at which theologians sought to examine their theological differences with sympathy. In the beginning Roman Catholics refused to participate; the Eastern Orthodox participated only through exiles in the Western dispersion; and the Nazi government refused to allow Germans to go far in participating. By the end of World War II in 1945 it was evident that there was a new atmosphere, and the World Council of Churches was formally constituted at the Amsterdam conference in 1948. The entire movement depended for most of its money and for part of its drive on the Americans; but its headquarters was in Geneva, and, under the guidance of its first General Secretary, Netherlands Reformed administrator W.A. Visser 't Hooft, it never lost sight of the fact that the traditional problems of divided Christian Europe had to be met if it was to succeed.

Formation of the World Council of Churches

In the years after 1948 the ecumenical movement brought Protestants into an ever-growing dialogue with the Eastern Orthodox and the Roman Catholics. After John XXIII became pope in 1958, the Roman Catholics at last began to participate in the ecumenical movement. Although the definitions of the second Vatican Council (1962–65) were unacceptable to most Protestants, they had a breadth quite unlike the definitions of the first Vatican Council in 1870 and encouraged those (usually liberal) Protestants who hoped in time to lower this greatest of barriers raised by the 16th century.

For specific information on ecumenical efforts in the second half of the 20th century, see the entries on individual denominations below. (E.C.N./M.E.M./W.O.C.)

THE PROTESTANT HERITAGE

The basic doctrines of Protestantism at the Reformation, in addition to those of the creeds, were: justification by grace alone through faith alone; the priesthood of all believers; the supremacy of Holy Scripture in matters of faith and order. There has been variation in sacramental doctrine among Protestants, but the limitation of the number to the two "sacraments of the Gospel," baptism and Holy Communion, has been almost universal.

In theory Protestantism has stood throughout its history for a principle of protest that calls under judgment not only the beliefs and institutions of others but also one's own movements and causes. On those grounds, however, most students of Protestantism would recognize that the Protestant tradition has not been substantially more successful than have other faiths at remaining self-critical or at rising above institutional self-defensiveness.

Within the spectrum of non-Roman Catholic Western Christianity a great variety of doctrinal views and polities have been expressed. Not all Western non-Roman Catholic Christians have been ready to be included in Protestantism. Some Anglicans and Lutherans, for instance, have been so eager to stress their continuity with the historic Roman Catholic Church and their distance from extreme Protestantism that they have asked for separate designations. Courtesy suggests that such appeals be taken seriously; however, ultimately habits of speech and sociological usage tend to predominate and, despite their protestations, these groups are usually included in the Protestant cluster.

Teaching, worship, and organization

COMMON PRINCIPLES AND PRACTICES OF THE MAGISTERIAL REFORMERS AND THEIR SUCCESSORS

Justification by grace through faith. The original Protestant leaders united in their contention that what separated them from the Roman Catholicism of their day was their teaching that man is justified by grace through faith. Devotion to this teaching has been central to Protestantism throughout its history. Although there have been subtle variations in the differing Protestant church bodies, a core of shared belief was at first easily discernible.

Concern for "justification"

Concern for "justification" was related to the obsession that in the 16th century was often expressed in terms of finding oneself on good terms with God. The metaphors were drawn from the courts of law. Aware of its shortcomings, its ignorance, its sin, and its guilt, mankind saw itself standing before a bar of justice presided over by God. Without help, the individual could expect nothing but God's wrath and condemnation. This meant that he would perish everlastingly, and his present life would be full of torment. Yet the Bible also presented mankind with a picture of a loving and gracious God, who may very well desire happiness for all. The question then was: how could the individual be sure that God would reveal his gracious, and not his wrathful, side? How could he have the confidence that he was included in the positive loving action of God?

The teaching of the Reformers becomes most intelligible when seen against the Western Catholic doctrines (*e.g.*, sin, grace, atonement), as they saw them. In the Protestant view the late medieval Catholic teaching held that a human being was brought back to God only when so much grace had been infused into his soul that he merited the favour of God. God could not have been expected to accept someone who was unacceptable, but he could impart something that would make humans acceptable. This something was grace, and its flow depended upon the merits of God's perfect Son, the man Jesus Christ. The church, according to medieval Catholicism, in a sense controlled the flow through the sacramental system and through its hierarchy.

To the Reformers the Roman Catholic sacramental system seemed to be part of a transaction that was always going on between man and God. In it, people made sacrifices designed to appease and please God. They would attend the mass, bring offerings, show sorrow, do penance—which might involve self-punishment or compensatory good works—until God would be gracious. The leaders of the church, from priests through bishops and popes, mediated the transaction. The Reformers believed that such an arrangement could easily be misused as a political instrument for forcing rulers to comply with the church's wishes and as a personal instrument for keeping people in uncertainty or terror. It was this vision of Catholicism that helped inspire the Protestant leadership to rebel and to define justification in other terms.

The terms for this Protestant teaching came from the Bible, especially from the New Testament, and even more so from the writings of St. Paul. In St. Paul they saw a religious hero and thinker who had endured a spiritual quest similar to their own. He could be described as having been brought up in a legalistic version of Judaism, a system in which he was constantly striving to please God by following his Law, particularly as set forth in the Old Testament through the Ten Commandments. Yet Paul failed and was assailed by doubts about his worthiness and his salvation. His conversion meant a radical turning and a free acceptance of God's favour "in Christ." This meant that in faith a person could be so identified with

Influence of the writings of St. Paul

Jesus Christ that when God looked at him, he saw instead the merit that Christ had won through his self-sacrifice on the cross. God looked, in short, at the sinner; but he did not see the sinner. He saw his perfect Son. So he could declare the person righteous; he could justify him—even though the person was still a sinner.

When taken out of the historical context of St. Paul's teachings in the letters to the Romans or the Galatians and transferred to their own times, the Reformers' teaching of justification relied heavily on the work of the Holy Spirit. The Holy Spirit, in effect, made Christ's action contemporaneous with the sinner's quest. God was working now on behalf of those in need. Through preaching, humanity learned of Jesus Christ's sacrifice and death. If the individual believed this historical narrative and, more importantly, if by the power of the Holy Spirit he believed that it was told and enacted for him, he stood before God in a new light. Grace was not infused into him to the point that he became acceptable and pleasing to God. Instead, while the individual was still a sinner, God accepted him favourably and justified him. Christ's death on the cross was then the only "transaction" that mattered between God and man. The sacraments reinforced the relation and brought new grace, but no pretense was made that the human subject had achieved satisfaction before God or produced enough merit to inspire God to act.

Man's freedom and the problem of "good works"

In the Reformers' view the new situation was one of freedom. Whereas Catholics constantly stood in fear as to whether they had provided enough merits, had achieved enough good works, or had pleased the church as God's bargaining agent, the Reformers' version had the believers standing before God completely freed of these nagging questions. They were liberated both from the terrors of sin, death, and the devil, on one hand, and, on the other hand, from the enslaving pride that went with the belief of human beings that they had achieved or at least had substantially cooperated in their own salvation.

This left the Reformers with a serious question, one to which their Roman Catholic opponents regularly referred. What had happened in this teaching of justification and freedom to the biblical accent on good works? Jesus himself, in the Synoptic Gospels (Matthew, Mark, and Luke), was constantly preoccupied with the effort of making people better, of having them bring forth "good fruit." Even Paul shared such concerns. Had the Protestant movement slighted these concerns in its desire to free human beings from the necessity of merits and good works?

The literature of Protestantism is rich in its expression of answers to such questions. The Reformers were virtually unanimous: good works did not produce appeasement of God or salvation, yet they inevitably flowed from the forgiven heart and were always the consequence of the justified person's life. The Law of God could never be used as the saving path along which human beings walked, as a sort of obstacle path or road map to God. Instead, the Law of God measured human shortcomings and judged them. A gracious God acting through his Gospel brought human beings back to him.

Man as simultaneously righteous and sinful

The Reformers' vision of human beings implied in such teachings was doublesided. They believed that from God's point of view the justified person was so identified with Jesus Christ that he shared Christ's perfection. The same person, throughout all his life, left to his own devices or when seen by God apart from Christ's sacrificial work, remained a sinner. The difference came through God's gracious initiative; nothing that a person did started the process of his justification. To the eyes of many in subsequent generations the result was an apparently pessimistic and gloomy view of human potential in Protestantism. The will was bound, apart from God's loving activity. No merits or good works would satisfy God. Sometimes the phrase total depravity was used to describe the human condition, though it must be said that the term had connotations in the 16th century that were different from those that it has today. It was used not so much to provide lurid connotations for descriptions of the depth of sin but rather to describe its extent; man as a total being was in trouble. Even good works, piety, religiousness, and efforts, apart from justification by grace through faith, fell under God's curse. On the other hand, the justified sinner could be described in the most lavish terms, as one who could be "as Christ" or even sometimes "a Christ."

Those who have heard this Protestant teaching outlined through the centuries have regularly seen the difficulties it raises insofar as the portrait of God's character is concerned. Protestants never came up with logically satisfying answers to the resultant questions, though they were convinced that they were faithful witnesses to biblical teachings concerning the mystery of God's nature. The central question: if everything depended upon God's initiative and yet the majority of people are not saved, does this not mean that God is responsible for creating humans only to have them suffer; is he not guilty of the worst kind of cruelty by being the sole agent of their damnation?

In facing the question Protestant leaders differed slightly from each other. Some said that whenever people were saved, it was to God's credit; whenever they were lost, it was through their own fault. They were free to hear the Word; they were free to respond and accept the gift of grace in Christ; their own hardness of heart kept them from freedom and new life. Others ran the risk of presenting cruel pictures of God's nature and action in their interest to witness to his sovereignty and initiative. The view that God predestined some people to be saved and others to be damned was called "double predestination." Some theologians argued that God did this predestining before humans fell into sin; others saw it as a new act of God consequent upon man's fall. Those Protestant parties that were generally non-Calvinist in outlook were usually less systematic and less logical in their statements. The non-Calvinists taught a doctrine called "single predestination." They shared the Calvinists' affirmation of God's total responsibility for human salvation; but they tended to be silent or to relegate to the area of mystery and unanswerable questions the issue of how God could then be other than responsible for human damnation. In general the Protestants saw themselves to be more successful at preserving the teaching of God's sovereignty and the corollary of human helplessness than they were at making his character attractive to all. They saw themselves overcoming this problem in biblical terms by a stress on his loving relation to humanity in sending his own Son, Jesus Christ, to suffer on its behalf.

Predestination

The "priesthood of all believers." If the teaching of justification had important consequences for the doctrines of God and of man in Protestantism, it was of at least equal import for any statement of the meaning of the church and especially of the relations between clergy and laity. The medieval system, sacramental and hierarchical, in effect gave the priests a monopoly in monitoring the transaction between God and man. The Protestant teaching of justification broke this down and the Protestant leaders reverted to what they held to be the biblical view, that all believers have a share in spreading the word of grace and the acts of forgiveness. The result was an emphasis not on the privileges of a priestly caste but rather on "the priesthood of all believers."

The Reformers viewed this teaching as based on the free-flowing sense of authority that existed between Christ and his Apostles, who had been pictured in the Gospels as being active apart from an elaborate clerical church order. At the same time they believed that their doctrine would effectively displace the Roman Catholic hierarchical thought and action. Now all people were to be enjoined to take responsibility for each other's salvation; any Christian could represent the needs of all others before God. Originally the priesthood of all believers was an enlargement of the view that all Christians had intercessory powers, that they could all pray for one another. But it came to refer to the Protestant view of an equality of status between clergy and laity and to the common calling of all Christians to be agents of God's Word and grace.

Priesthood and coresponsibility

The affirmation of the priesthood of all believers had widespread implications in society. In Protestant areas and nations the privileges of the clergy were limited and the scope of lay activity enlarged. All believers shared a "vocation" (calling), and priestly vocations were not considered to be more meritorious or nobler than lay vocations.

Monastic vocations were almost entirely swept away, and restorations of the monastic ideal have been rare and exceptional in Protestant history. Protestants kept, for the most part, a rite of ordination (though some Anabaptists dispensed with all acts that seemed to imply separation between a ministry of ordained persons and laymen) but did not regularly view it as a sacrament. That is to say, ordination conferred no special grace on men. In part a ministry was kept on a pragmatic basis; the clergy were to tend to the business of studying and preaching the Word, properly administering the sacraments, and caring professionally for the health of the church. A set-aside ministry was also derived from biblical precedent in the Book of Acts and early Christian letters.

Protestants, while acknowledging their belief in the equality of laymen and clerics in the priesthood of all believers, have not always seen themselves as particularly successful in clarifying the laity's role. In most cases laymen were not to be the preachers in public worship, and administration of the sacraments usually remained in clerical hands. By demanding of preachers expertise at expounding the Bible, Protestants often have made educational requirements a basis for ordained ministry, at the expense of a full lay involvement. Yet their views did greatly enhance both the theological and practical status of laymen, when contrasted to the situation in medieval Catholicism.

If all believers were priests, then no single church could monopolize the mediation of grace, since Protestants saw that there were believers in all churches, Roman Catholic and Protestant, Lutheran or Calvinist or Anglican. As a result the teachings inherited from medieval Catholicism about the visible and the invisible church were called into question. To many Reformers, most notably Luther, the church was always visible because it was made up of people. But its limits and borders were invisible since one could not examine the heart of others to determine exactly who were the true believers and who were the faithless. This inability to define the boundaries of the church led other Reformers, among them Calvin, to continue to employ the distinction between a visible church and an invisible one, the latter referring to the people who were saved, even if they were in churches where full doctrinal purity had not been achieved. People see the visible, humanly organized church of Christ, but they cannot simply identify this with the Bible's one, holy, catholic, and apostolic church, which is properly discerned only by God and hence invisible to humans. The visible church, in the Reformers' view, almost certainly contained a mixture of members of the invisible church, on the one hand, and hypocrites, or false believers, on the other.

The visible and invisible church

Authority of the Word. Justification by grace through faith and the priesthood of all believers were affirmations that challenged the inherited Roman Catholic views of authority because they seriously undercut the monopoly of the hierarchy in the system of grace. Downgrading the medieval system of authority left a vacuum that Protestantism hastened to fill. Full of variety and pluralism as the movement was from the first, it was rarely characterized by a love of anarchy or indiscipline, and the Reformers set to work at once to establish the locale and extent of authority in the church and the believer's life. Almost unanimously they saw final authority to reside in the Word of God, which tended in the minds of many to be simply equated with the Bible. The need of the Protestant movement to redefine authority enhanced its view of Scripture just as, one might argue, the rediscovery of scriptural teaching was seen to be the primary impetus behind the Protestant movement.

Later generations of Protestant thinkers sometimes resorted to scholastic philosophical definitions similar to those of the medieval Catholic theologians; in such definitions justification became the material (or substantive) principle of the Reformation, while the matter of scriptural authority became the equally important formal (or structural) principle. In some epochs debate about the nature of the Word of God or the Bible was even more preoccupying than was discussion of justification. Protestants often have portrayed medieval Catholicism as being a nonbiblical or even an antibiblical faith, one that denied the Bible to the laity. The expense of reproducing manuscripts led many libraries to chain books to the wall, and the Bible chained to the wall entered Protestant mythology as a symbol of the denial of lay access to the Bible in Roman Catholicism. In many circles Protestantism has been celebrated as a religion of the "open Bible" in opposition to the closed book of Catholicism.

Mythology aside, Protestants without exception concentrated on biblical teaching, and this led to a new passion for translating the Bible into the vernacular and disseminating it as widely as possible, an activity aided by the almost simultaneous invention of movable type and the resultant progress in printing technology. It was to be put into the hands of as many ministers and laymen as possible. Thus, they also had to be taught to read, and the Protestant movement claims some credit for hastening the modern impetus toward the ideal of universal literacy. While the Bible was ordinarily read in the churches and interpretation was shaped by the old and new traditions of these churches (Anglicans read the Bible's teachings on apostolic succession in a way different from that of Anabaptists, for example), what came to be called "the right of private judgment" was often exalted.

Interpretation of the Bible

While Protestants could agree that the Word of God was authoritative in matters of faith and that the Bible as the book inspired by the Holy Spirit had unique status, they did not agree on all interpretations of the Scripture, nor did they unite in a single doctrine of scriptural authority. The Anabaptists, and later the Quakers, stressed an immediate experience of God and thus tended to qualify the importance of the Bible in shaping Christian life. But even among the Lutherans, Calvinists, and Anglicans there were differences of opinion about the Bible.

Latter-day Protestant Fundamentalists have argued that scriptural authority, if it is to be believable and if it is indeed the Word of an all-knowing, perfect God, dare not be described as including errors, even errors in historical or geographic detail. They have backed up this contention with citations from John Calvin and Martin Luther. At the same time, subsequent Protestant contenders for a more open and critical attitude toward biblical literature are equally capable of citing the Reformers—Luther, for example. Luther often spoke of the Bible as "the cradle in which Christ lies" and as a book that derives its authority from Christ and not from an a priori (presumed) doctrine of inerrant inspiration. Their argument concludes that Luther could quite freely relegate some books of the Bible to secondary status or criticize the argument of others, even as he could complain of Paul's grammar or illogical argumentation, and point to errors of detail. Part of the confusion on this matter results from the fact that the Reformers in their era did not see questions regarding truth in the same terms as they have been recognized in the later scientific world, and, also, they could accept the full and governing authority of the Bible without elaborating theories concerning its perfection in detail.

The unquestionable elevation of the Bible as the authority in matters of faith led to a corollary downgrading of other authorities in the church. The hierarchy, and especially the pope, were hardest hit, and papal authority was denied in almost every sector of Protestantism. In place of papal authority more regard came to be paid, at least by conservative reformers, to the Fathers (doctrinal teachers and interpreters) of the early church, who were sometimes cited in the confessional writings of the various Protestant bodies. The Fathers were revered as guides rather than as final authorities. Similarly, a critical attitude toward councils of the church came to prevail. On the one hand, it was widely asserted that councils can quite often and do err, and historical studies pointed to contradictions between various conciliar statements and to unsubstantiated assertions by past councils. On the other hand, many formulas and creedal statements of the ecumenical councils, particularly as these referred to the Trinity or to the Person and work of Jesus Christ, were highly regarded, and many Protestant churches took the chief creedal statements of past councils into their own official body of teaching.

De-emphasis of institutional bases of authority

Canon law, the inherited body of legal materials that regulated faith and morals, quite naturally also suffered

because of the high regard for the Bible. In most Protestant circles it was difficult to make legislation binding upon conscience unless it was based on clearly affirmed biblical legal teaching; more important, accent on the Gospel of grace led most Protestants to want to undervalue the whole role of law in the life of the church. At the same time, new church order soon developed, and it must be said that Protestants often acted as legalistically as did the Roman Catholics, whom they were repudiating. Most Protestant bodies, notably the Anglicans, developed their own versions of canon law or rules of church order and discipline.

The ongoing reformation of the church. The church that is to be judged not by the pope but by a normative Bible, that is grounded in a priesthood of all believers and critically affirmative of Church Fathers and councils, and that rejects inherited codes of canon law differed vastly from medieval Catholicism. In few respects did it differ more than in its establishment of the principle of an ongoing reformation. While most of the Reformers, once established, tended to resist extensions of reformation that would jeopardize their status and definition, almost all Protestants, at least nominally, assented to the idea that *ecclesia reformata semper reformanda—i.e.,* that the church was always reformed and always in need of further reformation. The Protestant movement, then, was conceived as an unfinished product, constantly to be judged by a reading of the Bible, its polity continually subject to debate, its policy open to ongoing appraisal and change. It was in that climate that the sacramental teaching of Protestant churches was argued and developed.

Emphasis on the sacraments. On few points did Protestants disagree more than in their interpretation of the sacraments, but they did unite in their rejection of some aspects of Roman Catholic teaching. Since attack on the sacramental-hierarchical system of salvation was at the heart of their reform, almost nothing of it survived intact. In place of the churchly system the new accent fell on limiting sacramental teaching to those acts clearly commanded by Christ and connected with his promise in the Scriptures. One can argue that, since "sacrament" was not a biblical term, the debate had to do simply with definitions. Most Protestants defined sacraments, then, as acts that impart grace and the new life. They must combine the Word of God and some visible means (like bread, wine, and water); they must have been established by God and instituted by Christ. On these terms, five of the seven chief Roman Catholic sacraments failed to meet the definitional tests: marriage, ordination, confirmation, penance (now called repentance), and extreme unction (now called anointing of the sick). Not in every case did Protestants abolish these acts from their rites, but they ruled them out as sacraments. Thus the Protestant teaching on marriage was normally as "high" as Catholic doctrine and may be considered quasi-sacramental. But it was seen chiefly as a civil act blessed by the church, and it did not convey grace to the participants, nor was there a visible "means."

The sacraments commanded by Christ

Though Protestants—with a few exceptions, chiefly Anabaptist and Quaker—had little difficulty limiting the number of sacraments and perpetuating a high regard for those that survived the change in definition, they were far apart in their understandings of what went on in sacramental acts. Basically three views were debated. To the "right," as one might call it, was the Lutheran view, which critics considered as being quite close to Roman Catholicism. Luther seemed to bring with him something of a medieval worldview, in which symbols of the material world were transparent to another invisible, divine order. This attitude made it possible for him to make much of the material objects in the sacraments. When he connected them with biblical words, he was able to say of bread and wine that these are the body and blood of Christ, and of baptism, that it effected a change in the believers' status before God.

At the "left" was the view of Huldrych Zwingli and other Swiss Reformers, who accented the spiritual side and downgraded the material. In some respects they shared more of what has come to be considered a modern view of matter and spirit, in which the symbols were opaque, disengaged from an invisible "other order." Such teaching meant that what mattered most in the sacraments was the following of Christ's commands, the reminiscence of his participation in the world of his disciples, and the spiritual intentions brought to the acts of believers. For Zwingli the bread and the wine were symbols that merely represented the body and blood of Christ, and baptism was more a sign of a Covenant with God than a supernatural imparter of grace. Between the Lutheran and Zwinglian views were Calvinist and Anglican attitudes and definitions. All Reformers agreed, however, in their criticism of the Roman Catholic teaching called "transubstantiation," which held that the actual "substance" of the bread and wine in the Lord's Supper was turned into the body and blood of Christ. But they did not agree over the alternatives to that teaching, and debate over the sacrament did as much as any other theological factor to contribute to internal Protestant division.

Relationship between the community of the baptized and the political community. Equally varied were the attitudes toward civil authority among the various Protestant parties. Martin Luther expressed what in theory could have been a most radical theological view of the separation of civil and religious realms through his doctrine of "the two kingdoms." He could reduce his teaching virtually to an aphorism: God's Gospel ruled in the churchly realm and his Law ruled in the civil society. To rule the church by the Law or the civil realm by the Gospel would be to bring legalism to the sphere of grace and sentimentalism into the orbit of justice and thus dethrone God and enthrone Satan. In practice, however, the Lutheran Reformation worked to keep its ties to the civil order and was the established religion wherever it predominated in Germany and Scandinavia. In many territories princes actually took on the superintending roles that bishops had known in Roman Catholicism.

Separation of church and state

John Calvin made less of a theoretical effort to separate civil and religious realms. Under his plan Geneva was to be a theocracy in which the saints would rule. God's covenanted community was to be based on his Law, as revealed in the Scripture. Consequently, no detail of civil or community life was too remote, too secular, or too petty to escape inclusion by the Calvinists in the ecclesiastical sphere of supervision or regulation. Zwingli taught a variation of this version, one that asked the Christian to be a zealot or patriot in the civil society—a teaching that he confirmed with his blood, for he was killed in battle in 1531. In the Anglican approach there was also no attempt to separate the civil and religious realms; in England the church was given the mandate to press conscientious matters upon the sovereign and other civil authorities. These established Protestant views were to be subverted or countered by radical Reformers who did want a separation from civil spheres. These views were also constantly revised with the rise of the modern secular state.

Modes of expression of the ideas of the magisterial Reformers and their successors. Protestantism was forced to find means to propagate and sustain itself through time. Reformers had removed many of the inherited props or means and developed, within a century, parallel structures of most of those that had been repudiated along with Roman Catholicism. Lacking papal authority, canon law, and "international" connection with civil authority (as there had been in the old Holy Roman Empire), along with the binding power of church councils or a single philosophy on the basis of which to argue their case, they came up with alternatives or surrogates for most of these, though the new systems were more varied than the at least nominally homogeneous Catholic skein.

Most notable among the structural necessities was the formulation of "confessions," or creedal statements—and Protestants met frequently and regularly to write them—by which they could define their positions for the benefit of their adherents and their opponents. The Lutheran Augsburg Confession (1530), Reformed documents such as the Second Helvetic Confession (1566) and the Westminster Confession (1646), Anglican affirmations such as the Thirty-nine Articles (1563), and Anabaptist confessions such as that of Dordrecht (1632), all gave evidence

Protestant confessions

of the Protestant impulse to define their positions. The Protestant leaders recognized that their movement could not long exist or continue with the fervour and ferment of first-generational impulses.

Confessions of the church appealed to the minds of theologians, the administrative passions of leaders, and the legalistic spirit of those who would impose them as doctrinal standards, but they did not warm believers' hearts. Thus, Protestant leaders had to concern themselves with the affective side of church life in order to hold the attention of masses of people and to give them opportunity to express their faith and life in God. The chief instruments to achieve these aims were liturgies and hymns. The inherited liturgies included much of the Roman Catholic sacramental teaching. As such they were given over too much to an accent on the sacrificial character of the mass and thus had to be purged. Conservative Reformers retained the shell, or outline, at least, of these formulas for worship, though they took great pains to bring both these outlines and the nuances of expression into line with what they considered to be a more evangelical teaching. Since worship is perhaps the chief public expression of gathered Christians, all Reformers had to give attention to its detail.

Liturgies and hymns

Luther initiated the process with his *Formula Missae* ("Formula of the Mass") of 1523, a service that retained the Latin language; but he soon devised (in 1526) a *Deutsche Messe* ("German Mass"), a vernacular and folk expression of greater informality. At about the same time Zwingli was producing a Reformed order with two liturgies for the Word and the Lord's Supper in 1525, soon to be followed by Martin Bucer's work on Psalms and church practice in 1539 and Calvin's *Form of Church Prayers* in 1542 and 1545. The Anglicans were preserving stately forms of worship used in subsequent centuries, chiefly in *The Book of Common Prayer* of 1549 and 1552, and in Scotland John Knox helped formulate Presbyterian worship in *The Forms of Prayers* in 1556.

While Protestant orders were somewhat less ceremonial than the Roman Catholic liturgies they replaced, the human impulse to routinize ceremonies prevailed, and almost everywhere these forms for worship took on a more or less formal character. They differed from Catholicism chiefly in their elevation of the act of preaching the Word of God. Preaching was viewed as the means of grace whereby men were encouraged to repent and accept the grace of God through faith in Christ, just as the sermon was used to shape the community and give guidance. For some this accent on preaching meant a downgrading of the Lord's Supper; for others there was to be a parity, with the sacrament providing a necessary parallel means of conveying grace. Communion "in both kinds," with reception of both bread and wine, prevailed (whereas in the Catholicism of the era of the Reformers the cup was withheld from the laity), and, except in Anabaptist circles, the Catholic practice of infant baptism by means other than total immersion was retained. The Protestants, for the most part, took over existing Roman Catholic church buildings for worship, or they met in academic or civil halls or homes; but as time passed, they also took responsibility for erecting church buildings.

Hymnody played a major role in giving voice to Reformation sentiment, never more successfully than in Martin Luther's "A Mighty Fortress Is Our God," which came to be called "the battle hymn of the Reformation." The Genevan Reformation and the Presbyterian churches tended to prefer simple and sometimes stolid hymnody in the form of rephrased and parsed psalms, such as The Genevan Psalter of 1562. The rejection of hymns and attention to sung versions of Scripture also prevailed in early Anglicanism, not so much because of principle but because of the failure of Anglican Reformers to devote themselves to the propagation of their movement through song. The great tradition of Protestant hymn writing developed later, in the 18th and early 19th centuries.

Liturgies and hymns appealed to the heart and soul, but Protestant theologians also addressed the mind through an impressive outpouring of works in systematic theology and dogmatics. Calvin was the supreme systematizer of first and second generation Protestantism, and his *Institutes of the Christian Religion* (1536) is a classic on even the shortest shelves of Christian doctrinal literature. Luther was, of course, a first-rate theologian, but he made considerably less effort to be systematic, and his scores of volumes of theology usually grew out of comments on issues that agitated him or inspired or disturbed his movement at any moment. His disciple and colleague Philipp Melanchthon, in the *Loci Communes* of 1521, was much more concerned with systematic discipline.

Systematic theologies and dogmatics

In the 17th century the Protestant movement tended toward more rigid doctrinal expressions, as individuals interpreted the confessional statements of the earlier century with an almost fanatic attention to detail. Huge works of Lutheran and Reformed dogmatics poured forth from presses, most of them based on a kind of Protestant reversion to the type of scholastic philosophy that had prevailed in the late medieval period. Leaders in the period of Lutheran orthodoxy were Martin Chemnitz (1522–86) and Johann Gerhard (1582–1637); Reformed Orthodoxy was marked by the scholarship of Theodore Beza (1519–1605) or, in England, men like William Perkins (1558–1602). The ponderous and often lifeless writings of lesser orthodoxists than these were often expressions of internecine Protestant warfare. Debates raged over the sacraments, over the two natures of Christ, over the relations of ecclesiastical and civil realms, and over the part man played in salvation. Almost never did these debates lead to concord, and despite occasional irenic figures, such as Georg Calixtus (1586–1656) or Hugo Grotius, Protestantism was fated to remain divided, at least until the ecumenical movement in the 20th century began to produce new amity and common purpose or assent.

COMMON PRINCIPLES AND PRACTICES OF THE RADICAL REFORMERS AND THEIR SUCCESSORS

The interpretation of Protestantism up to this point has been, with only a few noted exceptions, based on the majority view among the 16th-century Protestant movements. No single term adequately covers the Lutheran-Calvinist-Anglican complex, though "magisterial," "establishment," "mainline," "conservative," and "classical" have frequently been applied to these movements. Of considerable parallel significance was the Protestant activity of another, and even more complicated, cluster of movements, for which also no single term can be agreed upon. Some historians speak of "the radical" Reformation or "the left wing of the Reformation"; others have concentrated on components, such as the Anabaptist-sectarian or the spiritual-mystic or the rationalist-unitarian versions. In almost every case, these were the expressions of the economically and socially deprived classes in the 16th-century societies, though their latter-day heirs have sometimes known or sought the favour of civil authority and social arbiters.

"The left wing of the Reformation"

The "radical" Reformation was radical in that it deliberately chose to repudiate as much as possible of traditional Roman Catholicism in various acts of "restitution" of what it held to be the obscured and eclipsed but true original apostolic church. The "conservative" or "magisterial" Reformation, on the other hand, tended to keep whatever it could of the medieval ecclesiastical tradition and to affirm continuities in the life of the church wherever possible.

The varieties of radical expressions are rich and bewildering. They grew in virtually every Protestant land, sometimes as an extension of the logic of the conservative Reformation but more often as original movements bearing a logic of restitution all their own. The radical Reformation also occurred in Catholic territories, such as Italy, where the mainline Protestant movement never knew much success. In Lutheran circles men like Karlstadt and Thomas Müntzer set out, in Luther's prime years, to shatter much of what he wanted to retain and to carry reform in new directions. Debates over the Lord's Supper and baptism led to new radical movements in Switzerland, southern Germany, and Bohemia-Moravia. In Strassburg a significant group of radicals, including Kaspar Schwenckfeld, Melchior Hofmann, and Sebastian Franck, gathered around 1529. The north of Germany and the Netherlands were havens of early Anabaptism (re-baptism), and in the southern Netherlands Menno Si-

mons spread the movement that has come to be called Mennonitism. In Poland and eastern Europe the radical Reformation often took spiritualist and unitarian (anti-Trinitarian) turns, as it did in Italy. "Radical reform" was also behind some of the Puritan and separatist movements in England. Because they were by nature competitive, free-formed, and varied, it is difficult to generalize about the radical Reformation movements, but some assertions common to major segments are possible, and the study of these movements is important because of the role they were to play in shaping modern Protestantism, especially as it developed in North America.

Generalizations about the radical reform movements

The gathered church. The radical Reformers were united in their opposition to established Protestantism's view of ecclesiastical continuity with the church of Christ in every age. The mainstream Reformers were radical in their rejection of what they regarded to be false teaching in the medieval church and almost never had kind words to say about any of its forms. But they did believe that God had kept a body of faithful teachers and respondents through the millennium or so after what they considered to be the "fall" of the church during the closing years of the Roman Empire, and this view of the succession of believers was integral to their doctrine of the church. Just as emphatic was rejection of this view in radical circles. Some radicals were willing and eager to trace a kind of continuity from John the Baptist down to the 16th century, but it was significant that they found virtually every evidence of true faith only in the sectarian movements that had separated themselves from official Roman Catholicism or that were condemned, harassed, and persecuted by Catholics. Among these were the Waldensians (a medieval religious movement espousing voluntary poverty and lay preaching); the Albigensians, also called Cathars (medieval sect espousing dualism and asceticism); some forms of Spiritual Franciscanism (branch of the Franciscan order espousing poverty); and other reform movements of the Middle Ages. Just as often, however, radicals taught that the true church had died not long after Christ and had to be restored as if from the foundation itself.

The repudiation of continuity was paralleled by rejection of a tie between the civil and ecclesiastical realms. The bond between these two, in the era of the Roman emperor Constantine (d. 337), was viewed by the radical Reformers as the root of the church's fall and later vicissitudes or death. From that experience, it was argued, the church ought to have learned not to let the spiritual infection of political authority prevail nor to permit any one to be regarded as a member of the church without an explicit personal affirmation of faith. In a widely used phrase, the church was to be "the believer's church," made up of assenting and consenting people of decision who chose to respond to God's Covenant. This view appeared in contrast to the view held by those who argued that baptism of infants, who of course could not make personal decisions, conferred church membership and that, thus, virtually entire populations of territories could be members.

The "believer's church"

The keystone of the concept of the believer's church is that people voluntarily choose to be members. No one can be coerced into it nor can one become a member automatically, as it were, through a sacramental act. It was on this ground that infant baptism was condemned by almost all radical Reformers. A result of this accent on voluntarism has been a strong stress on the will of the believer and the giving of a voice to all believers in the questions of the governance and destiny of the church.

The theological counterpart to the teachings that disengaged radical Protestantism from Catholic continuity or established life was the view that no human authority determined modes of church life. The church is Christ's, and not man's. As such it seeks to transcend territorial, racial, and ethnic bounds in theory, even if it is rarely consistent or successful in practice. As Christ's church it is capable of representing him fully in each place, and thus local governance or authority, and even autonomy, was universally stressed.

The radical Reformation almost always restored the sense of an apostolate (missionary outreach), whereas the conservative Reformers had often neglected the importance of a sense of witness and missionary activity, and some of them had even ruled it out from the church's present-day mandate. Anabaptists and spiritualists and "free" church (non-state) advocates tended to be missionary, even if this meant a kind of subversion of established Protestant churches, filled as these were—in radicals' eyes—with unbelievers or inadequate believers.

Relationships between church and state. Churches as disengaged as these were from established structures were in principle devoted to, and in practice successful in adhering to, ideas that called for sharp distinctions between Christian and non-Christian, sacred and secular, religious and worldly life. This is not to say that radicals took no interest in the civil or social realm; they often did, indeed. But they brought a special viewpoint. They were "eschatological"; that is, they almost always were moved by dramatic views of the future, in which Christ would come again or the Kingdom or Day of the Lord would be announced to change everything. Worldly conditions were temporary and were judged by the saints as ephemeral and corrupting, even if they found it necessary to live with or to employ earthly instruments in the meantime. At the same time, for the sake of the freedom and purity of the believer's church, its members advocated separation from the civil realm, permitting no intrusion by civil authorities in church affairs and seeking no direct involvement in administration of the state by ecclesiastical figures.

Concern for the "last times"

Because such a large number of radicals believed that Christ's new order was imminent, they generally took a negative view of most human means of facing problems. Many of them advocated a rejection of warfare and saw in the Gospels a support of pacifist positions. The modern "peace church" witness of Mennonites, Brethren, and Quakers was born of this impulse. Paradoxically, there were other radicals (such as Müntzer) who on occasion saw violence and warfare as legitimate means for them to help hasten Christ's new order.

Church discipline. Separation between the church and the world and membership based on clear commitment made it possible for radicals to insist on higher standards of church membership and stricter means of church discipline than could their magisterial counterparts. Social control was more feasible in these smaller and well-defined groups than in the established churches, and "the ban," as a form of excommunication, was the instrument which supported discipline. The use of the ban meant expulsion from the congregation of believers and, with it, social exclusion. The ban was not conceived as merely a punitive measure; brotherly admonition and discipline were to continue, with the hope that the wayward could be rescued.

Believers' baptism. A special word must be said concerning baptism since it gave its name Anabaptism to part of the movement and was one of the radicals' most dramatic points differentiating them from the rest of Protestantism. Infant baptism, from the radicals' viewpoint, cheapened the standard of church membership and was not clearly designated or foreseen in the New Testament documents that chartered the church. Michael Sattler (*c.* 1500–27), Menno Simons, and Balthasar Hubmaier (1485–1528) led the opposition to infant baptism. Radicals would follow Jesus, who underwent baptism as an adult, and they also would be "buried" (in water) with him, as St. Paul said baptized people would be. "New birth" would come from this act, and the reborn believers would restore the church.

Rejection of infant baptism

Doctrine of the ministry. The concept of ministry was also changed more drastically in radical groups than in the more established Protestant circles. When priests became Lutherans, Calvinists, or Anglicans, there could be a rather subtle transition in their calling. The Anabaptists and spiritual Reformers, however, wanted a clean break with the past. The minister was viewed chiefly as a prophet, not as priest. As an agent of a new order, anticipating Christ's fulfilled Kingdom, he was not to care about earthly prerogatives or routines. Some men, such as Menno Simons, believed that the only way to take on the new ministerial vocation was to repudiate their Roman Catholic ordination. But such conversions from Catholic to radical clergy were rare, and the radical wing of the

Reformation more frequently expressed its views on the ministry by simply placing a low valuation on ordination. The classical Reformers wanted university-trained, theologically expert ministers. The radicals, on the other hand, permitted laymen to be ministers: leaders such as Kaspar Schwenckfeld and Konrad Grebel (*c.* 1498–1526) were probably never ordained.

The suffering of persecution. The radical Reformation and the believer's church were made up of people who were prepared to suffer for their faith at the hands of both civil authorities and Catholic and other Protestant ecclesiastical leaders. The story of the rise of Anabaptism is one of persecution, of exiles and fugitives, and of a pilgrim church. The story of the rationalist form of Reformation, as in the case of Michael Servetus (anti-Trinitarian; *c.* 1511–53), often ended in something that can be called a Protestant Inquisition, in which men died for their ideas. Though some erratic personalities may have revealed a desire for martyrdom, more characteristic were those who upheld the idea of patterning one's life after Jesus, the great example. He had not known status or security and was eventually condemned to death; how could his true followers evade a similar path?

Doctrinal variations. Doctrinal varieties among the radicals were many, and it is hardly fair to cluster the various emphases. Certain features stand out, however. First of all, the role of Christ, central to Protestant Christianity, shifted subtly but significantly. The emphasis on Christ's priestly work, in which he brought sacrifice for men before the altar of God, was displaced by a new regard for his prophetic role. He had thundered against the powers of religion and civil society, against established forces, and against the rich; so would his followers. He was seen less as an agent in a divine-human transaction culminating in death on the cross as a sacrifice, and more as the supreme exemplar and leader.

Christ as the prophet and example

The radicals spoke critically of scholastic philosophy and the intellectualized theology built upon it, and therefore they displayed a distaste for the more arcane expressions of classical theology. Faustus Socinus (1539–1604) in Poland and Michael Servetus in Strassburg became shapers of modern Unitarianism. They believed that the doctrine of the Trinity was an unscriptural abstraction and that simple monotheism could best be protected if Christ were not defined as a full expression of the Godhead. Unitarianism remained a distinctly minority emphasis in the radical Reformation. The Bible was usually highly regarded, but whereas the magisterial Reformers tended to see it in the context of tradition, the radicals stressed contemporary personal experience and often allowed for or claimed new special revelation.

Protestantism's influence in the modern world

INFLUENCE ON NATIONALISM

Protestantism eventually became the majority faith throughout northwestern Europe and in England and English-speaking America. From there, in the great 19th-century Protestant missionary movement, it was carried into all parts of the world, joining Roman Catholicism as a minority presence in Asia and Africa and at the same time also establishing beachheads in largely Catholic Latin America. It is impossible to separate Protestantism from the general history of the North Atlantic nations, where it was firmly established for centuries and where its "free" churches or, after "separation of church and state," its voluntary churches, still predominated.

Thus it is possible to speak of Protestantism's contribution to modern nationalism. It shared in shaping this force initially by helping bring to an end the Holy Roman Empire, which was disintegrating already at the time of the Reformation but which finally collapsed in 1806. The old *corpus Christianum* (Christian body; *i.e.*, Christian society) did not survive; the presence of Protestantism spelled the doom of an international, transterritorial, unified Christianity under one head. Protestantism's desire to cultivate literacy and to spread regard for the vernacular served to remove the Latin linguistic bond of older Christendom and to encourage the rise of national boundaries based on languages. All but the radicals tended to make much of loyalty to the existing state, and Protestants often provided an ideological base for each new state as it rose to self-consciousness—as was the case in Prussia or in the United States.

INFLUENCE ON THE ARTS

Protestant attitudes toward the arts have been ambivalent and therefore have produced mixed results. For the most part, Reformed and spiritualist Protestants have been uneasy about the arts, fearing lest the symbol be confused with the reality—and lest, therefore, the symbol be idolized and the reality forgotten. Thus Calvin and Zwingli found little place for the visual arts, though Luther showed interest and was a friend of some artists of his time, including Lucas Cranach. Luther also revealed a more affirmative attitude toward music than did the Swiss Reformers, though through the centuries most of Protestantism encouraged the use of music. When Protestant historians want to point to past glories in the aesthetic realm, they cite men like John Milton in literature, Rembrandt in painting, and Johann Sebastian Bach in music, though such a group has few heirs in more recent centuries.

ECUMENICAL CONCERNS

While it is clear that Protestantism by nature had to allow for great variety, not all Protestants have rested content with division and separation. They were caught between two biblical mandates. One commanded them to seek the truth and not to express full fellowship with those they considered to be in error. The other stressed the values of Christian unity as a witness in the mission of the church and as a foretaste of the eschatological, or fulfilled, life of Christians when, all agreed, they will all be one. The ferment of the 16th century and the doctrinal formulations of the 17th century led to ever-increasing divisions and hardening of lines or positions. The 18th-century Enlightenment, which in its British and German forms lived off and fought against Protestantism just as the French forms similarly related to Roman Catholicism, tended to breed a spirit of consensus. The Enlightenment placed an exceedingly high value on toleration of differences even as its spokesmen worked for agreement on doctrines based on a search for what they viewed as natural in reason and law. Such a tendency inevitably served to minimize doctrinal differences among Protestants.

Protestant discontent with division and separation

The 20th century, however, has seen more effort toward producing consensus than did the previous three and a half centuries. The modern ecumenical movement, today thoroughly Protestant-Catholic-Orthodox in its outlook, was first born and institutionalized on Protestant soil by men who saw the mission of the church frustrated by competition and division. Beleaguered, huddled together like sheep in a storm, to use a familiar picture, they sought each other's company.

At the same time modern transportation and communication techniques effectively reduced their world and made uniting symbols accessible. A theological recovery was fused with a new vision of common tasks to produce a Protestantism eager for common statement and often for common action in an ecumenical era. The ecumenical movement has led to denominational mergers and to conciliar organizations, on both confessional and transconfessional lines.

In the meantime, the openness of Roman Catholicism, particularly exemplified in the career of Pope John XXIII (1881–1963), led to new amity and concord between Protestants and Roman Catholics. In the last third of the 20th century both of the old warring parties, without formally repudiating their polemical positions of the 16th century, have tended to move beyond their terms and to find new bases for meeting. Modern Catholic biblical commentators speak in what sounds much like Protestant terms of grace and faith. Protestants have new appreciation for a Roman Catholic view of the interconnectedness of the components of the church. More and more, Protestants view the Scriptures as rooted in a tradition and tradition as rooted in the Scriptures. Thus they have a new sympathy for Catholic views of tradition—even as some Catholics

criticize unreflective responses to ecclesiastical authority on coercive lines in their own communion. Protestants and Eastern Orthodox Christians, generally spatially quite separated, have begun to understand each other through agencies and organizations such as the World Council of Churches.

CONCLUSION

"The end of the Protestant era"

In the latter half of the 20th century many heirs of Protestantism, among them the philosophical theologian Paul Tillich, began to speak of "the end of the Protestant era," or of the times as being "post-Protestant." This does not mean that they all wavered in their faith in Protestantism's general witness. Tillich, for one, argued that "the Protestant principle" of prophetic criticism had to be included in any authentic expression of church life and that it was a genuine value in the secular world. But these thinkers believed that the cultural dominance of Protestantism on its own historic soil was waning.

From the Renaissance onward and increasingly during the Enlightenment the adherents of Protestantism saw their thought-world repeatedly challenged on many fronts. During the 19th century, with the rise of industrialism and urbanization, a changing world presented new problems to societies and cultures shaped by traditional Protestantism. Meanwhile, ideologues, some of whom were avowed "god-killers," rose up on Protestantism's territory to challenge its deepest beliefs: the economic theorist Karl Marx, the evolutionary theorist Charles Darwin, and the philosophical nihilist Friedrich Nietzsche, to take only three examples, were thoroughly at home with the Protestant experience and were able to use it as a foil to develop many of their own views.

In the 20th century Protestantism has become uncertain about its "foreign mission" of expansion in a postcolonialist, anti-imperialist world. The modern appreciation for values in non-Christian religions has led many Protestants to adopt positive attitudes toward these at the expense of the desire to extirpate or displace them with an expanding Protestantism. Totalitarian forces, particularly in Nazi Germany, absorbed some Protestant emphases and changed them beyond recognition, or they persecuted those Protestants who radically opposed suppression.

The attractions of modern life, secularization, and a crisis of faith all have contributed to a general Protestant decline, beginning with a measurable decrease in church membership, first on the Continent in the 19th century and then in England around the turn of the century. Therefore, while huge majorities of the population (as in Scandinavia and England) are baptized members of established Protestant churches, only a small percentage of these are attendants at worship services or responsive to the disciplines and mandates of the church. Those who use church attendance and support of ecclesiastical appeals as indicators of Protestant fortunes unite with those who see that Protestant dogma no longer defines belief, and its divisions no longer excite Western man—and then note the end of the Protestant era.

On the other hand, Protestantism is deeply integrated into so many elements of Western culture that it can be expected to continue to assert subtle influence. It has experienced ebb and flow or revival and decline periodically and now may be going through an extended period of decline. Yet even to speak in these terms may betray a Western provincialism that does not do justice to major trends. Countering all phenomena that elicit words about decline are at least two forces. One is the strength of conservative and evangelistic forms of Protestantism: Pentecostalism, Evangelicalism, and Fundamentalism. While historical antecedents of these movements were often world-denying, sectarian, and withdrawn, late 20th-century versions include men and women eager to shape their surrounding culture. They give evidence that they may be doing so, or may begin to do so, in forms that not many would have foreseen a few decades ago.

The other compensatory force is the growth of Protestantism in sub-Saharan Africa, Latin America, and many parts of Asia. Some of these new Protestant churches have begun to take indigenous forms that have little to do with the forms initially introduced by missionary forces and to witness far beyond Protestantism's conventional Western bases. (M.E.M.)

THE MAJOR PROTESTANT DENOMINATIONS

Lutheran churches

Lutheran churches are those religious bodies that trace their distinctive interpretation of the Christian Gospel to Martin Luther and the 16th-century movements that issued from his reform. They take their place alongside Anglican and Calvinist communions to make up one of the three major branches of Protestantism.

The Lutheran churches, originally in Germany but quickly spreading to Scandinavia, did not wish to be called after their founder. He had seen his work as an evangelical (*i.e.*, Gospel-centred) reform within the Western Catholic church. The name Lutheran came from opponents of Luther and his reforms, but the epithet eventually came to be turned into a badge of honour among partisans of the reformer's interpretation.

Still, many of the leaders attempted to adopt other terms such as "Evangelical," which has subsequently become part of the official name of the church in various nations and territories. Others preferred, and prefer, to be called

The Church of the Augsburg Confession

The Church of the Augsburg Confession, a title that recalls the Lutheran document presented by evangelicals to the emperor at the Diet of Augsburg in 1530. In the 20th century many have chosen to speak of their church as an "evangelical Catholic" movement, yet "Lutheran" they became and remain.

For several decades after 1530 this oldest and largest Protestant body hardly broke the European geographic bounds that were set for it. It was a negligible force in Presbyterian Scotland, Anglican England, the Reformed Lowlands and Switzerland, or in Catholic France, Spain, and Italy. There were early Lutheran movements in central Europe, as in Hungary, where the Reformed came to dominate in 1543, and in Transylvania. But Lutheranism prospered most in the many territories that were eventually to make up modern Germany and then the northern lands: Finland, Sweden, Norway, Denmark, and Iceland.

From this significant but well-defined territory, Lutheranism moved with substantial numbers into North America after the 1740s and in the 19th century from European and North American bases into much of the rest of the world. Still, most of its theological, intellectual, cultural, and political expression as well as the major trends in its development are best measured from northern Europe and especially from the German territories that it shared with Reformed and Roman Catholic Christians.

Justification by grace through faith

Lutherans claim to see their movement centred in the understanding that, thanks to the saving activity of God in Jesus Christ, they are themselves "justified by grace through faith." Early Lutherans invoked this theme against both Catholic and Reformed Christianity, both of which, though on differing grounds, they professed to see stressing salvation in part through good works or moral earnestness. This would be an endeavour to help the believer make a claim upon God and thus, thought Lutherans, would deprive Christians of the security of faith and would arrogate to human beings activities that belonged only to God. In Lutheranism the bond between God and the redeemed was entirely at God's initiative and through God's grace. The believer trusts this God. With most other Protestants, Lutherans based their teachings not on churchly authority but on the divinely inspired Bible.

HISTORY

The post-Reformation in Germany. A generation after Luther the churches and territories that followed him

theologically took part in a diet that produced the Peace of Augsburg (1555). This action accepted the principle *cuius regio, eius religio,* which meant that whoever governed a region determined its religion. It was a far remove from the modern policy of separation of church and state, for which the reformers could have taken little credit. Instead there developed what is often called "territorialism" in confessional life. Where once there had been one empire and one church, there now were numerous nations or principalities, each with its own official church.

The confessional church on territorial grounds compromised the very nature of a confession, which was to have been a freely accepted creedal statement. When a ruler chose to take his people into Lutheranism, they had to follow or suffer penalty, no matter what their deepest convictions. Yet if they had to accept territorial conformity, they were free to change the character of its faith through time. Lutheranism passed through numerous phases and appeared to be quite different in each.

Conventionally Lutheranism has been seen as having passed through a series of rather clearly defined stages, with movements of thought and practice termed Orthodoxy, Pietism, the Enlightenment, and the like. This convention points to complex realities with sufficient accuracy and thus serves well for understanding the complex movement after more than four and a half centuries.

Orthodoxy. Orthodoxy came to dominate first. Whatever the ordinary people who rejected Catholic "work-righteousness" and were attentive to "grace" and "faith" were thinking, their pastoral and professorial leaders—and, for that matter, their princes, for politics was much involved in the new definition—grappled with orthodoxy in doctrine and practice. They did so perhaps out of a Germanic love for order and precision and more clearly because the first generation of Lutherans had left them with a rather unstable mix with which they had to deal.

One strand, most faithful to Luther himself, was more ready to live with the risky faith and the paradoxes that coloured his preaching and life, but it then converted the experience of such faith into rather rigid doctrine. In the second generation these "Gnesio-Lutherans" or "Genuine Lutherans," who gathered at centres like the university at Jena, followed impulses to bring order and precision to bear on the thought and the creeds they had inherited. In the eyes of many historians this party lost much of the drama and dynamic of Luther's witness. But it was also a belligerent faction, one that brought passion to its claims on orthodoxy.

Philippists

Its opponents were called Philippists, after Philipp Melanchthon, the chief scholar at Luther's side and the author of the Augsburg Confession itself. Melanchthon was a humanist with a pacific outlook, who appeared to be a compromiser, as did his intellectual heirs, in the eyes of the Gnesio-Lutherans. They accused Philippists of "synergism," the contention that human beings could cooperate in the work of salvation. They also saw Reformed tinges in the Philippists' doctrine of the Lord's Supper.

By 1577 leaders of the Lutheran parties had agreed on a statement called the Formula of Concord and in 1580 sealed the Lutheran confessions in the *Book of Concord,* which has been respected ever since and holds varying degrees of authority in Lutheran churches. In the century following, Lutheran scholars, led typically by Johann Gerhard, wrote the multivolumed *Loci Theologici,* code names for books that stressed a proper doctrine or place for all Christian teachings. Scholastic in style, these books of dogma characteristically began with arguments proving the existence of God and the full authority of the verbally inerrant Bible.

Pietism. Orthodoxy bred reaction, and this Orthodoxy, soon perceived as rather sterile, did not satisfy pastors and people alike. From the 1670s into the 1760s Pietism flourished, originating again at universities, such as Halle, and spreading from thence to other schools and congregations. As the name implies, this movement stressed the piety of the individual or of the small groups of Lutherans who gathered as smaller "churches within the church" for prayer, Bible reading, moral scrutiny, and works of charity. Philipp Jakob Spener, a leader among the Pietists, wanted to remain orthodox but nonetheless engaged in criticism of what Pietists saw to be the barren larger Lutheran church of which they remained a part.

Pietism downplayed doctrinal definition and led to movements that helped make up the 18th-century German Enlightenment. The tendency of Pietism was to minimize supernatural and miraculous elements in Christianity and to stress reason and morality. Although ordinary worshipers seem to have been sustained by the Scriptures, hymns, and liturgies that retained these elements, scholars of Lutheranism initiated radical theological traditions that have characterized German universities ever since.

19th-century developments. In the 19th century the Enlightenment lived on in more romantic forms that gave a greater place to emotions. Some of these were philosophically idealistic, some Germanically nationalistic. At least two schools should be singled out. One, in the tradition of G.W.F. Hegel, saw Christian development against a huge screen of "thesis" and "antithesis," and under the great historian F.C. Baur at Tübingen posed Hebraic versus Hellenic, Catholic versus Protestant motifs and movements. This school began to cast doubt on fact and event in history and soon began to speak in terms of biblical myth. Out of it issued radical movements that led to theological extremism, as found in D.F. Strauss's *Life of Jesus Critically Examined,* an essay on the impossibility of writing a life of Jesus.

Higher criticism of the Bible

The second school, in a Neo-Kantian spirit, stressed biblical fact and event and issued in a quest for the historical Jesus. Under Albrecht Ritschl a number of Lutheran and Reformed theologians developed a theology that stressed morality and the will. Out of their efforts came the well-known late 19th-century German liberal theology with its devotion to Jesus as teacher and doer of good.

Thus a kind of intellectual schism emerged among 19th-century German Lutherans; one school pioneered in the application of historical methods to biblical studies, developing what came to be known as higher criticism of the Bible, while the other, in a conservative reaction, established more pietistic training centres for the clergy.

Events on the political level also caused schism. Frederick William III's successful efforts to form a Prussian Union church with the Reformed in 1817, while at first meeting with approval, soon prompted a critical reaction. Dissension came partly over Frederick William's new church order, according to which the territorial regent was placed into the position of chief bishop of the church, not because he was the head of state but because he was the person of highest status in the congregation. This decision was in violation of the Lutheran tradition regarding the relationship of church and state, and "confessionalists," who that year celebrated the third centenary of the Lutheran Reformation, promoted "back to Luther" movements. Some Lutherans refused to become part of the Union and formed the "Old Lutheran" church; others chose emigration.

Lutheranism and civil obedience

Much of Lutheranism, however, remained obedient to the civil order, perpetuating a tradition begun by Luther himself. After the break with Rome, Luther had remained closely tied to the growth of German nationalism and welcomed the protection of German princes. Many of these became, in effect, "prince-bishops," with considerable church power. The Lutheran churches came to be established by law and supported by taxes in Scandinavia and in many parts of Germany. Further, Luther had a fear of anarchy and a predisposition to grant considerable power to the state as an instrument of order. While he himself was "civilly disobedient" in the face of the emperor in the 1520s, neither his theological vision nor his personal inclination led him to endorse revolution or radical critique of the state. Lutheranism, therefore, was generally obedient to the civil order, and its clergymen were often cast in roles that made them seem to be lower-level civil servants.

Lutheranism in eastern Europe and Scandinavia. Lutheranism was carried rather quickly from Germany to Bohemia and Austria, Poland and Hungary (where it remained a minority party), and then to the Scandinavian nations. There much of the proselytizing impulse

came through universities, especially when scholars from north European schools studied in Germany and carried Lutheran ideas back north. Already in the 1520s and 1530s Denmark made its move to Lutheranism under the influence of its successive kings. A former monk, Hans Tausen, who became a convinced Lutheran, was the major theological influence.

Norway was in many ways a dependency of Denmark when about 1525 King Frederick I encouraged Lutheran preaching in Bergen. Norway followed Denmark into the orbit of Augsburg Confessionalism, as did Iceland, by importing Lutheranism from Denmark. Sweden's political restlessness similarly led that emerging nation to turn Lutheran under King Gustav I Vasa after 1523. Olaus Petri, a student at Wittenberg during the years of Luther's reform, brought conviction and passion to the task of spreading Lutheran ideas in Sweden. Although the King and Petri or his reformer colleagues often were in conflict, Petri's reformation ideas prospered. In Finland, Michael Agricola, still another former Wittenberg student, translated the New Testament and books of worship and helped Finland make a transition to Lutheranism before his death in 1557.

Scandinavian universities and churches generally followed developments parallel to those in Germany, albeit with a special accent on 19th-century Pietist revivals in Norway and Sweden that would make their impact through immigration to America. The major drama of Scandinavian Lutheranism occurred in 19th-century Denmark. There N.F.S. Grundtvig represented a romantic "folk-church movement," and Hans Martensen a kind of official church idealism. Søren Kierkegaard issued a devastating critique of both in the name of an existentialist encounter with Jesus, bypassing established and, he thought, dead Christendom.

Lutheranism in North America. American Lutheranism inherited both the Orthodox Confessionalism and the Pietism of the Continent. Enlightenment rationalism, however, was rarely advocated. The Lutherans began to go in large numbers to New York, the Carolinas, and especially Pennsylvania in the 1740s, where they were often gathered by the Lutheran patriarch Henry Melchior Muhlenberg. Many who went to America were poor, and some arrived as exiles or protesters against imposed conformity. Handicapped by their relatively late arrival and the fact that they spoke languages other than English and tended to live in rural enclaves, they had less impact on politics and culture than might be assumed, given the record of their counterparts in Europe.

Because Lutherans came from many nations, spoke many different languages, were propelled by a variety of motives, and were guided by leaders either unaware of or competitive with one another (as, for instance, "Pietists" versus "Confessionalists"), they tended to be isolated. As they became aware of one another, they became contentious. In the mid-19th century, for instance, a shaping influence was Samuel S. Schmucker, a Gettysburg (Pennsylvania) Seminary professor, who advocated Americanization and cooperation with the Reformed evangelical churches. Partly in reaction, the more Lutheran Confessional-minded Charles Porterfield Krauth, also at Gettysburg, stressed Lutheran distinctiveness. More militant in his defense of 17th-century orthodoxy was the great shaper of the Missouri Synod, Carl F.W. Walther, who was president of both the synod and its principal seminary, Concordia, at St. Louis. Walther advocated a policy that forbade Lutherans from communing or praying together if their synods were not in complete doctrinal agreement with one another.

Divisions among Lutherans

The majority of American Lutherans were of German descent and were often suspect in the Anglo-Saxon milieu. They did not, in the main, support Prohibition and other Protestant social causes. Some retained the German language and were, generally falsely, suspected of German loyalties during World War I. In response they set out to prove themselves superpatriots and after the war they became more Americanized than before.

Through two centuries American Lutherans gathered about 8,000,000 Christians into scores of church bodies. In the 20th century, especially after 1918, they had a tendency to merge, and 5,500,000 of them united three bodies to become the Evangelical Lutheran Church in America (ELCA) in 1988. The largest non-ELCA group was the 2,500,000-member Lutheran Church–Missouri Synod, which saw trends toward liberalism and ecumenical expression in the larger body that it did not welcome. Canadian Lutheranism, about 300,000 strong, is divided chiefly into two bodies parallel to the ELCA and the Missouri Synod in the United States and is strongest in Ontario and the Western provinces.

In the United States the ELCA constituency is chiefly northern. One large wing thrives in and around the states where Lutheranism first arrived: Pennsylvania, New York, Virginia, and the Carolinas. Another resulted from 19th-century immigrations from Scandinavia and Germany to the upper Midwest, with Minnesota having the largest number. The Missouri Synod has less strength in the East and is strongest also in the upper Midwest and around the Great Lakes. After World War II, partly through population mobility and partly through conscious efforts to found new congregations, Lutheranism came to be more of a national presence.

Lutherans in America developed an extensive network of seminaries, beginning at Gettysburg and Philadelphia, when it was seen that they could not depend upon clergy sent by agencies in Germany and Scandinavia. Most Lutheran groups founded colleges of their own, many of which remain strong church-related liberal arts institutions. The Missouri Synod and a smaller and still more isolated Wisconsin Synod established a flourishing network of parochial elementary and, in some cases, high schools. Originally these were shaped by a defensive mentality bent on sheltering the young from public school life. In more recent decades the schools have tended to attract non-Lutheran constituencies and to see themselves less as competition than as complements to public schools.

Educational institutions

Lutheranism in the 20th century. In the 20th century, after the moral collapse of World War I, Lutherans along with the continental Reformed reacted against the humanistic liberalism that they felt failed to do justice to the radical difference between the divine and the human. There was a revival of biblical theology, often on existentialist grounds, as in the work of Rudolf Bultmann at Marburg, Ger. Historians like Karl Holl helped inaugurate a Luther renaissance. In Sweden Gustaf Aulen and Anders Nygren inspired theological revivals that stressed certain profound motifs in Christian and Lutheran thought.

To its shame, much of Lutheranism was silent or even concurrent when Hitler came to power and provided some intellectuals and some pastors for Nazi Church ("German Christian") leadership. At the same time, some of the neo-orthodox Lutherans joined with the Reformed in 1934 to establish a Confessing Church. In Martin Niemöller and Dietrich Bonhoeffer in Germany, Kaj Munk in Denmark, and Bishop Eivind Berggrav in Norway, it brought forth anti-Nazi heroes and, in some cases, martyrs.

"Mission fields" around the world—established in the 19th century from the Continent and from North America chiefly in the first half of the 20th century—later became younger churches. After the middle of the 20th century many of these showed a vitality that was disappearing or had gone from churches in Europe. In South West Africa/Namibia and elsewhere in Africa thousands gathered to worship and to use their churchly vision for meeting their political problems. Thus in South West Africa/Namibia, a territory under the domain of South Africa, Lutheran church members participated in the leadership of revolutionary organizations, contributing to the establishment of an independent Namibia in 1990. The Lutheran Christians thus developed a pattern different from the characteristic Lutheran one of passivity in politics or conservatism in the face of change.

In its northern homelands Lutheranism was anything but an expanding force after the mid-20th century. Yet, through the Lutheran World Federation and countless vital agencies and institutions, Lutherans continued to find ways of expressing the faith they had heard Martin Luther proclaim. They also took responsible parts in the formal ecumenical movement of the century in their endeav-

our to stress both their "evangelical" and their "catholic" sides. (M.E.M.)

TEACHINGS

Book of Concord

Lutheran Confessions. The official teaching of the Lutheran churches is that of the *Book of Concord* (1580), which contains the three ancient creeds (Apostles', Nicene, and Athanasian), the Augsburg Confession, the Apology of the Augsburg Confession, Luther's Schmalkaldic Articles, Luther's Small and Large Catechism, Melanchthon's "Treatise on the Power and Primacy of the Pope," and the Formula of Concord. Of these Lutheran symbols only the Augsburg Confession and Luther's Small Catechism are accepted by all Lutheran churches. No general confessions of faith were adopted after 1580 by the Lutheran churches, although other doctrinal statements have served a confessional purpose for particular churches.

Partly because of the circumstances of its composition and partly because the Reformers understood their work to be a restoration of Christianity amidst contemporary corruptions, the Augsburg Confession emphasizes the continuity of the Lutheran teaching with the ancient Christian Church.

Justification. The teaching centres in the Gospel, or "justification": the doctrine that men "are justified freely on account of Christ through faith when they believe that they are received into grace and their sins forgiven on account of Christ, who by his death made satisfaction for our sins"; God "imputes [this faith] as righteousness in his sight" (Augsburg Confession, IV). Modern Lutheran theologians, among them Rudolf Bultmann and Paul Tillich, have applied this doctrine about grace to doubt as well as to guilt and have called attention to the change in the cultural and religious situation since the 16th century. Thus, Tillich interpreted justification through faith as a person's accepting his having been accepted in spite of unacceptability.

Law and Gospel

This doctrine ("the article by which the church stands or falls") provides the key for understanding the Bible (Apology, IV. 3–5) as a book that has two kinds of content—law and promises. Law demands a perfect inward as well as outward obedience to the divine will, which reason can never achieve. As such it drives men to despair, but the despair is conquered by the promise that God justifies the unjust man. This means that in Lutheran theology, in the act of being justified before God, the human being recognizes no positive or constructive role for the law. "The law always accuses," always destroys what the sinner had thought would impress God. God then effects a new creation by producing the new and justified person in Christ. Theologically, the doctrine of justification gives a Christocentric (*i.e.,* what honours Christ) stress and a practical (*i.e.,* whether afflicted consciences are consoled) emphasis to the other articles of faith.

Human nature. Lutheranism has a doctrine of human nature that defines the natural state as one in which humans do not fear or love God and are self-seeking. Human beings have freedom of will concerning the outward observance of laws (civic righteousness) but not before God (where they are inevitably unrighteous). They have a knowledge of God but not a true knowledge (they think, for example, that righteousness is what God has rather than what God gives).

Similarly, the meaning of predestination is to be sought not in the hidden counsel of God but in his revelation (Formula of Concord, Epitome XI). Lutheran teaching differs from the Calvinist double predestination by accepting the formal inconsistency of saying that believers are predestined to salvation without saying that unbelievers are predestined to damnation, for the purpose of the article on predestination is to console the troubled conscience. The mechanism of predestination has been the subject of controversy within Lutheranism (whether the decision of God is made "in view of faith"), but the basic position expressed in the symbols has been maintained.

Church, sacraments, and ministry. In opposition to the claim that the Roman Catholic Church was the only legitimate ecclesiastical organization, as well as to the biblicist demand to restructure the Christian Church according to the New Testament pattern, the Augsburg Confession (Art. VII) defines the church as the "congregation of saints [believers] in which the gospel is purely taught and the sacraments rightly administered." "Gospel" is interpreted to mean that God justifies believers on account of Christ, not on account of their merits (Augsburg Confession, V). Right administration includes the practice of communion under both kinds (bread and cup). For the unity of the church it is sufficient to agree concerning the gospel and administration of the sacraments. This is the formula Lutherans use to build ecumenical relations with other churches. But it also brings difficulties, for the meaning and degree of "agreement" are always difficult to define and measure.

The church

Luther regarded the church as essentially hidden or invisible. Although it is as weak and sinful an institution as any other one, it is possible to believe that God works in and through the church because it is founded on God's word.

This doctrine has undergone transformations since the 16th century. Orthodoxy and Pietism understood the invisibility of the church to mean that only God knows who among the assembled people are true believers (the invisible church as distinguished from the visible congregation). In the 19th century a sacramental-institutional conception was formulated by some Lutheran theologians (*e.g.,* the German leader Wilhelm Löhe), a congregational conception by others (*e.g.,* the conservative American C.F.W. Walther), a national or folk conception by still others (*e.g.,* the Danish leader N.F.S. Grundtvig), and a historical-evolutionary conception (*i.e.,* the church as the first actualization of the Kingdom of God to be progressively realized in history) by others. Though these differences radically divided the Lutheran bodies in the 19th century, particularly in America, today Lutherans tend to live with different conceptions of church polity without letting such matters divide them.

The sacraments

Of the three sacraments (baptism, Lord's Supper, penitence-absolution) recognized by Luther and the Lutheran confessional writings, which are called symbols, in the *Book of Concord,* the Lutheran churches generally hold to two by combining absolution in part with baptism (daily repentance is the repeated actualization of baptism) and in part with the Lord's Supper (confession and absolution). The criterion used in determining the number of sacraments was that they were actions instituted by Christ and connected with God's promise (Apology XIII). The symbols do not define the relation between word and sacraments except to say that they come together, and both have the effect of creating and strengthening faith. This is a rejection of the view that sacraments are effective *ex opere operato* (operative apart from faith) and that they are only memorial actions.

The Formula of Concord's teaching on the Lord's Supper is that Christ is bodily present "in, with, and under the bread and wine" (Solid Declaration, 35 ff., adopting Luther's terminology over Melanchthon's "with the bread and wine"). The Formula of Concord left open the question whether Christ is present in the sacrament because he is present everywhere, as one party contended, or whether he is present in the sacrament because he chooses to be.

In the 19th century some Lutherans (*e.g.,* Gottfried Thomasius) distinguished word and sacrament by saying that the sacraments are intended for man's natural life as the word is for his conscious personal life. This view in some cases was carried so far (*e.g.,* in Martensen and Friedrich Stahl) as to subordinate the word (as the presentation of salvation) to the sacrament (as the participation in salvation).

The ministry is conceived of as a service in word and sacrament but not a special status. Every baptized Christian is a priest by status (universal priesthood of believers), but the public preaching and administration of sacraments devolves upon "rightly called" ministers, who are priests by office.

Church and state. The Lutheran churches generally have understood the relation of church and state on the basis of God's two ways of ruling in the world (two kingdoms). Through the "laws, orders, and estates" of the world God rules by compelling external obedience through

fear and threat of punishment. Through preaching and sacrament he rules in apparent weakness by converting the human heart. This conception has provided Lutherans with a basis for understanding the constitutional separation of state and church in the United States.

The two domains of power and grace are interdependent because the word alone cannot preserve peace and justice—the civil government must even protect the freedom of the church to proclaim the Gospel—and civil power cannot effect salvation. Lutheranism has rejected the view that civil power is of itself evil, as well as the view that civil obedience has merit for salvation in the sight of God.

The two domains of power and grace

To define a citizen's relation to government one may say that in ordinary circumstances a Christian obeys the powers that be (except in matters of faith) as the agent of God's rule. But if a law or government is unjust, a Christian has the right and duty to resist it, passively accepting the consequences of disobedience for himself but actively defending his fellow man against that law or government. If the government is tyrannical, a Christian not only resists but rebels. Those Christians who also are holders of civil power have an obligation to resist and oppose misuse of such power by other rulers (as the territorial princes opposed the emperor in Luther's time). Lutheran scholars are not in complete agreement, but many would associate this view with Luther himself. In the 20th century it was developed by figures like the Norwegian bishop Eivind Berggrav, who resisted the Nazis.

In the 19th century the romantic view of the national state as expressing the spirit of a people was widely influential, but later it became suspect because of the demonic character of nationalism in the 20th century.

Scripture and tradition. The Lutheran Confessions, unlike the Reformed, have no article on Scripture, although the Formula of Concord does designate the Scriptures of the Old and New Testaments as the "sole and most certain rule" for judging teachings and teachers.

Toward tradition the attitude of Luther and the confessions was conservative; they retained whatever did not conflict with the Gospel of justification through faith. They viewed the written tradition of the church fathers as useful for interpreting the Scriptures but not as a source or norm of teaching. Some Lutheran theologians in the 19th century developed an organic view of the relation of the two (*i.e.,* Scripture contains a truth that is unfolded in the course of history) not unlike that of the Roman Catholics Johann Möhler and John Henry Newman.

Ethics. Lutheran teaching on ethics is determined by the perspective of the two kingdoms—the domain of law is not to be confused with that of the Gospel—and by the relation of faith and love implicit in justification. Works of love are the result, rather than the condition, of faith. Human beings have freedom from concern with self by the act of God and are enabled to direct their concern to other human beings. The works a person is to do are specified in part by his status in the world (as parent, ruler, subject, and other roles). Though early Lutherans thought of status in more natural terms (as "orders of creation"), recent Lutherans have given the concept a historical reference (*e.g.,* a person's particular destiny). A person's calling is to do well whatever his status requires. A second factor defining the works a person is to do is the concrete need of fellow human beings.

Man's works

Controversies. Lutheran teaching has been shaped in part by the theological controversies in its history, almost all of which were at one time divisive. They had to do with such questions as the relation between divine and human agency (synergistic controversy, predestinarian controversy); whether works are indifferent, necessary, or dangerous for salvation (antinomian controversy, Majoristic controversy); whether in a state of confessional disagreement any questions are neutral (adiaphoristic controversy); what the nature of the sacramental presence is; whether the divine power resides in the Scriptures only when they are being used or also apart from their use (Rathmann controversy); what are sufficient grounds for church unity (syncretistic controversy); and whether God's election of believers is made "in view of faith" or not (predestinarian controversy).

WORSHIP AND ORGANIZATION

Liturgy and music. The worship service also was affected by the theology of the Reformers. Luther's "German Mass" of 1526 reflects changes that began about 1523. Apart from shifting the emphasis from sacrifice to thanksgiving, Luther's chief innovation here was to take the words of institution out of the framework of prayer and make of them a proclamation of the Gospel. This change has been preserved to the present day, although there is now a tendency to put the words again into a eucharistic prayer.

Because of the Reformers' emphasis upon the importance of the word, the sermon took an essential place in the service. Preaching is usually based upon a biblical text, a biblical story or doctrine, or a theological theme. Partly in reaction to the 19th century, there is an effort to keep preaching biblically oriented, though not necessarily tied to specified texts.

The term mass, at first retained, is not normally used except in the Church of Sweden (*högmessa*) as a name for the main service of worship. The other minor services disappeared from use during the 17th century, though some have been recovered in the liturgical reforms of the last century. Only Matins and Vespers are used with any regularity.

The basic order of service in most Lutheran churches is the same. It consists of two main portions (preaching and sacrament) in which the Kyrie, Introit, Gloria, Credo, and Agnus Dei are incorporated. Under the impact of the liturgical movement in the 20th century the didactic emphasis has given way to an emphasis on celebration in the service. Liturgical revisions (Swedish order of 1942, German of 1954, in the United States in 1941, 1958, and 1978) have brought an even greater uniformity in the basic order. They have also restored communion as a normal part of the regular Sunday service.

The service, sacraments, and rites

Lutherans observe two sacraments, baptism and the Lord's Supper (communion, Eucharist). The common practice is to baptize children and adults who have not been baptized previously. The frequency of communion has increased in recent years, but there are still many congregations where it is celebrated only once a month or less often. Though the usual practice has been that only those who have been confirmed may participate in the Eucharist, in 1970 the Lutheran Church in America and the American Lutheran Church approved participation for 10-year-old baptized children, whether they have been confirmed or not.

The rites of the Lutheran churches are confirmation, ordination, marriage, and burial. In the rite of confirmation (usually between the ages of 10 and 15) a member makes public profession of the faith received in baptism. In the rite of marriage the church ceremony may replace the civil ceremony or it may serve as an invocation of blessing on the civil ceremony. Ordination of the clergy does not endow its members with a special character or give them a special status, but it sets them apart for the particular office of preaching the word and administering the sacraments. This rite is interpreted either institutionally (*i.e.,* preaching is an order instituted by Christ and transmitted from generation to generation [succession]) or congregationally (*i.e.,* the congregations call certain of their members to assume the functions of preaching and administering the sacraments for them). In 1970 the Lutheran Church in America and the American Lutheran Church approved the ordination of women, a practice carried over in the Evangelical Lutheran Church in America, which came into being from a merger in 1988. There is no sacrament of extreme unction, but there is a burial service for the dead.

An important role was played in the Reformation by hymns, which not only conveyed the evangelical teaching but also allowed for popular participation in the church services. The best-known Lutheran hymns come from the period of the 16th and 17th centuries (*e.g.,* "A Mighty Fortress Is Our God" by Luther, "All Glory Be to God on High" by Nikolaus Decius, "O Sacred Head Now Wounded" by Paul Gerhardt, "Wake, Awake, for Night Is Flying" by Philipp Nicolai, "Now Thank We All Our

Hymnody

God" by Martin Rinkart). But each nation has made its contribution (*e.g.*, Thomas Kingo in Denmark and Norway), and Lutheran hymnals today include hymns from many ages, nations, and communions.

Among the composers of choral music (cantatas, motets, masses, settings of the Passion of Christ) Johann Sebastian Bach ranks highest (*e.g.*, *Mass in B Minor*, *St. Matthew Passion*, and *St. John Passion*). But other composers of importance were Michael Praetorius, Heinrich Schütz, and Dietrich Buxtehude. To this music should also be added the Scandinavian folk tunes (*e.g.*, L.M. Lindeman in Norway).

Education. Education of the laity and clergy was an early problem for the Reformers. The means developed to meet it have had a formative influence on Lutheranism to the present day.

To instruct the people in Christian teaching, Luther not only translated the Bible into the vernacular but also wrote his Small and Large Catechisms (1528–29). The small one was to be used by heads of households to instruct those under their care. It includes not only the three parts that had been in use before (the Ten Commandments, the Creed, and the Lord's Prayer) but also three additional parts on baptism, the Lord's Supper, and absolution. Each topic in the various parts is connected with an explanation in the form of an answer to the question, "What does this mean?"—a device Luther used in order to avoid mechanical memorization.

The catechisms

The Small Catechism, with various expositions, has remained a basic instructional tool in the Lutheran churches, though it has been supplemented by other materials (*e.g.*, Bible courses, Sunday school literature, projects).

In the 20th century efforts have been made to connect the secular world and the Christian tradition by establishing institutions such as the academies for laity in Europe (which provide opportunity for regular meetings of persons from specific vocations to discuss the relevance of Christianity to those vocations) and the church colleges in the United States.

Organization. The polity of the Lutheran churches varies from country to country. The Church of Sweden has maintained the episcopal succession unbroken, and congregations there are given great freedom to appoint their own pastors. The Danish Church lost but later regained the episcopacy. In Norway there is a closer tie between church and state than in the other Scandinavian countries. Since 1869, by an arrangement with Russia, the Finnish Church is independent of state control but is supported by public funds.

Until the end of World War I the churches in Germany were under secular authority, administered by a commission of laity and clergy, a system that grew out of the emergency situation of the Reformation. After the collapse of the government in 1918, the churches drew up new constitutions placing the congregations under a General Synod in some provinces and under a bishop in others; and the several provincial churches (*Landeskirchen*) were united in the German Evangelical Church Federation (1922). At the end of World War II, after the conflicts under Hitler, the Evangelical Church in Germany was organized under Bishop Theophil Wurm and Pastor Martin Niemöller, adopting the Declaration of Barmen (1934) as a binding statement. The United Evangelical Lutheran Church of Germany, formed in 1948, became a unit within the Evangelical Church in Germany.

In the United States the Lutheran churches have the same denominational standing as other churches. The polity is congregational, but in a complex form in which congregations yield some authority to synods on regional and national levels. Elected heads are called presidents in some Lutheran bodies, but in the largest, the Evangelical Lutheran Church of America, they are bishops.

Besides these larger Lutheran churches there are a number of Lutheran free churches in Europe (*e.g.*, Evangelical Lutheran [Old Lutheran] Church, Germany) and in the United States (*e.g.*, Church of the Lutheran Confession), which have complete congregational autonomy.

The Lutheran World Federation, established in 1947, is a cooperative organization. (R.P.S./M.E.M.)

Reformed and Presbyterian churches

Reformed and Presbyterian churches share a common origin in the Reformation in 16th-century Switzerland. Reformed is the term identifying churches regarded as Calvinistic in doctrine. The term presbyterian designates a collegial type of church government by pastors and by lay leaders called elders, or presbyters, from the New Testament term *presbyteroi*. Presbyters govern through a series of representative consistories, from the local congregation to area and national organizations, commonly termed sessions, presbyteries, synods, and assemblies.

Meaning of Presbyterian and Reformed

A slogan for the Lutheran Reformation was "by faith alone." Reformed Christians added the principle "to God alone the glory." Reformed Christians emphasized that God's word alone and no mere human opinion should be the norm for faith. "To God alone the glory" determined attitudes toward church government and worship, the design and furnishing of church buildings, and even secular authority. Reformed churches are confessional in nature, and during the 16th and early 17th centuries a number of manifestos of faith were written. Some of these confessions were theses for debate, such as Zwingli's *Sixty-Seven Articles* of 1523. Others, such as the Zurich Consensus of 1549, sought unity between groups on controversial doctrines. The very names of the Geneva, Helvetic, French, Belgic, and Scots confessions indicate the relationship of Reformed churches to the rising sense of nationhood in 16th-century Europe. A harmony of confessions prepared in 1581 shows the agreement among national churches as well as between Reformed confessions and the Lutheran Augsburg Confession. Some national confessions had international significance. The Second Helvetic Confession became standard for churches in countries east of Switzerland. The Heidelberg Catechism had great importance in the churches of the Netherlands and wherever the Dutch settled. The Westminster Confession of Faith, produced in 1648 by a committee appointed by the English Parliament, had its greatest influence among Presbyterian and Congregational churches outside of England.

HISTORY

This section treats developments within the Reformed and Presbyterian churches after the Reformation. For a discussion of the emergence of these churches, see above *History of the Protestant movement*.

After the Reformation in Europe. *Reformed churches in eastern Europe.* Reformed Christianity in eastern Europe had great strength among Hungarians. By 1576 the government of the Hungarian Reformed Church emerged with superintending bishops chosen by church councils of pastors and elders. In 1606 István (Stephan) Bocskay, prince of Transylvania, secured recognition of the rights of Hungarian Reformed churches in territories under both Habsburg and Turkish rule, and Reformed faith was identified with Hungarian nationalism. The Transylvanian town of Debrecen became known as the Calvinist Rome. Transylvania, a sovereign state at the Peace of Westphalia ending the Thirty Years' War in 1648, fell under Habsburg domination later in the century. This resulted in a Counter-Reformation against Protestants, which was lightened by toleration in 1781 and equality under the law in 1881. Partitioning of Hungary in 1919 and 1945 left a significant number of continuing Hungarian Reformed churches in Romania, Czechoslovakia, the Soviet Union, and Yugoslavia as well as in the present state of Hungary.

The Thirty Years' War was devastating to the Hussite Unity of Brethren in Bohemia, who had identified with the Reformed tradition during the Reformation. Protestantism survived underground until limited toleration came in 1781. Two Czech Brethren denominations exist in present-day Czechoslovakia. A Christian Peace Movement, which has gained international significance, developed from these churches in Prague during the 1950s.

Though Poland produced an influential Reformed theologian in Jan Łaski (d. 1560), the Counter-Reformation reduced Reformed churches to the status of a small sect in Poland by the 17th century. In 1648 there were still more than 200 Reformed congregations, but by the late 20th

century there were only eight congregations in Poland, five in Lithuania, and one in Latvia.

Congregational churches in Bulgaria and Evangelical churches in Greece are members of the World Alliance of Reformed Churches.

Reformed churches in France. French Calvinists, or Huguenots, set the pattern for presbyterian organization on a national level at a synod of the Reformed Church of France in 1559. During the religious wars of the next decades they developed a theory of resistance to the unjust state, but the end of effective resistance came with the fall of La Rochelle in 1628. Huguenots remained as a weakened, tolerated minority in France. On Oct. 18, 1685, Louis XIV revoked the Edict of Nantes, which had granted Huguenots limited toleration. At least 250,000 French Protestants emigrated to Prussia, Holland, England, and America. After the suppression of the Camisard (French Protestant peasant) revolt in 1715, Louis XIV announced the end of the practice of Protestantism in France. Yet that very year a group met in Nîmes to plan restoration of the Reformed Church. With the 1789 French Revolution equality under the law came to Protestants. Napoleon placed Reformed congregations under state control, with pastors on state salary.

Revocation of the Edict of Nantes

A national synod did not meet again until 1848. At that time a free Evangelical Synod was organized, separating from the state-recognized church over the issue of state support. In 1905 state support of the old synod was withdrawn, and the two synods were united in 1938.

When Alsace was annexed to France in 1648, a number of Reformed Christians were brought into the French nation. But the Reformed Church in Alsace-Lorraine, whose history has been different from that of the Reformed Church of France, remained a separate organization. Outside of French-speaking Switzerland, French Reformed churches are the largest Protestant group in the Latin countries of Europe, each having a Reformed Church. French Reformed Christians have played a role in the World Council of Churches, in liturgical and theological renewal, in relating the church to technology and urbanization, and in Catholic–Protestant and Communist–Christian dialogue.

Reformed churches in Germany. The Peace of Westphalia in 1648 established the legality of Reformed churches in German states, according to the pleasure of the ruling prince. At the end of the 17th century Reformed Christians in the Palatinate faced an attempt at their destruction. Many fled to the Netherlands, America, and Prussia, where Reformed churches were established. The Hohenzollern Elector of Brandenburg was converted to Calvinism in 1609. Hohenzollern rulers permitted the establishment of Reformed churches among refugees and also continued Reformed churches in territories that came later under Prussian rule.

Frederick William III of Prussia in 1817 proposed a union of Reformed and Lutheran churches. Reformed theologian Friedrich Schleiermacher led ministers in support of this union but shared with them a concern for the loss of Reformed systems of self-government to monarchial absolutism. The union became a pattern for a majority of Protestants in Germany. Distinctively Reformed territorial churches are still to be found in northwestern Germany. The Reformed Church of Anhalt joined in the union Evangelical Church in 1981.

A Reformed Alliance was organized in Germany in 1884 to preserve the Reformed heritage. A synod held in Altona in January 1934 drew up a confession in opposition to Nazi corruption of the Gospel. This led to the Barmen Synod of May 1934, in which Christians of Lutheran, Union, and Reformed background joined in the Barmen Confession of Faith. This confession was the basis for resistance to Hitler by the Confessing Church. After World War II the Confessing Church ceased, but its work continued to be an inspiration to churches in both West and East Germany. The Reformed Alliance remains active in unified Germany.

The Barmen Confession

Reformed churches in England and Wales. The failure of the Puritans either to complete establishment of a presbyterian system during the Westminster Assembly in 1648 or to continue a looser arrangement of independent churches under Cromwell opened the way in 1660 to an episcopal restoration in the Church of England. Those Reformed Christians who could not accept this became persecuted Nonconformists. The Glorious Revolution of 1688–89, which expelled the Roman Catholic sovereign James II, gave English Presbyterians, Independents, and Baptists limited toleration outside the state church. Many Presbyterian congregations became Unitarian during the next century. This movement was checked by the Evangelical Awakening of the 18th century, which reinvigorated the Nonconformist groups.

The free-church spirit changed the Reformed ethical emphasis on parish discipline to formation of voluntary societies dedicated to preserve the sabbath, to suppress vice, to abolish slavery, and to work for moral reform.

In 1972 the United Reformed Church was formed out of the Congregational Union of England and Wales and the Presbyterian Church of England. The Presbyterian (Calvinistic/Methodist) Church of Wales, formed in the 18th century, has a substantial membership.

Reformed churches in Scotland and Ireland. The refusal of the Episcopal bishops of the Church of Scotland to accept the legitimacy of William and Mary in 1688 resulted in presbyterian government for the church. State interference in the appointment of pastors along with evangelicalism gave rise to secessionist movements in the 18th century, culminating in 1843 in a major schism and the formation of the Free Church of Scotland under Thomas Chalmers. In 1900 secession and free churches became the United Free Church, which in turn reunited with the Church of Scotland in 1929.

Secessionist movements

In Ireland the Presbyterian Church has roots both among Scottish settlers and also among English Puritans of the early 17th century. Although the church is represented in all of Ireland, most of its membership resides in Northern Ireland, where Irish nationalism is a crucial issue.

Reformed churches in The Netherlands and Switzerland. The Peace of Westphalia in 1648 ended the Eighty Years' War for the independence of the Netherlands. The Reformed Church, which was identified with Dutch nationalism, constituted the majority church within a nation that had remarkable tolerance for religious minorities.

Closer state control of the church followed the Napoleonic era. This and an enervated theology prompted two secessions from the Dutch Reformed Church, the first in the 1830s and the second in the 1880s. These secession churches united as the Gereformeerde Kerken in The Netherlands, which exist alongside the traditional Hervormde Kerk. Abraham Kuyper, the scholarly neo-Calvinist leader of the second of these secessions, served as prime minister of The Netherlands with a conservative coalition in Parliament from 1901 to 1905. The two main bodies of Reformed Protestantism in The Netherlands cooperate on many levels.

Nineteenth-century evangelical secession and 20th-century reunion occurred in Swiss Reformed churches, which continue to be organized along cantonal lines. A Christian Socialist movement was developed in the early 20th century. Karl Barth and Emil Brunner, whose theological influence went far beyond Switzerland and the Reformed tradition, emerged from that movement with less utopian political realism.

Reformed and Presbyterian churches in the United States. *The colonial period.* Persons of Reformed background were important in shaping and directing the political and religious course of the 13 American colonies. In 1611 Alexander Whitaker, son of a Reformed theologian, began to establish churches in Virginia. Elder William Brewster, in the 1620 Plymouth Colony, used the writings of the English Presbyterian Thomas Cartwright as his guide in church government. A Dutch Reformed Church was organized on Manhattan Island in 1628, and the Massachusetts Bay Colony in 1630 was a new model Reformed church and commonwealth. In the 17th century Waldensian refugees came to Staten Island, and Huguenots settled in New York and New England. These were followed by Scots-Irish immigrants, who settled throughout the colonies, and by German Reformed refugees from the Palatinate.

The 18th-century Great Awakening—led by Calvinist preachers Jonathan Edwards, Theodore Frelinghuysen, George Whitefield, and Gilbert Tennent—encouraged an evangelical Christianity often at odds with establishment attitudes. Hence revival-seasoned clergy learned to fight for the free expression of religion. These evangelicals joined with deists in supporting religious liberty in the constitutional foundation of the United States.

Calvinist culture

The 19th century. Most religious groups in the new nation had a Calvinist viewpoint and pattern of life, favouring constructive activity rather than idle enjoyment. Art, music, literature, and recreation were approved only if edifying. Sunday was a quiet day with minimal farm chores, freedom from business cares, Sunday school, church, and conversation among friends. A disciplined nation might receive the blessing of God and enjoy peace and prosperity. Revivalism was seen as the means by which people could be brought under the Lord's discipline. Revivals then bore fruit not only in disciplined souls but also in movements for women's rights, abolition of slavery, and temperance. Saving souls and building a better world came to be two aspects of the Kingdom of Christ in America.

The 20th century. After the Civil War (1861–65) conflict developed between those who adapted Darwinism to theology and those who saw evolution as a threat to biblical authority, between those who championed higher biblical criticism and those who opposed it. This conflict peaked in a fundamentalist-modernist controversy in the 1920s with fundamentalists withdrawing to the edges of American denominational life. In the 1980s television preachers gave the fundamentalist perspective not only new popularity but also political significance.

Mainline denominations, however, have been in numerical decline. Reformed Christianity is still concerned about achieving a more just society and at the same time is working for the redemption of individuals. There is debate over goals and methods.

Reformed and Presbyterian world mission. *Asia.* In 1622 an institute was founded in Leiden (the Netherlands) to prepare missionaries for the Dutch Indonesian colonies. Building upon work begun by Catholics, Presbyterian missionaries established churches in Indonesia that by the late 20th century comprised at least one-third of all Asian Reformed and Presbyterian Christians.

Presbyterian churches in Korea

Presbyterian churches in Korea have been established for more than 100 years and are second in Asian membership to the Reformed churches of Indonesia. Not only have these churches grown rapidly in South Korea, but through immigration they constitute the fastest growing segment of Presbyterian churches in the United States. Identified with Korean nationalism in the past, these churches have found themselves in tension with the government of South Korea. In 1986 contact was made with Presbyterian Christians in North Korea after 40 years of isolation.

The strong Presbyterian Church in Taiwan has been identified more with the native Taiwanese than with church members coming from mainland China after 1945. Conflict with the government has resulted in the jailing of Taiwanese Presbyterian leaders.

Presbyterian and Reformed churches exist in Japan, Thailand, Malaysia, Singapore, Burma, India, Pakistan, Iran, Syria, and Lebanon. There is also a strong Presbyterian and Reformed component in larger United churches in Japan, the Philippines, India, and China. With new tolerance in the 1980s in the People's Republic of China, a resurgence of the United Protestant Church of Christ in China has taken place. Church buildings have been reopened and new congregations formed.

Africa. Reformed churches in Africa date from Dutch settlement in South Africa in 1652 as well as from settlements by Huguenot and German Reformed refugees somewhat later. With British occupation in South Africa in 1806 Scots brought Presbyterianism. By the late 20th century half of the Presbyterian and Reformed membership in Africa was in the Republic of South Africa. White Dutch Reformed churches have been closely identified with the government policy of apartheid. At the meeting of the World Alliance of Reformed Churches in Ottawa, Can., in 1982 apartheid was declared heresy. Two of the white Reformed denominations then were suspended from the alliance, and the Reverend Allan Boesak, a Colored Reformed pastor and leader of the anti-apartheid forces, was named president of the World Alliance. A confessional statement, the Kairos Document, drawn up in 1985 by Reformed, Congregationalist, Presbyterian, and other church leaders, affirmed a theology unconditionally opposed to the state theology of South Africa. It has been compared to the 1934 Barmen Confession in Germany calling for resistance to the state.

Other African nations with large Presbyterian church membership include Madagascar, Kenya, Zaire, Cameroon, Malawi, Egypt, and Ghana. Churches from 16 other African nations belong to the World Alliance.

Other areas. In Canada, Australia, and New Zealand, as well as in the Pacific Islands and West Indies where there were former British colonies, there are both Presbyterian churches and United or Uniting churches with Presbyterian components.

In 10 countries of Latin America there are member churches of the World Alliance, but half of the Presbyterian and Reformed membership is found in Brazil. Since most of the Presbyterian membership in these countries is of middle-class background, liberation theologies that identify with the concerns and needs of the poor have created controversy. There is a small but vigorous Presbyterian-Reformed Church in Cuba.

The success of the world mission can be seen in the vanguard of Reformed theology. For most of the 20th century influential Reformed theologians included such white, male, North Atlantic leaders as Barth, Brunner, John and D.M. Baillie, Reinhold and H. Richard Niebuhr, Hendrik Kraemer, and Jürgen Moltmann. This type of leadership has begun to make room for theologians from Asia, Latin America, and Africa, such as C.S. Song, Kosuke Koyama, Mariane Katoppo, Yong-Bok Kim, Elsa Tamez, and Allan Boesak. Reformed theology has become global.

Global theology

Reformed Christians in the ecumenical movement. Since the time of Martin Bucer and John Calvin the Reformed movement has had leaders who were untiring in efforts toward church unity. In the 17th century the Scot John Dury and the Czech John Amos Comenius were notable for their ecumenical efforts. While later Pietism and Evangelicalism divided churches, people were also encouraged to put aside differences for common goals. Mission societies received support and sent missionaries from diverse denominational backgrounds. In the past 150 years Presbyterian and Reformed churches have not only reunited among themselves but also have formed close links with churches of other historical backgrounds. In the United States discussion and the adoption of consensus papers have taken place since 1961 by a Consultation on Church Union that included Reformed, Presbyterian, Congregational, Methodist, Episcopal, and Disciples churches.

The World Council of Churches was organized in 1948. Reformed and Presbyterian churches participate in local and regional councils of churches and interfaith groups. Since the second Vatican Council (1962–65), called by Pope John XXIII, there has been increased dialogue with Roman Catholics. The insights coming through ecumenical and interfaith relationships make for more global, more dynamic, and more relevant teaching and practice in Reformed and Presbyterian churches.

TEACHINGS

Doctrines. Reformed churches consider themselves to be the Roman Catholic Church reformed. Calvin in his *Institutes* spoke of the holy Catholic Church as mother of all the godly. Bullinger in the Second Helvetic Confession made it clear that Reformed churches condemn what is contrary to ecumenical creeds. Interpretations of the early Church Fathers and decrees and canons of councils "were not to be despised, but we modestly dissent from them when they are found to set down things differing from, or altogether contrary to, the Scriptures." Universal articles of Christian faith, such as the doctrines of the Trinity, the humanity and divinity of Christ, and the sin of man and the saving work of Christ, are affirmed in Reformed faith.

Reformed churches share with Lutheran and other

Justification by grace through faith

Protestant communions the concept of justification by grace through faith as central to the Gospel. The essence of faith is God's forgiving love coming as a gift through Jesus Christ. As with Lutherans, the true treasure of the church is this good news of the grace of God. Scripture is the authoritative witness of the good news, but, as was stated in the Westminster Confession, "authority thereof is from the inward work of the Holy Spirit, bearing witness by and with the word in our hearts." Calvin said: "There is no doubt that faith is a light of the Holy Spirit through which our understandings are enlightened and our hearts are confirmed in a sure persuasion." Such understanding is shared by Lutheran and Reformed Christians.

The church and the sacraments. Calvin tried unsuccessfully to mediate between the Lutherans and Zwinglians, holding that Zwingli had been more concerned to show how Christ was not present than how he was and affirming, with Luther, the real presence of the resurrected Christ in communion. In the 1980s Lutheran and Reformed churches in Europe and the United States came to recognize each other's ministries of word and sacrament.

Both Calvin and Bucer, more than Luther, were concerned to keep the "profane" from receiving communion. This encouraged the development of church discipline, and the use of elders to oversee discipline within the parish became a feature of Reformed church life. In the struggle to maintain that discipline, Calvin's successor, Theodore Beza, asserted that the presbyterian form of government was ordained by Christ.

Scripture and tradition. Before the Reformation, humanists rejected arguments that appealed to the authority of church tradition. They made the authority of Scripture central in the church. Following them, Reformed Christians insisted that no authority in the church was on a level with Scripture; by Scripture all tradition was to be judged.

Church and state. The position in the Swiss Reformation was that church and state should render reciprocal service yet remain distinct. The church invisible consisted of God's elect, but the membership of a visible church approximated the population of the corresponding state. Beyond borders national churches kept communion with each other in spite of differences of custom.

Obedience was required of Christians, even to unworthy rulers, unless the ruler commanded disobedience to God. On such occasions, God rather than man must be obeyed. But even then, the private individual should not actively resist the ruler. It was the responsibility of lesser magistrates to bring such rulers into line. Sixteenth-century resistance of Huguenots in France, Protestants in Scotland, and Puritans in England was justified on this basis.

English Puritans asserted that the government of the state should be patterned after their form of government in the church. This teaching was one source of modern constitutional government. Another source in Reformed tradition was the belief that no one person should be trusted with unlimited power, a doctrine James Madison built into the U.S. Constitution.

There has been a constant Reformed hope that the kingdoms of this world may be brought closer to the will of God and that this would result in a better justice for all. This view requires that church people become involved in politics.

The sovereignty of God and double predestination. There has been no argument in Reformed theology about the positive side of the doctrine of predestination concerning the election of those whom God wills to save. Difference of opinion, however, arose over whether God determines who is reprobated. Bullinger did not believe that it was God's will that "one of these little ones should perish." He maintained that Christians should always hope for the best for all. Calvin affirmed "double" predestination, meaning that both reprobation and election are within the active will of God. His reason found this appalling but scriptural. To call God, thereby, unjust was to judge One who is the very standard of justice.

In his *Institutes* Calvin discussed predestination in the context of the love and grace of Jesus Christ. Later theologians expounded predestination more abstractly as an aspect of God's sovereignty. Arminianism rose in protest to this. The defenders of double predestination thought that Arminianism would cut the nerve of the Protestant doctrine of justification by grace alone and lead people back to popery. Hence, at Dort in 1618, double predestination was affirmed as Reformed orthodoxy.

WORSHIP AND ORGANIZATION

Liturgy. In the Reformation earlier liturgies were modified by using the vernacular, removing anything that implied the reenacting of sacrifice in the mass, providing for congregational confession, and emphasizing the preaching of the word. Following Erasmus' recommendation, the singing of Psalms became characteristic of Reformed worship. While most Reformed churches today use a broad spectrum of vocal music, some hold exclusively to Psalms.

Music and sermons

Stress on preaching reached its peak among English Puritans. Some clergy preached two hours on an Old Testament text on Sunday morning, two hours on a New Testament text in the afternoon, and devoted the evening to discussion of the day's sermons with the congregation. Calvin held that the Eucharist should be celebrated weekly, though others believed that it was too sacred for such frequent use. Care was taken to instruct participants and to prepare them for confession. The Eucharist was served around a table.

In the 20th century attention has been given to relating worship to the social and material needs of human beings as well as to communicating the word to human hearts and minds. At the Iona Community in Scotland, for example, where worship is directed to those intending to work in economically deprived areas, and at the Taizé Community in France new forms of worship are being developed. In recent years there has been emphasis upon celebration in response to the good news of God, a greater appreciation of the arts in worship than in the past, and a concern for inclusive language.

Religious education. The requirements of Reformed life have demanded an educated clergy and an informed laity. Besides academic training for pastors, the early practice was for them to meet often and for one to interpret Scripture and for the others to engage in critical discussion. Queen Elizabeth I suppressed the custom in England, for she believed that four sermons a year were quite enough and that gatherings of pastors might be subversive.

Lay education was accomplished through preaching the word and teaching the catechism, such as Calvin's Little Catechism, which was designed for teaching the young. Others, such as the Westminster Larger Catechism, were used to instruct pastors and teachers. More recently catechetical instruction has given way to inductive forms of education, with emphasis on the age level at which instruction takes place. There is also concern to relate the Christian faith to the daily life of the larger community.

Present organization of Reformed and Presbyterian churches. In Presbyterian churches a local congregation is ruled internally by a session moderated by the pastor and composed of laity (elders) elected from the congregation. A presbytery formed of pastors and elders representing each congregation rules over local congregations on a district level. In other Reformed churches the district association has less power and the local congregation more than in Presbyterian churches. In Hungarian Reformed churches a presiding bishop moderates the presbytery.

Beyond the district level are regional synods or conferences and national assemblies. These bodies are usually composed of an equal number of clergy and laity. Since 1875 there has been a World Alliance of Reformed Churches, which was joined in 1970 at Nairobi, Kenya, by the International Congregational Council to form the World Alliance of Reformed Churches (Presbyterian and Congregational). There are about 160 member denominations.

Although a few Reformed groups still have a special relationship to the government of their nation, there is little difference in practice between established and free Reformed churches.

Social ethics. Reformation leaders were involved in the total life of their communities. Calvin's relation to the education, health and welfare services, refugee settlement,

industry, finance, and politics of Geneva is well documented. The historian R.H. Tawney, impressed by this, has called Calvin a "Christian socialist." The English Puritans believed that if they could reshape the political and church life of the nation, God's blessing would come upon the land instead of war, famine, and pestilence. Concern to achieve greater social justice for humankind has been normative among Presbyterian and Reformed churches. Such concern in the past has been seen as resulting sometimes in petty rules and harsh administration, but in new forms that concern is still a living force.

Types of Reformed piety. In Zwingli, Calvin, William the Silent, and Cromwell, a classic type of Reformed piety was manifest. Those persons saw themselves as God's instruments in redeeming human affairs, even at cost to themselves, and they had high expectations of others. Living under God's mercy, they showed little fear of the powers of this world and were ready to make choices on a pragmatic basis.

In a less heroic mold were Reformed Christians who did not expect to change history but who encouraged the development of godliness in those about them, beginning with themselves. The increasing emphasis in the late 16th century upon the personal experience of saving faith helped the Reformed tradition to become a nursery for Pietism in the late 17th and 18th centuries. Along with a more confessional orthodoxy and a more rationalistic liberalism, such Pietism remains to the present. A new style of worldly Christianity is emerging with Christ, standing for and with the oppressed, as the model. (J.C.S.)

Anglican Communion

Anglicanism refers to the form of Christianity practiced by the churches of the Anglican Communion. This loosely organized family of religious bodies represents offspring of the Church of England, one of the major branches of the 16th-century Protestant Reformation. It is a form of Christianity that includes features of both Protestantism and Catholicism. It prizes traditional worship and structure but operates autonomously and flexibly in different locales. Anglicans possess few firm rules but a cluster of historic pieties and procedural loyalties. *The Book of Common Prayer,* a compilation of the church's liturgical forms originally issued in the 16th century, represented the achievement of autonomy from Rome and remains the hallmark of Anglican identity. The prayer book derives from ancient English spirituality and embodies the uniqueness of Anglican Christianity.

HISTORY

Christianity in England. The Church of England, mother church of the Anglican Communion, has had a long history. When Christianity began in England is uncertain, but it probably was not later than the early 3rd century. The church was well enough established by the 4th century to send three British bishops—of Londinium (London), Eboracum (York), and Colonia Linum (Lincoln)—to the Council of Arles (in modern France) in 314.

Missionary efforts of the English Church

In the 5th century, after the Romans had withdrawn from and the Anglo-Saxons had invaded Britain, Illtud performed missionary work in Wales and Patrick in Ireland. Though isolated from continental Christianity in the 5th and 6th centuries, Christianity in the British Isles grew due to the influence of monasticism. About 563 Columba founded an influential monastic community on the island of Iona off Scotland. In 597 a monk named Augustine went to England at the request of Pope Gregory the Great to oversee the development of English Christianity. Augustine's archbishopric at Canterbury soon became the symbolic seat of England's church. Subsequent mission work, such as that of Aidan around 634 in northern England, solidified the church's life. The early Catholic Church in England was a distinctive fusion of Romano-British, Celtic, and Roman influences. It retained powerful centres in the monasteries and lived in tension with the medieval monarchy. The martyrdom of Thomas Becket demonstrated the church's concern to preserve its integrity over the throne in the 12th century. The writings of John Wycliffe (d. 1384) questioned the form of the medieval church and became an early protest against Rome's control over England's church.

Under King Henry VIII in the 16th century the Church of England broke with the pope. Henry wished no Reformation but intended to substitute his royal authority over the English Church for that of Rome. Upon Henry's death Archbishop Thomas Cranmer began changes that allied the Church of England with the Reformation. His *Book of Common Prayer,* which appeared first in 1549, revised traditional forms of worship to incorporate Protestant ideas. When Elizabeth I assumed the throne in 1558 the Reformation in England triumphed. The theologian John Jewel (1522–71) wrote that England's church had returned to ancient precedent. Richard Hooker (1554?–1600) offered a defense of English Church order against Puritans and Catholics in England. In the 17th century Puritan opposition achieved powerful political form. But the Restoration of 1660 ended the Puritan commonwealth and began more than a century of great influence for the Church of England. Until the early 19th century it dominated England's religious life and became closely allied with the power of the throne.

The Church of England became a considerable social and spiritual force, its piety permeating English life. The church generated impressive forms of philanthropy, and clergy commonly performed the duties of civil servants. Anglican influence spread to colonial areas in India and North America. But the church's hold on English religious life began to wane in the 18th century despite impressive reform efforts. John Wesley, Charles Simeon, John Newton, and other Evangelical clergy prompted a surge of new religious fervour. Evangelical laity such as William Wilberforce and the Clapham Sect fought slavery and encouraged social reform. In the early 19th century the Anglo-Catholic (High Church) Oxford Movement led by John Henry Newman, John Keble, and E.B. Pusey attempted a recovery of ancient liturgy and a response to social concerns. The church made impressive efforts to encompass the diversity of modern English life while retaining its traditional identity.

Developments in worldwide Anglicanism. From the time of the Reformation the Church of England expanded, following the routes of British exploration and colonization. It served native peoples and expatriates alike, and all initially considered themselves loyal to the see of Canterbury. The Church of England's great missionary societies were important agents of its growth beyond England. The Society for Promoting Christian Knowledge, founded in 1699, the Society for the Propagation of the Gospel in Foreign Parts, 1701, and the Church Missionary Society, begun in 1799, achieved global identity. These societies undertook mission work among indigenous people of English colonies and began the process of transferring authority in church matters to local leadership. Anglicanism thus came to function as a decentralized body of national churches loyal to one another and to the forms of faith inherited from the Church of England.

Development of local autonomy

Social and political circumstances often hastened the development of autonomy. The American Revolution compelled the organization of the Episcopal Church, which completed its structure by 1789. The first American bishop, Samuel Seabury, was consecrated in Scotland in 1784. The Anglican Church of Canada had its own separate organization in 1893, though it was known as the Church of England in Canada until 1959, just as the Anglican Church of Australia continues to be so designated.

Initially Anglicanism's growth followed the outline of the British Empire. Vigorous missionary work produced strong church life in such diverse places as Nigeria, Kenya, South Africa, India, and Australia. In China and Japan, British, American, and Canadian Anglicans combined their efforts. The church left an impressive legacy of educational institutions and medical facilities. Here and there native peoples became clergy and even bishops. Samuel Crowther of Nigeria became the first black bishop in 1864.

Consolidation and indigenization characterized later Anglican mission. By the late 19th century Anglican bishops began meeting once a decade for the Lambeth Conference

at the archbishop of Canterbury's residence in London. The immediate cause of the initial meeting in 1867 was a controversy that arose in one of the colonial churches. The archbishop of Cape Town, Robert Gray (who was High Church, or traditionalist), wanted the bishop of Natal, John Colenso (who was Low Church, or evangelical), to be arraigned on charges of heresy for holding what were then regarded as advanced views of the creation stories in the opening chapters of Genesis. The controversy centring on Bishop Colenso aroused intense feelings and anxieties over a wide range of issues—doctrinal, personal, and organizational—among all the Anglican churches throughout the world. Bishop Colenso was convicted and deposed in the church courts but upon appeal to the civil courts of England won his case and retained his church properties. What began as a jurisdictional dispute in South Africa became a matter of concern for all Anglicans. The issue of the relationship between the church's various branches required clarification. Lacking an authoritative centre, however, Anglicans have continued to rely upon consultation and consensus to coordinate matters of belief and practice.

Bishop Colenso

The end of colonialism and the rise of newly independent nations compelled Anglicans to rethink their identity and mission. Once the church of the colonizer, Anglicanism has spawned a host of self-directing churches linked by common form and historic allegiance to the Church of England. In most cases Anglicanism has been able to adapt in an affirmative way to new and changing social circumstances. In 1947 Anglicans joined several Christian bodies to create the Church of South India, a unique ecumenical union. Frequently Anglicans have been articulate opponents of injustice. Archbishop Janani Luwum of Uganda was martyred for opposition to the rule of Idi Amin. In South Africa the Anglican Church has consistently opposed apartheid, and Archbishop Desmond Tutu won the Nobel Prize for Peace for 1984 for his stand on behalf of racial equality. Anglicans rarely become revolutionaries, for the church views its task as working through existing structures for justice.

The Church of England has evolved a similar posture since the mid-19th century. Still the nation's official church, it has experienced attrition and attempted to redefine its place in English life. A succession of powerful leaders have enhanced the church's claim of being the nation's soul. In the latter 19th century Christian Socialism was an effort to draw compassionate attention to social problems. Sparked by the theologian F.D. Maurice, the movement later was led by clergy such as Stewart Headlam and Henry Scott Holland. In the 20th century Archbishop William Temple underscored that the church was a community of worship in step with modern life. The scholar and lay theologian C.S. Lewis restated the tenets of Christian belief in a sensitive response to modern doubt, and John A.T. Robinson affirmed the searching quality of modern Christian experience.

TEACHINGS

Doctrinal views. What has come to be known as the Lambeth Quadrilateral defines Anglicanism's essential beliefs. First suggested by an American, William Reed Huntington, in 1870, the Quadrilateral stated four marks essential to the Anglican conception of Christian identity—Scripture, the Nicene Creed, baptism and Holy Communion, and the episcopate. The Lambeth Conference of 1930 further clarified the nature of Anglicanism when it described the Anglican Communion as:

The Lambeth Quadrilateral

> a fellowship within the One Holy Catholic and Apostolic Church, of those duly constituted Dioceses, Provinces or Regional Churches in communion with the See of Canterbury, which uphold and propagate the . . . faith and order as they are generally set forth in the Book of Common Prayer . . .; promote within each of their territories a national expression of Christian faith, life and worship; and are bound together not by a central legislative and executive authority, but by mutual loyalty sustained through the common counsel of the Bishops in conference.

The Anglican Communion thus holds to the Catholic faith as expounded by the Holy Scriptures and by the early Church Fathers. It respects the authority of the state but does not submit to it; and it equally respects the freedom of the individual. In its relationship to the world the Anglican Communion does not seek to evade the challenges of the world or to live a life separate from it. Basing its doctrines on the Bible, the Anglican Communion allows a remarkable latitude of interpretation by both clergy and laymen.

Though the Church of England holds close to the spirit of the Thirty-nine Articles (a 16th-century doctrinal document that allows for broad interpretations), subscription to them is not required of the laity, and adherence by the clergy is expected only in a general way. Other churches or councils of the Anglican Communion take different views of the Articles, but none regards them as having, for example, the status of the historic statements of belief as set forth in the Apostles' or Nicene creed, nor do they accord them the status given to other 16th-century doctrinal statements, such as the Augsburg Confession of the Lutheran churches or the Westminster Confession of the Reformed and Presbyterian churches.

The ministry. Anglicans accept a threefold order of ministry, which consists of bishops, priests, and deacons. Though holding to the view of succession from the Apostles, Anglicans are not committed to any one theory regarding the conveyance of that ministry. Anglicans attempt to balance the clerical point of view with forms of authority that include the laity. Even bishops rarely are able to function without the advice and consent of other clergy and laity.

Bishops, priests, and deacons

WORSHIP AND ORGANIZATION

Anglican worship. Worship is the centre of Anglican life. Anglicans view their tradition as a broad form of public prayer, and they attempt to encompass diverse Christian styles in a traditional context. Although the prayer book is the most apparent mark of Anglican identity, it has undergone many revisions and wears national guises. The prayer book of 1662 represents the official version in the Church of England, but a 1928 version and a later Alternative Service Book are commonly used.

A few overseas Anglicans still rely upon the English prayer book of 1662, but most have their own versions, increasingly in languages other than English. All forms hold to the essential, historic elements of the prayer book but incorporate local idioms. In recent years there has been a recovery of ancient liturgical styles and vestments and a heightened emphasis upon the Eucharist as the central act of Christian worship. Experimental rites have appeared in different parts of the Anglican world. Change in Anglican worship has meant increased variety, new roles for laity, and a tendency toward freedom of expression while holding to the essence of the church's traditional forms.

Comprehensiveness in doctrine and practice. Often said to be the middle way between Roman Catholic and Protestant churches, the Anglican Communion is comprehensive in matters of doctrine and practice. While asserting the importance of the apostolic succession of bishops and *The Book of Common Prayer*, it nevertheless allows a considerable degree of flexibility in most doctrinal and liturgical matters. Thus, within the communion there are several schools of thought and practice, including High Church, Anglo-Catholic, Low Church, evangelical, and others. The various churches of the Anglican Communion, though autonomous, are bound together by a common heritage and common doctrinal and liturgical concerns, and there has always been a considerable amount of interchange of ecclesiastical personnel.

Authority and structure. Having no central authority and no one person, such as the pope of the Roman Catholic Church, from whom it can expect final authority, the Anglican Communion consists of national, autonomous churches that are bound together by intangible links best described as ties of loyalty between the see of Canterbury and each other. Although the archbishop of Canterbury is respected throughout the Communion and his words carry great moral authority, he exercises no jurisdiction over any part of the Communion other than over the diocese of Canterbury and over the Church of England as a whole through the authority vested in synods

Role of the archbishop of Canterbury

and convocations. Like a family the Anglican Communion changes its form and shape, increasing when new provinces (areas of jurisdiction) are formed and decreasing when schemes of union with non-Anglican churches are consummated. The basic unit of the Anglican Communion is the diocese, a geographic area over which a bishop presides. Dioceses generally form part of a larger unit known as a province, but even these are far from uniform in configuration. A province, for example, may be part of an autonomous church: the Anglican church in Australia has five provinces and the one in Canada four; the churches of England and Ireland have two each; and the Protestant Episcopal Church in the United States of America has nine. Some provinces, however, include whole countries, such as Japan, South Africa, West Africa, Kenya, Uganda, and Tanzania; other provinces cover a number of countries, such as the provinces of Central Africa, the West Indies, and the Southern Cone of America. On occasion, one diocese covers a whole country or even several countries, such as the diocese of Polynesia.

Titles of Anglican ecclesiastical leaders

Variations occur in the titles of the heads of the various provinces or national churches. England has two archbishops (Canterbury and York), known as metropolitans, as does Ireland. Canada has a primate (who has no province) and four metropolitans; Australia has five archbishops, one of whom—while having jurisdiction over a province—is known as the primate. The Church of Japan and the Episcopal Church of Brazil each has a primate who also has a diocese, and the United States has a presiding bishop and a primate, both without a diocese. To complicate organizational matters, the Scottish Episcopal Church has a primus (primate), and the Protestant Episcopal Church in the United States has presidents (who are elected only for three-year periods) of its nine provinces. Several branches of the church exist apart from provincial or national churches, though usually in reliance upon either the Church of England or the Episcopal Church in the United States. A number of dioceses in Central America, such as Guatemala, Honduras, Nicaragua, and El Salvador, participate in the Episcopal Church, though with the goal of eventual autonomy in a Central American province. The Church in Hong Kong is a special adjunct to the Council of the Church of East Asia, and the Church of Bermuda is an extra-provincial see of the Church of England. In some areas, such as China (Three Self Movement), India (Church of North India and Church of South India), and Pakistan, Anglicans have participated in the creation of ecumenical forms of church union.

Internal developments. The mother church of the Anglican Communion, the Church of England, has maintained close connections with the state; it has representative bishops in the House of Lords and can properly be called the established church, even though, contrary to much popular opinion, it is in no sense supported financially by the state. The Church of England itself is without question the church of the English people, even though many of the country's citizens do not so regard it. Only in England do Anglicans comprise a majority, accounting for more than one-half of the world Anglican population.

Apart from its assured position in the life of England, the Anglican Communion has never had much of a worldwide structure. Indeed, the Anglican Communion has been characterized by its lack of structured cohesion. Even meetings of Anglican Church leaders have been restricted, except in very recent times, to the meetings of the Lambeth conferences, which are held only once every 10 years, and to Pan-Anglican congresses, which involve clergy and laity as well as bishops. Only three such meetings have been held in the 20th century: in London in 1908, Minneapolis, Minn., in 1954, and Toronto in 1963. At two- or three-year intervals between Lambeth conferences meetings of the Anglican Consultative Council are held. While it has no real authority, the council gives cohesion to the Anglican Communion between Lambeth conferences. The council replaced the Lambeth Consultative Body, whose members were the primates or presiding bishops of the various national churches and also replaced the Advisory Council on Missionary Strategy, which came into being after World War II. The Lambeth Conference of 1968 recommended formation of the Anglican Consultative Council, and that body has assumed primary responsibility for coordinating the global Anglican network. The council is an advisory body of about 60 members, including bishops, clergy, and lay people. It shares information, coordinates policy, and develops unified mission strategies. Though lacking binding authority, the council has the archbishop of Canterbury as its president, and it increases the Anglican tendency toward consultation in matters of faith and life. The Lambeth Conference of 1978 recommended that the primates (heads) of all Anglican provinces meet regularly, and they have since done so in various countries of the Anglican Communion.

Anglican Consultative Council

The importance of conversation among Anglicans has been underscored by the extent of change in some branches of the Anglican Communion. In the second half of the 20th century most churches of the Anglican world revised their versions of *The Book of Common Prayer*. Decentralized and autonomous, Anglican branches have this freedom, although they are constrained by a sense of coordinating their efforts. In the United States revision of the Episcopal Church's prayer book was extensive. The new prayer book of 1979 incorporated years of liturgical study, of trial drafts, and of discussion. It offered unprecedented liturgical options, including use of modern English liturgies and opportunities for informal worship. The book generated controversy, which abated only slowly.

Equally controversial was the admission of women to the church's priesthood and the prospect of women bishops. Women had been ordained priests in Hong Kong in 1944 and in 1971. By the mid-1970s large numbers of women in various parts of the Anglican world called for the priesthood to be opened to them. The impact was greatest in the United States and Canada, where women became a significant percentage of seminary students. American Episcopalians approved women as priests in 1976 after heated debate. While several other Anglican churches took a similar course, the Church of England hesitated to study and to debate the issue. Opponents of the ordination of women feared the loss of the church's Catholic heritage. Advocates saw a chance for Anglican leadership in expanding the ministries open to women in the church. The Lambeth Conference of 1988 confronted the possibility that a woman would be chosen bishop in the United States, forcing the issue of women's ministries into the international context.

Global mission has remained a priority for Anglicans, but the gradual penetration of Latin America has been a recent feature. While recognizing and respecting the pervasive influence of Roman Catholicism in the area, Anglicans have found a niche among unchurched people. Social mission, education, and provision of indigenous leaders have characterized this phase of Anglican expansion. There has also been impressive growth in Africa and Asia, all sparked by indigenous leadership, and Anglicanism has thus become as much a non-Western as a Western form of Christianity.

Relations with other churches. Because the Anglican Communion consists of a cluster of related churches, it does not, as a worldwide Communion, have membership in the World Council of Churches; each of the Anglican churches, however, holds such membership. This type of ecumenical relationship is in keeping with one of the consistent goals of Anglicanism. Anglicans see themselves as catalysts for Christian unity, and the Anglican blend of Catholic liturgy and Protestant procedure affords the basis of broad ecumenical encounter. Within Anglicanism there is a common point with virtually all other expressions of the Christian faith. Anglicans readily engage Roman Catholic, Orthodox, and Protestant leaders in theological discussion and joint liturgy. Ecumenical processes involving the Catholic Church have been regular and intensive, though without prospect of organic reunion. The Anglican/Roman Catholic International Theological Commission has met regularly as have committees involving the Lutheran and Reformed traditions. For Anglicans ecumenical discussion is the appropriate context for advancing Christian mission.

World Anglicanism. The Anglican Communion has tried to establish itself as the middle way in Christianity,

attempting to bridge the gulfs between Protestant, Roman Catholic, and Eastern Orthodox churches. In 1947 Anglican dioceses were included in the new Church of South India, a communion that also included mission churches of the Methodists and Congregationalists. In other areas the Anglican Communion has special interchurch relations, as with the Lusitanian Church in Portugal, the Mar Thoma Syrian Church in India, the Old Catholic churches in Europe and the United States, the Philippine Independent Church, and the Spanish Reformed Episcopal Church. In the United States Anglicans took part in the Consultation on Church Union. In 1974 the Church of England and English Roman Catholics, Baptists, United Reformed, and Methodists agreed to form a national commission for discussions about practical reunion. Statements issued by Archbishop of Canterbury Robert Runcie and Pope John Paul II following a historic meeting between the two in England in 1982 emphasized the importance of the reconciliation effort. (R.S.De./W.L.Sa.)

Baptists

Baptists are Protestant Christians who share the basic beliefs of most Protestants but who insist that only believers should be baptized and that it should be done by immersion rather than by the sprinkling or pouring of water. (This view, however, is shared by others who are not Baptists.) Although Baptists do not constitute a single church or denominational structure, most adhere to a congregational form of church government. Some Baptists lay stress upon having no human founder, no human authority, and no human creed.

HISTORY

Origins. Some Baptists believe that there has been an unbroken succession of Baptist churches from the days of John the Baptist and the Apostles of Christ. Others trace their origin to the Anabaptists (a 16th-century Protestant movement; see above *History of the Protestant movement*) on the European continent. Most scholars, however, agree that Baptists, as an English-speaking denomination, originated within 17th-century Puritanism as an offshoot of Congregationalism.

Particular Baptists and General Baptists

There were two groups in early Baptist life: the Particular Baptists and the General Baptists. The Particular Baptists adhered to the doctrine of a particular atonement—that Christ died only for an elect—and were strongly Calvinist (following the Reformation teachings of John Calvin) in orientation; the General Baptists held to the doctrine of a general atonement—that Christ died for all people and not only for an elect—and represented the more moderate Calvinism of Jacobus Arminius, a 17th-century Dutch theologian. The two currents were also distinguished by a difference in churchmanship related to their respective points of origin. The General Baptists had emerged from the English Separatists, whereas the Particular Baptists had their roots in non-Separatist independency.

Both the Separatists and the non-Separatists were congregationalist. They shared the same convictions with regard to the nature and government of the church. They believed that church life should be ordered according to the pattern of the New Testament churches, and to them this meant that churches should be self-governing bodies composed of believers only.

They differed, however, in their attitude toward the Church of England. The Separatists contended that the Church of England was a false church and insisted that the break with it must be complete. The non-Separatists, more ecumenical in spirit, sought to maintain some bond of unity among Christians. While they believed that it was necessary to separate themselves from the corruption of parish churches, they also believed that it would be a breach of Christian charity to refuse all forms of communication and fellowship. While many non-Separatists withdrew and established a worship of their own, they would not go so far as to assert that the parish churches were devoid of all marks of a true church.

Growth in England and abroad. Although the Particular Baptists were to represent the major continuing Baptist tradition, the General Baptists were first to appear. In 1608 religious persecution induced a group of Lincolnshire Separatists to seek asylum in Holland. A contingent settled in Amsterdam with John Smyth (or Smith), a Cambridge graduate, as their minister; another group moved to Leiden under the leadership of John Robinson. When the question of baptism arose during a debate on the meaning of church membership, Smyth concluded that, if the Separatist contention that "churches of the apostolic constitution consisted of saints only" was correct, then baptism should be restricted to believers only. This, he contended, was the practice of the New Testament churches, for he could find no scriptural support for baptizing infants. Smyth published his views in *The Character of the Beast* (1609) and in the same year proceeded to baptize first himself and then 36 others, who joined him in forming a Baptist church. Shortly thereafter Smyth became aware of a Mennonite (Anabaptist) community in Amsterdam and began to question his act of baptizing himself. This could be justified, he concluded, only if there was no true church from which a valid baptism could be obtained. After some investigation Smyth recommended union with them. This was resisted by Thomas Helwys and other members of the group, who returned to England in 1611 or 1612 and established a Baptist church in London. The parent group in Amsterdam soon disappeared.

The question of baptism

The Particular Baptists stemmed from a non-Separatist church that was established in 1616 by Henry Jacob at Southwark, across the Thames from London. In 1638 a number of its members withdrew under the leadership of John Spilsbury to form the first Particular Baptist Church.

The two decades from 1640 to 1660 constituted the great period of early Baptist growth. Baptist preachers won many adherents around the campfires of the Puritan leader Oliver Cromwell's army. The greatest gains were made by the Particular Baptists, while the General Baptists suffered defections to the Quakers. After the Restoration of the Stuarts in 1660 both groups were subjected to severe disabilities until these were somewhat relaxed by the Act of Toleration of 1689.

During the following decades the vitality of the General Baptists was drained by the inroads of skepticism, and their churches generally dwindled and died or became Unitarian. The Particular Baptists retreated into a defensive, rigid hyper-Calvinism. Among the Particular Baptists in England renewal came as a result of the influence of the Evangelical Revival, with a new surge of growth initiated by the activity of the English Baptist clergymen Andrew Fuller, Robert Hall, and William Carey. Carey, in 1792, formed the English Baptist Missionary Society—the beginning of the modern foreign missionary movement in the English-speaking world—and became its first missionary to India. A New Connection General Baptist group, Wesleyan in theology, was formed in 1770, and a century later, in 1891, it united with the Particular Baptists to form the Baptist Union of Great Britain and Ireland.

Survival and expansion of Particular Baptists

By the end of the 19th century Baptists, together with the other Nonconformist churches, were reaching the peak of their influence in Great Britain, numbering among their preachers several men with international reputations. Baptist influence was closely tied to the fortunes of the Liberal Party, of which the Baptist David Lloyd George was a conspicuous leader. After World War I English Baptists began to decline in influence and numbers.

Baptist churches were established in Australia (1831) and New Zealand (1854) by missionaries of the English Baptist Missionary Society. In Canada, Baptist beginnings date from the activity of Ebenezer Moulton, a Baptist immigrant from Massachusetts who organized a church in Nova Scotia in 1763. In Ontario the earliest Baptist churches were formed by United Empire Loyalists who crossed the border after the American Revolution, while other churches were established by immigrant Baptists from Scotland and by missionaries from Vermont and New York.

Development in the United States. Baptist churches in the English colonies of North America were largely indigenous in origin, being the product of the leftward movement that was occurring among the colonial Puritans at

the same time as it was in England. While some emigrants went to the New World as Baptists, it was more typical for them to adopt Baptist views after their arrival in the colonies, as happened in the case of Henry Dunster, the first president of Harvard College, and Roger Williams.

Colonial period. The first Baptist Church in North America was established at Providence in 1639 by Roger Williams shortly after his banishment from the Massachusetts Bay Colony. Although Williams' general Calvinist theological position was roughly analogous to that of Spilsbury, prior to becoming a Baptist he had adopted the narrower Separatist view of the church. Williams soon came to the conclusion that all churches, including the newly established church at Providence, lacked a proper foundation, and that this defect could be remedied only by a new apostolic dispensation, when new apostles would appear to reestablish the true church.

The defection of Williams left the church with no strong leadership and thus made it possible for it to be reorganized on a General Baptist platform in 1652. There was scattered General Baptist activity throughout the colonies, but the only large cluster of General Baptists was in Rhode Island, where the churches were united into an association in 1670. The early General Baptists never gained great strength. Most of their churches decayed, and some, including the Providence church, were reorganized as Particular Baptist churches. The half dozen churches that survived never entered the mainstream of American Baptist life and exerted little influence upon its development.

The earliest strong Particular Baptist centre in the colonies was at Newport, R.I., where, between 1641 and 1648, a church that had been gathered by the physician and minister John Clarke adopted Baptist views. Except for a church that had a brief existence at Kittery, Maine, there were only two other Particular Baptist churches in New England for the better part of a century. One was at Swansea, Mass.; the other was organized at Boston in 1665. Another Particular Baptist church was established at Charleston, S.C., in 1683 or 1684.

The centre of Particular Baptist activity in early America was in the Middle Colonies. In 1707 five churches in New Jersey, Pennsylvania, and Delaware were united to form the Philadelphia Baptist Association, and through the association they embarked upon vigorous missionary activity. By 1760 the Philadelphia association included churches located in the present states of Connecticut, New York, New Jersey, Pennsylvania, Delaware, Virginia, and West Virginia; and by 1767 further multiplication of churches had necessitated the formation of two subsidiary associations, the Warren in New England and the Ketochton in Virginia. The Philadelphia association also provided leadership in organizing the Charleston Association in the Carolinas in 1751.

Philadelphia Baptist Association

Although this intercolonial Particular Baptist body provided leadership for the growth that characterized American Baptist life during the decades immediately preceding the American Revolution, that growth was largely a product of an 18th-century religious revival known as the Great Awakening. Though they participated directly in the Awakening only during its last phase in the South, Baptists attracted large numbers of recruits from among those who had been "awakened" by the preaching of others. In addition to strengthening and multiplying the "regular" Baptist churches, the Awakening in New England produced a group of revivalistic Baptists, known as Separate Baptists, who soon coalesced with the older New England Baptist churches. In the South, however, they maintained a separate existence for a longer period of time. Shubael Stearns, a New England Separate Baptist, migrated to Sandy Creek, N.C., in 1755 and initiated a revival that quickly penetrated the entire Piedmont region. The churches he organized were brought together in 1758 to form the Sandy Creek Association. Doctrinally these churches did not differ from the older "regular" Baptist churches, but what the older churches saw as their emotional excesses and ecclesiastical irregularities created considerable tension between the two groups. By 1787, however, a reconciliation had been effected.

In several of the colonies, Baptists laboured under legal disabilities. The public whipping of Obadiah Holmes in 1651 for his refusal to pay a fine that had been imposed for holding an unlawful meeting in Lynn, Mass., caused John Clarke to write his *Ill News from New England* (1652). Fourteen years later Baptists of Boston were fined, imprisoned, and denied the use of a meetinghouse they had erected. Payment of taxes for support of the established church was a cause of continuing controversy in New England, while the necessity to secure licenses to preach became an inflammatory issue in Virginia.

In the 19th century. The problem of travel had made it difficult for the Philadelphia association to serve as a bond uniting Baptists, and the rapid multiplication of churches made it impossible. It has been estimated that immediately before the American Revolution there were 494 Baptist congregations; 20 years later, in 1795, Isaac Backus estimated the number at 1,152. The initial expedient of the Philadelphia association had been to organize subsidiary associations, but during the war the churches, left to their own devices, proceeded to organize independent associations. By 1800 there were at least 48 local associations, and the main problem was to fashion a national body to unite the churches. The final impetus in this direction came from an interest in foreign missions. Among the first missionaries of the newly organized Congregational mission board were Adoniram Judson and Luther Rice, who had been sent to India. On shipboard they became convinced by a study of the Scriptures that only believers should be baptized. Upon arrival at Calcutta, Judson went on to Burma, while Rice returned home to enlist support among American Baptists. As a result of Rice's efforts a General Convention of the Baptist denomination was formed in 1814. Its scope was almost immediately broadened to include, in addition to the foreign mission interest, a concern for home missions, education, and the publication of religious periodicals. In 1826 the General Convention once again was restricted to foreign mission activities, and in the course of time it became known as the American Baptist Foreign Mission Society. Other denominational interests were served by the formation of additional societies with similar specialized concerns, such as the American Baptist Home Mission Society and the American Baptist Publication Society.

Formation of General Convention

The unity achieved through these societies was disrupted by the slavery controversy. During the decade prior to 1845 various compromises between the proslavery and antislavery parties in the denomination were attempted, but they proved to be unsatisfactory. As a result a Southern Baptist Convention was organized at Augusta, Ga., in 1845. Although its constitution provided for boards of home and foreign missions, education, and publication, its energies were devoted largely to foreign missions. Consequently, the American Baptist Home Mission Society and the American Baptist Publication Society continued to operate in the South after the Civil War. Later the Southern Baptist Convention began to develop its own home mission and publication work and to protest the intrusion of the older societies in the South. The final separation between Baptists of South and North was formalized in 1907 by the organization of the Northern Baptist Convention (in 1950 renamed the American Baptist Convention and after 1972 called the American Baptist Churches in the U.S.A.), which brought together the older societies and accepted a regional allocation of territory between the Northern and Southern conventions.

Development of black churches. Black churches constitute a major segment of American Baptist life. Many slaves were converted and became members of Baptist churches during the Great Awakening. While there were black Baptist churches prior to the Civil War, they rapidly multiplied following the Emancipation Proclamation (1863), an edict freeing the slaves in the United States. State and regional conventions were formed, and the National Baptist Convention was organized in 1880. By 1900 black Baptists outnumbered black adherents of all other denominations. Throughout the Jim Crow years of segregation and exclusion from most aspects of American life, black churches were the focal point of black communal life. In the civil rights struggle of the 1960s the major leadership,

National Baptist Convention

including that provided by Martin Luther King, Jr., came out of black churches.

Developments in education. From the beginning, American Baptists displayed an interest in an educated ministry. The Philadelphia association in the 18th century collected funds to help finance the education of ministerial candidates. Hopewell Academy was established in 1756, and in 1764 Brown University was founded in Rhode Island. After 1800, educational institutions multiplied rapidly. The educational advance culminated in 1891 in the founding of the University of Chicago.

During the 20th century. After 1900, Baptists were troubled by theological controversies that led to the formation of several new Baptist groups. Some of the tensions arose over questions of structure of church organization, some arose over refusals to adopt an authoritative creedal statement, some were created by converts among new immigrants, and some were the product of dissatisfaction with the affiliation of the American Baptist Convention with interdenominational and ecumenical bodies. Questions of organizational structure were involved in the formation of the American Baptist Association in 1905 by churches located primarily in Oklahoma, Texas, and Arkansas. Two other groups were products of the Fundamentalist controversy: the General Association of Regular Baptist Churches, organized in 1932, and the Conservative Baptist Association of America (1947).

During the post-World War II, period the Southern Baptist Convention abandoned its regional limitations. Because of increasing mobility of population, it became necessary for the convention to follow its members to the growing urban centres of the North and West. By the second half of the 20th century Southern Baptists had become the largest Protestant body in the United States, and their churches were located in every part of the country.

Following World War II, Southern Baptists increasingly isolated themselves from other Christian churches, feeling no need to cooperate with them in common enterprises. During these years they also developed centralized operations through the boards and agencies of the Convention. Participation in the "Cooperative (mission) Program" and utilization of the materials and activities supplied by the Sunday School Board became badges of loyalty. These programs were carefully devised and eminently successful in promoting numerical growth.

Fundamentalist Southern Baptists

Meanwhile, dissident Southern Baptists, based initially in the old southwest of Tennessee, Mississippi, Louisiana, Arkansas, and especially Texas, began to become influential elsewhere. They were heirs of an older isolationism that had long been kept in check but gained major new impetus from a radical fundamentalism developing strength in the South after World War II. Led by a small coterie of Texas strategists, the dissidents put a plan into operation in 1979 by which they gained control of and imposed their views on the bureaucracy and theological seminaries of the Southern Baptist Convention. No room for a difference of opinion was left except at the local level.

Growth outside the United States. While Baptists have been troubled by divisive tendencies during the 20th century, there has also been a tendency toward greater unity and cohesiveness through the Baptist World Alliance. The 19th century was a period of great Baptist missionary activity. The endeavour in Asia was led by William Carey in India, Adoniram Judson in Burma, and Timothy Richard and Lottie (Charlotte) Moon in China. The initial Baptist presence in Africa began in 1793 when David George, a former slave from South Carolina, reached Sierra Leone by way of Halifax, N.S. More organized activity was initiated in 1819 by black Baptists of Richmond, Va., who sent Lott Cary to Sierra Leone in 1821 and then shifted his base of operations to Liberia in 1824. By the late 20th century there were major concentrations of Baptists in Zaire, Nigeria, and Cameroon. Of later origin is the Baptist community in Latin America.

The pioneer Baptist in Europe was Johann Gerhardt Oncken, who organized a church at Hamburg in 1834. Oncken had become acquainted with Barnas Sears of Colgate Theological Seminary, who was studying in Germany, and with six others he was baptized by Sears. From this centre, evangelistic activity was extended throughout Germany, and missions were established elsewhere in eastern Europe. Baptist activity was initiated independently in France, Italy, and Spain. Swedish Baptist beginnings date from the conversion of Gustaf W. Schroeder, a sailor baptized in New York in 1844, and Frederick O. Nilsson, also a sailor, who was baptized by Oncken in 1847.

The expansion of the Baptist community in Asia, Africa, Latin America, and Europe led to the formation of the Baptist World Alliance in London in 1905. The purpose of the alliance is to provide mutual encouragement, exchange of information, coordination of activities, and consciousness of the larger Baptist fellowship.

Developments in Russia

The most notable growth occurred in Russia, where a Russian Baptist Union was formed in 1884 as the result of influences stemming from Oncken. Another Baptist body, the Union of Evangelical Christians, was organized in 1908 by a Russian who had come under the influence of English Baptists. Persecution of Baptists, which had been severe, was relaxed in 1905, and within the remaining disabilities a moderate growth occurred. The Revolution of 1917, with its proclamation of liberty of conscience, marked the beginning of a period of astonishing advance: by 1927 the Russian Baptist Union numbered some 500,000 adherents, while the Union of Evangelical Christians embraced more than 4,000,000. The Soviet constitution of 1929 subjected them to pressure once again, however. Membership in the two groups, which combined in 1944 to form the All-Union Council of Evangelical Christians–Baptists in the U.S.S.R., declined sharply, but an estimated membership of more than 500,000 in the 1980s testified to the tenacity with which these believers held their faith.

TEACHINGS

History. Initially Baptists were characterized theologically by strong to moderate Calvinism. The dominant continuing tradition in both England and the United States was Particular Baptist. By 1800 this older tradition was beginning to be replaced by evangelical doctrines fashioned by the leaders of the evangelical revival in England and the Great Awakening in the United States. By 1900 the older Calvinism had almost completely disappeared, and Evangelicalism was dominant. The conciliatory tendency of Evangelicalism and its almost complete preoccupation with "heart religion" and the experience of conversion largely denuded it of any solid theological structure, thereby opening the door to a new theological current that subsequently became known as Modernism. Modernism, which was an attempt to adjust the Christian faith to the new intellectual climate, made large inroads among the Baptists of England and the United States during the early years of the 20th century, and Baptists provided many outstanding leaders of the movement, including Shailer Mathews and Harry Emerson Fosdick. Many people regarded these views as a threat to the uniqueness of the Christian revelation, and the counterreaction that was precipitated became known as Fundamentalism (a movement emphasizing biblical literalism).

As a result of the controversy that followed, many Baptists developed a distaste for theology and became content to find their unity as Baptists in promoting denominational enterprises. By 1950, outside the South, both Modernists and Fundamentalists were becoming disenchanted with their positions in the controversy, and it was from among adherents of both camps that a more creative theological encounter began to take place. While the majority of Baptists remained nontheological in their interests and concerns, there were many signs that Baptist leadership was increasingly recognizing the necessity for renewed theological inquiry.

Contents. The unity and coherence of the Baptists is based on six distinguishing, although not necessarily distinctive, convictions they hold in common.

Central aspects of Baptist doctrine

1. The supreme authority of the Bible in all matters of faith and practice. Baptists are a non-creedal people, and their ultimate appeal always has been to the Scriptures rather than to any confession of faith that they may have published from time to time to make known their commonly accepted views.

2. Believer's baptism. This is the most conspicuous conviction of Baptists. They hold that if baptism is the badge or mark of a Christian, and if a Christian is a believer in whom faith has been awakened, then baptism rightly administered must be a baptism of believers only. Furthermore, if the Christian life is a sharing in the life, death, and resurrection of Christ, if it involves a dying to the old life and a rising in newness of life, then the act of baptism must reflect these terms. The sign must be consonant with that which it signifies. It is for this latter reason that Baptists were led to insist upon immersion as the apostolic form of the rite.

3. Churches composed of believers only. Baptists reject the idea of a territorial or parish church and insist that a church is composed only of those who have been gathered by Christ and who have placed their trust in him. Thus the membership of a church is restricted to those who—in terms of a charitable judgment—give clear evidence of their Christian faith and experience.

4. Equality of all Christians in the life of the church. By the doctrine of the priesthood of all believers Baptists not only understand that the individual Christian may serve as a minister to other members but also that each church member has equal rights and privileges in determining the affairs of the church. Pastors have special responsibilities, derived from the consent of the church, which only they can discharge, but they have no unique priestly status.

5. Independence of the local church. By this principle Baptists affirm that a properly constituted congregation is fully equipped to minister Christ and need not derive its authority from any source, other than Christ, outside its own life. Baptists, however, have not generally understood that a local church is autonomous in the sense that it is isolated and detached from other churches. As individual Christians are bound to pray for one another and to maintain communion with one another, so particular churches are under similar obligation. Thus, the individual churches testify to their unity in Christ by forming associations and conventions.

6. Separation of church and state. From the time of Smyth, Baptists have insisted that a church must be free to be Christ's church, determining its own life and charting its own course in obedience to Christ without outside interference. Thus Smyth asserted that the

> magistrate is not by virtue of his office to meddle with religion or matters of conscience, to force and compel men to this or that form of religion or doctrine, but to leave Christian religion free to every man's conscience.

Baptists were in the forefront of the struggle for religious freedom in both England and the United States. They cherished the liberty established in early Rhode Island, and they played an important role in securing the adoption of the "no religious test" clause in the U.S. Constitution and the guarantees embodied in the First Amendment.

Few Baptists have been willing to become so sectarian as to deny the Christian name to other denominations. With the exception of the Southern Baptists, most Baptists cooperate fully in interdenominational and ecumenical bodies, including the World Council of Churches.

WORSHIP AND ORGANIZATION

Baptist worship is hardly distinguishable from the worship of the older Puritan denominations (Presbyterians and Congregationalists) of England and the United States. It centres largely on the exposition of the Scriptures in a sermon and emphasizes extemporaneous, rather than set, prayers. Hymn singing also is one of the characteristic features of worship. Communion, received in the pews, is customarily a monthly observance.

Baptists insist that the fundamental authority, under Christ, is vested in the local congregation of believers, which admits and excludes members, calls and ordains pastors, and orders its common life in accord with what it understands to be the mind of Christ. These congregations are linked together in cooperative bodies—regional associations, state conventions, and national conventions—to which they send their delegates or messengers. The larger bodies, it is insisted, have no control or authority over a local church; they exist only to implement the common concerns of the local churches.

The pattern of organization of the local church has been undergoing change during the 20th century. Traditionally the pastor was the leader and moderator of the congregation, but there has been a tendency to regard the pastor as an employed agent of the congregation and to elect a lay member to serve as moderator at corporate meetings of the church. Traditionally the deacons' functions were to assist the pastor and to serve as agents to execute the will of the congregation in matters both temporal and spiritual; there has been a tendency, however, to multiply the number of church officers by the creation of boards of trustees, boards of education, boards of missions, and boards of evangelism. Traditionally decisions were made by the congregation in a church meeting, but there has been a tendency to delegate decision making to various boards. The relationship of local churches to the cooperative bodies has been undergoing similar change, and this has occasioned continuing discussion among all Baptist groups. (W.S.H.)

Congregationalists

Congregationalists are members of a group of churches that arose in England in the late 16th and 17th centuries. Originally they were frequently called Independents, as they still are in Welsh-speaking communities. The main centres of Congregationalism traditionally were in Britain and the United States, but in the 20th century Congregationalists have joined with others to form united churches in these and several other countries.

Congregationalism has occupied a position among the churches somewhere between Presbyterianism and the more radical Protestants, such as non-Fundamentalist Baptists and Quakers. Its distinctive emphasis has been on the right and responsibility of each properly organized congregation to make its own decisions about its own affairs, without having to submit them to the judgment of any higher human authority. Although this was not always true in the early days in America, Congregationalists have generally been distrustful of state establishment of religion and have been workers for civil and religious liberty. Their emphasis on the rights of the particular congregation and on freedom of conscience arose historically from their strong Protestant convictions concerning the sovereignty of God and the priesthood of all believers. This attitude has given them an openness of outlook that has led many of them to theological and social liberalism and to active participation in the ecumenical movement.

HISTORY

England. The "Congregational way" came into prominence in English life during the 17th-century Civil War, but its origins lie in 16th-century Separatism. Robert Browne is sometimes taken as its founder, but he was an erratic character who changed his views more than once. Congregational ideas were in the air, finding expression independently of him. The Separatists (those advocating separation from rather than reform of the Church of England) were severely persecuted under Elizabeth I; three of them—John Greenwood, Henry Barrow, and John Penry—were the first Congregational martyrs. Some of the Separatists settled in Holland to escape persecution, and it was from among these that the *Mayflower* Separatists later set sail for the New World (see below, *United States*).

The Separatists

At the time of the Long Parliament, beginning in 1641, many exiles returned to England, and the Independents, as they were now called, became increasingly active. They were particularly influential in the army, having Oliver Cromwell himself associated with them. They began to move away from the Presbyterians, with whom they had initially cooperated, and to draw closer to the Baptists and the Fifth Monarchy Men (a Puritan millennialist sect). They reached the peak of their influence during the Commonwealth in the 1650s, and their leaders, Hugh Peter, John Owen, and Thomas Goodwin, held positions of eminence. With the death of Cromwell (1658) they lacked the conviction and power of initiative to hold the country

together, and in the confused period before the recall of King Charles II in 1660 their political influence collapsed.

The advent of Charles II was a disaster for Congregationalists, and the Act of Uniformity of 1662 was the first of a series of determined efforts to root them out from English life. "Black Bartholomew," St. Bartholomew's Day, Aug. 24, 1662, when some 2,000 ministers of various Protestant groups who rejected the authority of the Church of England were ejected from their livings, has always been regarded as a great turning point in the history of English Dissent. All Nonconformists were subjected to a persecution that, although severe, was not so intense as to imperil their existence. In this time John Owen and others produced some of the classical statements of Congregational belief; John Milton produced his greatest poems; and John Bunyan, although his closest affinities were with the Baptists, imprinted some of the characteristic religious attitudes of the Dissenters indelibly on the English consciousness.

Persecution of Nonconformists

The accession of William and Mary in 1688 and the consequent Toleration Act of 1689 meant that the survival of the Congregationalists was assured, although still under civil disabilities. Their fears were renewed by the advent of Queen Anne (1702). The Occasional Conformity Act (1711) forbade Dissenters from qualifying for public office by occasionally taking communion at the Anglican parish church, and the Schism Act (1714) was directed against their schools. The death of Queen Anne in 1714, before the Schism Act could be fully implemented, was considered providential by the Dissenters. They supported the new regime and the Whig ascendancy and for the next 50 years enjoyed a modest prosperity. Most of them belonged to the economically independent sections of society and lived in London and the older provincial towns. They were especially active in education. After 1662 Dissenters were debarred from the universities, and many ejected ministers started small schools and colleges called academies, which gradually became more numerous and influential. Their curricula, influenced by the educational theories of Francis Bacon and John Amos Comenius, were more relevant than those of the comatose universities, and they were the precursors of many later educational developments.

Religious zeal was declining as the 17th century waned, and rationalism became more influential. Deism and Arianism (a heresy denying the divinity of Christ) were widespread, the latter especially among the Presbyterians, some of whom gradually became Unitarian. That Congregationalism did not go the same way was in no small measure due to the influence of Philip Doddridge, minister of Northampton, who was a theologian, pastor, social reformer, educationist, and author of the devotional classic *The Rise and Progress of Religion in the Soul* (1745).

The quality of Congregationalism in the early 18th century has sometimes been disparaged, but its limitations were those of a small community in the aftermath of a period of great intensity of experience. A change came with the rise of Methodism and the Evangelical Revival (*c.* 1750–1815), which had a profound, if unobtrusive, influence on Congregationalism. Many ministers were deeply affected by the revival, and many people who were reached by the Methodist preaching found their way into the already existing Congregational churches. Thus the great evangelist George Whitefield had close relations with Congregationalism, and many of the churches founded by Selina Hastings, countess of Huntingdon, a leading figure in the revival, made and long retained a connection with Congregationalism. By 1815 the character of Congregationalism had been significantly changed in an Evangelical direction, especially in the developing industrial areas of Lancashire and Yorkshire.

Influence of Methodism and Evangelical Revival

The outstanding result of the Evangelical Revival in Congregationalism was the founding of the London Missionary Society (1795). Its purpose was not so much the spreading of Congregationalism overseas as the proclaiming of "the glorious gospel of the blessed God," leaving the churches it founded to find their own form. Its main support was always Congregational, and it has now been incorporated into the Council for World Mission of the United Reformed Church. Through its agency, churches have been established in Africa, Madagascar, India, China, Papua New Guinea, and on islands in the South Seas. Many of these are now united in wider bodies, of which the most notable is the Church of South India.

In the first half of the 19th century Congregationalism was involved in a period of expansion and consolidation. Increased numbers brought many poorer people into the churches, and a new political and social radicalism began to emerge. Voluntarism, which opposed the state support of denominational education, and the Liberation Society, which advocated disestablishment, found widespread support. The Congregational Union of England and Wales, linking the churches in a national organization, was founded in 1832 and the Colonial (later the Commonwealth) Missionary Society, for promoting Congregationalism in the English-speaking colonies, in 1836.

Congregational churches shared fully in the ecclesiastical prosperity of the Victorian era. Many new buildings were erected, often in ambitious Gothic style, and the cult of the popular preacher developed. Able ministers, among whom R.W. Dale of Birmingham was outstanding, deeply influenced the public life of Victorian cities. The links of the churches with the Liberal Party were greatly strengthened, and the civic disabilities of Dissenters were steadily removed. Thriving churches in new suburbs developed into hives of social, philanthropic, and educational activity. The picture of the philistine (unimaginative) Dissenters drawn by the poet and critic Matthew Arnold in *Culture and Anarchy* (1869) contained a measure of truth, but the work's lack of historical perspective led it to underestimate the zeal for self-improvement and the desire for a richer life that existed in Victorian Congregationalism.

Links with the Liberal Party

The Liberal victory of 1906 represented the peak of the social and political influence of Congregationalism. After that, Congregational churches shared in the institutional decline of most British churches, but they continued to show theological and cultural vitality. In October 1972 the majority of English Congregationalists and Presbyterians united to form the new United Reformed Church, which was joined in 1981 by the Churches of Christ, the small British counterpart of the American Disciples of Christ.

Wales, Ireland, and Scotland. Welsh-speaking Congregational churches did not join the United Reformed Church but have their separate organization in the Union of Welsh Independents. These churches grew up originally in the countryside but transplanted themselves with remarkable success to the developing industrial valleys in the 19th century. The churches have been strong centres of distinctively Welsh culture, and their ministers have often been national leaders. Their influence diminished in the 20th century as population moved away from old centres of strength, but Welsh Congregationalists maintain their tradition of preaching, poetry, and hymnody.

Congregationalism in Scotland has been less prominent, and in Ireland it has struck only a very small root. In Scotland it arose in the 19th century out of dissatisfaction with the lack of missionary zeal of the Church of Scotland and soon united with a similar group called the Evangelical Union. Numerically small, it has made a distinctively liberal contribution to Scottish life and has given many notable sons to the church-at-large, among them the missionaries David Livingstone and Robert Moffat and the writer George Macdonald, as well as Peter Taylor Forsyth.

United States. It was in the United States that Congregationalism achieved its greatest public influence and numerical strength, and, through the New England experiment, in setting up communities based on Congregational-type religious principles, it was a major factor in determining the character of the nation. The New England settlement had two roots, in the Separatism of Plymouth Colony and in the Puritanism of Massachusetts Bay. The first Separatists came on the *Mayflower* in 1620 from the exiled church at Leiden, Holland. The Puritans wished to reform the Church of England rather than to leave it, and they left England in order to build a "godly commonwealth" that would be an example to old England of what a new England, truly reformed according to the Word of God, might be. They were closer in spirit to the English Presbyterians than to the Separatists, but there

The New England experiment

was enough affinity between the two groups to enable them to live together in comparative harmony and to reject more radical leaders such as Roger Williams and Anne Hutchinson. In 1648 the two groups united to produce the Cambridge Platform, a declaration of faith that accepted the theological position of the Westminster Confession but maintained a Congregational polity. (The English Congregationalists produced a similar statement, the Savoy Declaration, in 1658.)

The original experiment demanded a radical commitment of an intellectual and spiritual intensity that made the New England colony unique in history. As the community became established and a second generation grew up, it became difficult to maintain the high standard, and the rigorous conditions for church membership had to be relaxed. This need found expression in the famous Half-Way Covenant, which said that those who had been baptized but could not enter into full church membership on the basis of the kind of religious experience considered appropriate were accepted as church members but not admitted to communion or allowed to have voting rights.

The community was keenly interested in education from the outset, and one of its earliest acts was to start a college to maintain the succession of learned ministers. Thus was founded Harvard College (1636), the first of a long line of colleges begun under Congregational auspices in America.

The Great Awakening

The gradual loss of religious fervour caused great distress and self-questioning to the Congregational leaders, but a quickening of new life came with the 18th-century Great Awakening, the widespread revival movement that started in 1734 under the influence of Jonathan Edwards. The Awakening, however, threw into relief the differences emerging between two wings in Congregationalism. On the one side were those who maintained the Calvinist tradition, creatively restated by Edwards and his followers, with a greater emphasis on the affective elements in religion. On the other was a rapidly growing Unitarianism, parallel to a similar movement in England. By the early 19th century many of the oldest Congregational churches had become Unitarian, including 12 of the 14 in Boston. Unitarianism was not so prevalent in Connecticut, where Congregationalism had quickly taken root and remained the established church until well into the 19th century.

Although the loss to Unitarianism was serious, Congregationalism remained vigorous in the 19th century and was active in the westward expansion of the nation. The Presbyterians were almost nonexistent in New England but strong in the Middle Atlantic states, where Congregationalism had little root. The two bodies adopted a Plan of Union in 1801 for joint missionary activity in the developing territories. One of the reasons for the ultimate breakdown of this arrangement after half a century was the growing liberalism of Congregationalism. The characteristic theologian of this period was Horace Bushnell, who challenged the traditional substitutionary view of the Atonement (that Christ's suffering and death atoned for man's sins), and whose well-known book, *Christian Nurture* (1847), questioned the necessity of the classical conversion experience. Such influential preachers as Henry Ward Beecher and Washington Gladden popularized similar ideas. The so-called Kansas City Creed of 1913 summed up the liberalism of this period, which represented a radical break with the Calvinist past.

American Congregationalists have engaged in widespread missionary activity, particularly in the Middle East and in China before the Communist Revolution. A national Congregational organization was formed in 1871, and powerful Boards for Home Missions and Education were established, through which Northern Congregationalists did a great deal for black education in the South, where there were hardly any indigenous Congregational churches.

Modern American Congregationalism has shown itself singularly ready to unite with other churches. Union with a relatively small body called the Christian Church, which was concentrated in the upper South, was achieved between the world wars, and a more notable union was achieved with the Evangelical and Reformed Church in 1961. This was a strong community of German Lutheran and Reformed background, which claimed the eminent theologians Reinhold Niebuhr and Paul Tillich among its ministers. The new church body is known as the United Church of Christ. A minority of Congregational churches refused to join the union, and these remain separate.

United Church of Christ

Congregationalism has not succeeded in becoming a popular worldwide form of church life, although it has been represented in most English-speaking countries. Congregationalists were prominent in the formation of the Church of South India in 1947. They have also become part of the United Church of Canada and of the Uniting Church in Australia. Through the International Congregational Council, united with the Reformed Alliance since 1970, they have had fraternal ties with churches of similar outlook in Europe, notably the Remonstrant Brotherhood of Holland and the Swedish Mission Covenant Church.

TEACHINGS

Throughout their history Congregationalists have shared the faith and general outlook of evangelical Protestantism in the English-speaking countries, but normally in a more liberalized way than would be customary among their nearest neighbours, the Presbyterians, the Methodists, and the Baptists. The English historian Bernard Manning once described their position as decentralized Calvinism, in contrast to the centralized Calvinism of Presbyterians. That description contains much truth about their doctrines and general outlook until well into the 19th century, but it underestimates the Congregational emphasis on the free movement of the Spirit. This provides a link with the Quakers and partly explains the Congregational distrust of giving binding authority to creedal statements. The other part of their distrust is explained by their anxiety to accord supreme authority to Scripture. They have not been slow to produce declarations of faith. In addition to the Savoy Declaration, the Cambridge Platform, and the Kansas City Creed already mentioned, lengthy statements have also been produced both by the United Church of Christ and by the English Congregationalists. No great authority is claimed for any of these, and in recent generations most Congregationalists have regarded the primitive confession, "Jesus is Lord," as a sufficient basis for membership.

Distrust of binding creedal statements

Similarly, they have always stressed the importance of freedom. Even in the days of their Cromwellian triumph they were tolerant by the standards of the time, and through the activities of the Protestant Dissenting Deputies, who had the right of direct access to the monarch, they contributed greatly in the 18th century to the establishment of the rights of minorities in England. Both in England and America the long-faced and repressive Puritan of tradition owes as much to the caricatures of political opponents and literary rebels as to actual fact.

WORSHIP AND ORGANIZATION

Practices. Congregationalism has always attached importance to preaching because the Word of God as declared in Scripture is regarded as constitutive of the church. Baptism and the Lord's Supper are considered to be the only sacraments instituted by Christ. Infants are baptized, normally by sprinkling. The Lord's Supper is normally celebrated once or twice a month and has not always been given a central place, often following a preaching service after a brief interval in which many of the congregation leave. In recent times, the unity of sermon and sacrament as parts of the same service has been much more strongly emphasized, and there has been a tendency to assimilate Congregational and Presbyterian practice to each other. Traditionally public prayer has been extempore, but in the 20th century service books and set forms have been increasingly used. Since the 18th century and the work of the great Congregationalist hymn writer Isaac Watts, hymns have featured prominently in Congregational worship. The English compilation, *Congregational Praise* (1951), worthily maintained the tradition.

Polity. The distinctive organizational tenet of Congregationalism has been that of the spiritual autonomy of the particular congregation. The congregation, however, is not thought of as any casual gathering of Christians but as a settled body with a well-defined constitution and proper offices that has tried to order itself in harmony with the

The Congregational principle

New Testament understanding of the nature of the church. The claim is made that if a church in a particular place possesses the Bible, the sacraments, a properly called and appointed minister and deacons, and members who have made a genuine Christian profession, no earthly body can be more fully the church than this. It follows that, as it is responsible to God for its life in that place, so it must have freedom to discern and obey God's will for itself, with no dictation from outside. Although this view carries with it respect for the rights of the individual conscience, it is not spiritual individualism but an attempt to treat the visible and corporate character of the church as concretely as possible.

It has always been recognized that this principle did not involve ecclesiastical isolation. "The communion of the churches with each other" was a frequent 17th-century theme. But the precise way in which churches should be related to the association and councils through which they expressed their communion has often caused uneasy debate. In the 19th century, thinking about this relation was affected by the individualism of the age, while in the more centralized and mobile 20th century, with the widespread movement toward mergers and redeployment, the positive role of councils has been stressed. The authentic Congregational principle would appear to be that, whatever adaptations of organization may be necessary in changing circumstances, responsibility and the freedom to fulfill it must always be as specific and personal as possible.

The "gathered" church

The idea of the "gathered" church is integral to traditional Congregationalism. It is a recognition that the primary agent in church foundation is not human but God's Spirit. Arising in protest against the Anglican territorial conception of the church, according to which all residents of a particular neighbourhood should be counted as members of the local Anglican church, it insisted that it was the duty and privilege of the believer to discover who else in the vicinity was called by Christ and then to walk together with them in church order, which was thought of not primarily as a matter of organization but of common style of life. Where the state or prelacy tries to impose another principle, "the crown rights of the Redeemer" (Christ) in his church—a great phrase among Congregationalists—are impugned. How far the principle of the gathered church can be honestly applied in churches with large formal memberships is a problem modern Congregationalists have not solved, but great responsibilities remain with particular churches. All members are deemed to have equal rights and are expected to exercise them through membership of a church meeting that is empowered to deal with all matters pertaining to that particular church's life. Church meetings have not always been very vigorous and, especially in the United States, many of their powers have been delegated to officers or committees, but efforts have been made to restore them to their important place.

Ordination to the Congregational ministry has been through the ratification of the call of the individual by acceptance for training by the churches acting together, and then by the call from a particular church to act as its minister. This practice has been retained in most of the new united churches. The churches corporately set standards of training, which, particularly in the United States and Canada, is frequently conducted in interdenominational seminaries or universities.

Associations or unions of churches

Until new patterns were established by mergers, nearly all Congregational churches were linked together in association or unions on local, provincial, and national levels. In recent times these have appointed superintendent ministers or moderators, who exercise a general ministry to the churches over a large area; but it would be misleading to think of their role as equivalent to that of diocesan bishops, since they are not regarded as the sources of ecclesiastical order and have no formal authority over independent churches. It is a Congregational principle that the service of the Word and the sacraments, rather than one's place in a system of ecclesiastical administration, confers authority on a minister.

All offices in Congregational churches were open to women before it became widespread practice. The first woman was ordained in 1917. Churches are mainly financed by the contributions of members. There are substantial denominational funds for ensuring minimum stipends to finance missionary work and pensions, but even these depend heavily on contributions from the churches as well as on endowments.

Congregationalism in the modern world. Congregationalism has flourished most in settled communities of manageable size, in provincial cities, or in the substantial suburbs of larger cities. It has played a prominent part in the civic life of such places, especially in the 19th century, and it has proved itself a rich seedbed for educational and cultural aspirations. It has not itself always enjoyed the fruits of these aspirations because many of the children it has produced have moved on to spheres where the organized churches have found difficulty in keeping pace with them. Many prominent American and English politicians have been Congregationalists, among them Hubert Humphrey and Harold Wilson. John Milton and Robert Browning stand closest to the distinctive Congregationalist outlook among the numerous major artists of Congregationalist connection or upbringing.

Influence and decline

Congregationalism has clearly not succeeded in establishing itself as one of the major forms of churchmanship in the modern world. Congregational ideas and practices have, however, had a deep influence on many other churches. Congregationalism has also been a major factor in shaping the institutions and the general culture of the United States and, to a lesser degree, of Britain and the Commonwealth. Its expansion and vitality in England in the 19th century were closely linked with the rise of new middle-class groups, but with the increase of social mobility, the centralization of business organizations, and the decline of the continuity of family style of life from one generation to the next, its churches have suffered heavily in deterioration of numbers and direct social influence. The decline has not been as marked in the United States, where Congregational churches have shared in the general ecclesiastical prosperity, although even there they have not expanded at anything like the rate of most other large groups of churches. Most of the historic Congregational churches are now incorporated in reunited churches belonging to the Reformed family. Whether what is distinctive in Congregationalism can be effectively maintained under the pressure of modern urban mobility in more centrally organized churches is to be determined. (D.T.J.)

Friends

Friends (or Quakers) are a Christian group that arose in mid-17th-century England, dedicated to living in accordance with the "Inward Light," or direct inward apprehension of God, without creeds, clergy, or other ecclesiastical forms. As most powerfully expressed by George Fox (1624–91), Friends felt that their "experimental" discovery of God would lead to the purification of all of Christendom. It did not; but Friends founded one American colony and were dominant for a time in several others, and though their numbers are now comparatively small, they continue to make disproportionate contributions to science, industry, and especially to the Christian effort for social reform.

HISTORY

The rise of Quakerism. There were meetings of the kind later associated with the Quakers before there was a group by that name. Small groups of Seekers gathered during the Puritan Revolution against Charles I to wait upon the Lord because they despaired of spiritual help either from the established Anglican Church or the existing Puritan bodies—Presbyterians, Congregationalists, and Baptists—through which most of them had already passed. To these Seekers came a band of preachers, mostly from the north of England, proclaiming the powers of direct contact with God. Fox and James Nayler were perhaps the most eminent of these, but Edward Burrough, William Dewsbury, and Richard Farnworth also were active. The cradle of the movement was Swarthmore (Swarthmoor) Hall in northwestern Lancashire, which after 1652 became the centre of an evangelistic campaign by traveling ministers. Within

a decade perhaps 20,000 to 60,000 had been converted from all social classes except the aristocracy and totally unskilled labourers. Heaviest concentrations were in the north, Bristol, the counties around London, and London itself. Traveling Friends and Cromwellian soldiers brought Quakerism to the new English settlements in Ireland; Wales and especially Scotland were less affected.

Persecution of Quakers

The Puritan clergy, in England and New England, greeted the rise of Quakerism with the fury that an old left often reserves for a new. Friends' religious style was impulsive and nonideological; Quakers seemed to ignore the orthodox views of the Puritans and pervert their heterodox ones. Though most Friends had passed through varieties of Puritanism, they carried the emphasis on a direct relationship between the believer and God far beyond what Puritans deemed tolerable. The Restoration of Charles II in 1660 was only a change of persecutors for the Quakers, with their former tormentors now sharing some of their sufferings. From the Quaker Act of 1662 until the de facto toleration of James II in 1686 (de jure toleration came in the Toleration Act of 1689), Friends were hounded by penal laws for not swearing oaths, for not going to the services of the Church of England, for going to Quaker meetings, and for refusing tithes. Some 15,000 suffered under these laws, and almost 500 died in or shortly after being in prison, but they continued to grow in numbers until the turn of the century.

At the same time Quakers were converting and peopling America. In 1656 Quaker women preachers began work in Maryland and in the Massachusetts Bay Colony. The magistrates of Boston savagely persecuted the visitors and in 1659 and 1661 put four of them to death. Despite this, Quakerism took root in Massachusetts and flourished in Rhode Island, where Friends for a long time were in the majority. There were also many Friends in New Jersey, where English Quakers early secured a patent for settlement, and in North Carolina. Yearly meetings were established for New England (1661), Maryland (1672), Virginia (1673), Philadelphia (1681), New York (1695), and North Carolina (1698). The most famous Quaker colony was Pennsylvania, for which Charles II issued a charter to William Penn in 1681. Penn's "Holy Experiment" tested how far a state could be governed consistently with Friends' principles, especially pacifism and religious toleration. Toleration would allow colonists of other faiths to settle freely and perhaps become a majority; consistent pacifism would leave the colony without military defenses against enemies who might have been provoked by the other settlers. Penn, entangled in English affairs, spent little time in Pennsylvania and showed erratic judgment in selecting his non-Quaker deputies, who were almost always at odds with the Quaker-dominated legislature. Penn also went bankrupt through mismanagement; but the Quaker influence in Pennsylvania politics remained paramount until 1756, when legislators who were Friends could no longer find a saving formula allowing them to vote support for military operations against the French and Indians fighting settlers in western Pennsylvania. Voltaire's description of Penn's agreements with the Indians as the only treaties never sworn to and never violated was exaggerated; but Friends' relations with the Indians were more peaceful than those of other settlers.

William Penn's "Holy Experiment"

The age of quietism. The achievement of religious toleration in the 1690s coincided with a quietist phase in Quakerism that lasted until the 19th century. Quietism is endemic within Quakerism and emerges whenever trust in the Inward Light is stressed to the exclusion of everything else. It suits a time when little outward activity is demanded and when the peculiar traditions of a group seem particularly worth emphasizing. In the 18th century Friends had gained most of their political objectives. Their special language and dress, originally justified as a witness for honesty, simplicity, and equality, became password and uniform of a group now 75 to 90 percent composed of second- and third-generation Quakers. Strict enforcement of rules prohibiting marriage without parents' consent or to nonmembers led to the disownment, according to one estimate, of a third of the English Friends who got married in the latter half of the 18th century. More were disowned than converted, and since most members were the children of members, it is not surprising that Friends eventually came to recognize a category of "birthright" membership, which seemed to relax the expectation of conversion.

Seemingly self-absorbed in other ways, Friends in the age of quietism intensified their social concerns. English Friends were active in the campaign to end the slave trade, and American Friends, urged on by John Woolman and others, voluntarily emancipated all their own slaves between 1758 and 1800. Meetings, though slow to adopt this concern, pursued it thoroughly; in Rhode Island Stephen Hopkins, who was governor nine times, was disowned because he would not free his one slave.

The impact of evangelicalism. Cooperation with other Christians in the antislavery cause gradually led Friends out of their secluded religious life. They also came closer to other Protestants through the evangelical movement originally associated with John and Charles Wesley. Evangelical Friends were concerned with emphasizing the inerrancy and uniqueness of the Bible, the incarnation and atonement of Christ, and other characteristic Protestant doctrines which, although seldom denied outright by Friends, had tended to be subordinated to the quietistic emphasis on the Inward Light. In the early 19th century most leading English Friends were sympathetic to evangelical ideas, although they did not lose their unity with more traditional-minded Friends.

Evangelical Friends

In the United States unity proved more difficult. Friends had gone west—from Virginia and North Carolina because of difficulties over slavery, but also from Pennsylvania. As new yearly meetings were formed—Ohio (1812), Indiana (1821), Iowa (1863), Kansas (1872), Oregon (1893), California (1895), and Nebraska (1908), among others—ties with the London Yearly Meeting, the "mother" meeting, became weaker, and no American yearly meeting had a predominant position. Leaders of the Philadelphia Yearly Meeting, mostly rich merchants with strong ties to England, were sympathetic to evangelicalism; but many poorer country Friends left the meeting, no longer feeling a unity with the beliefs of the Philadelphia ministers and elders or with the way they exercised their authority. Elias Hicks (1748–1830), whose name was applied to these separatists, placed extreme emphasis on the Inward Light; he wrote that it might be a good thing if God withdrew the Bible, since he could inspire worshipers to write new scriptures that would probably be better than the originals. Since the various American yearly meetings corresponded with one another, the Hicksite separation spread to other yearly meetings that had to decide to which portion of the Philadelphia Yearly Meeting to write. A pastoral visit to the United States (1837–40) by the leading English evangelical Friend, Joseph John Gurney (one of the few systematic theologians ever produced in the Society of Friends), led to a further separation when the evangelical or "Gurneyite" New England Yearly Meeting disowned John Wilbur, an orthodox quietist Friend.

Schism is often a sign of religious vitality, and so it proved then. Whether Hicksite, Wilburite, or Gurneyite, all branches of Quakerism began to show vigour unknown in their days of torpid unity. With more vital preaching, many converts not devoted to the inherited peculiarities of Quaker tradition joined Friends; to them it seemed more important to assure a saving ministry than to preserve the traditional mode of worship. There thus grew up, especially in the Midwest and Far West, "pastoral meetings" in which a paid minister assumed the functions of delivering a sermon and exercising pastoral care of members. Such meetings often called themselves "Friends' Churches"; congregational singing was a part of the service, which might have only a few moments of silence, and baptismal and marriage ceremonies were introduced. In doctrine, worship, and polity they were not unlike Congregational churches, though they remained faithful to Friends' social testimonies. Even in England, where such innovations were not introduced, Friends, under the influence of the evangelical revival, discontinued disownment for irregular marriages and curtailed the powers of elders and overseers, which had been a profoundly conservative force.

Pastoral meetings

The 20th century. Friends in 1900 were divided into

three groups. Yearly meetings of evangelical, or "orthodox," Friends were in fellowship with one another and with the London and Dublin yearly meetings. In the United States these Gurneyite meetings in 1902 formed the Five Years' Meeting (now the Friends United Meeting). The "conservative" American yearly meetings, in fellowship with one another, maintained traditional Quaker customs and mode of worship. The Hicksite yearly meetings, which formed the Friends General Conference in 1902, remained the most open to modern thought. During the century these divisions have been much softened. Theological distinctions have receded in importance, and the habit of cooperation in such agencies as the American Friends Service Committee has drawn Friends together.

The 20th century has also seen the extension of Quakerism to Africa and continental Europe. Quakerism took root in the Netherlands in the 17th century but died out in the mid-19th, as did groups in Congéniès, France, and Bad Pyrmont, Germany. Quaker relief work in World War I and its aftermath produced new yearly meetings in Germany, The Netherlands, France, Sweden, and Switzerland, but numbers remain small.

The influence of Quakers. Quaker customs and the exclusion of Friends from many professions in England concentrated their secular achievements. Plainness meant that painting, music, and the theatre were proscribed. For a century trust in the Inward Light inhibited the foundation of colleges (though in the 19th century American Friends founded colleges like Earlham, Haverford, and Swarthmore; and individual Friends founded Bryn Mawr College, Cornell University, and Johns Hopkins University). Friends' schools emphasized science; the chemist John Dalton, the geneticist Francis Galton, the anthropologist E.B. Tylor, the astronomer Arthur Eddington, and Joseph Lister, discoverer of antisepsis, were Friends. In trade Friends were trusted and got customers; they trusted one another and extended credit; thus the many successful Quaker firms and banks, of which Barclay's and Lloyd's are the best known. Friends also pioneered in inventions, developing the puddling process for iron and the safety match and promoting the first English railroad line.

Disdaining formal education and a clerical intelligentsia, Friends, not surprisingly, often failed theologically (that is, could not solve some of the intellectual problems of their faith). But they would agree with the 19th-century Danish religious philosopher Søren Kierkegaard that "the highest of all is not to understand the highest but to act upon it."

TEACHINGS

The "public testimonies" of Friends from the very beginning included the plain speech and dress and refusal of tithes, oaths, and worldly courtesies. To these was added in a few years an explicit renunciation of participation in war; within the next century bankruptcy, marriage out of meeting, smuggling, and dealing in or owning slaves also became practices for which an unrepentant Friend would be disowned. These latter, especially those relating to slavery, became matters for discipline because a comparative minority of Friends persuaded the rest that they were inconsistent with Friends' principles.

Relief work

But not all social concerns were corporate in this sense or were enforced by sanctions. Friends' relief work, for example, has usually arisen from an individual response to suffering, often as the result of war. From the time of the American Revolution Quakers have been active in ministering to refugees and victims of famine—so much so that the entire Society of Friends is sometimes taken for a philanthropic organization; yet this work, recognized in 1947 by the award of the Nobel Peace Prize to the American Friends Service Committee and the (British) Friends Service Council, has mobilized many non-Quakers and thus exemplifies the interaction between the Quaker conscience and the wider world.

Yet the Society of Friends is grounded in the experience of God, out of which philanthropic activities may flow. There have always been Friends whose concerns went well beyond what meetings were willing to adopt. Most Friends were not abolitionists before the American Civil War; they probably did not approve of the Underground Railroad nor share the early feminist views of Lucretia Mott and Susan B. Anthony. (Most of the early suffragist leaders in America were Quakers.) The two American presidents of Quaker background were both Republicans: Herbert Hoover and Richard M. Nixon. Often the issue has been the relationship between private witness and public policy. Some Quaker pacifists make an absolute personal stand against war (for example, by refusing to register for selective service and thus forfeiting conscientious objectors' status); others are more willing to sacrifice absolute purity by working for an alleviation of international tensions even at the cost of less rigorous application of their principles.

WORSHIP AND ORGANIZATION

The Inward Light. Trust in the Inward Light is the distinctive theme of Quakerism. The Light should not be confused with conscience or reason; it is rather that of God in everyone, which allows human beings an immediate sense of God's presence and will for them. It thus informs conscience and redirects reason. The experience of hearkening to this inner Guide is mystical, but corporate and practical. Meetings to worship God and await his word (always open to anyone who wishes to come) are essential to Quaker faith and practice. Although the inward Seed can work in a solitary person, Friends do not meditate like monks, isolated in their cells. It is in the pregnant silence of the meeting of true waiters and worshipers that the Spirit speaks. Sometimes the meeting is too dull or worldly for any message to be heard, and sometimes there are altogether silent meetings. Although these are spiritually beneficial to the participants, ideally someone has reached a new understanding that demands to be proclaimed. He or she—for Friends have always given women equality in worship—speaks or prays and thus ministers to the meeting, which weighs this "testimony" by its own experiences of God. Friends historically have rejected a formal or salaried clergy as a "hireling ministry." If God can provide his own living testimony, the Bible and the learning necessary to read it can take a subordinate place, and creeds and outward sacraments can be dispensed with altogether. But despite their emphasis on silent waiting and their distrust of "creaturely" activity, Friends are no more habituated to passive than to solitary meditation. Often the "opening" of the Inward Light is a "concern" for the sufferings of others and a mandate laid upon the conscience to take action to alleviate that suffering. Such concerns typically are laid before a meeting and thoroughly considered; there must be a consensus for any corporate action. But slow as such action sometimes is, Friends have taken the lead in opposing slavery, brutality in prisons and insane asylums, oppression of women, militarism, and war.

Clergy and sacraments

Polity. Insofar as George Fox was the founder of Quakerism, he was so chiefly because of the system of meetings for church business that he established in the years immediately after 1667, which essentially stands today. Most important is the monthly meeting, which considers all applications for membership, in some localities manages Friends' properties, and acts on members' concerns. Generally, in the United States each congregation has a monthly meeting; in England and in some parts of the United States several meetings for worship combine in monthly meeting. Several monthly meetings form quarterly meetings, which are combined in yearly meetings.

This array is less hierarchical than it sounds. Any Friends can attend any meeting, which tries to remain open to the concerns or the service they can perform (much in the spirit of a meeting for worship). There is an official, the clerk, but the responsibility of the clerk is not to preside in a parliamentary manner but rather to feel for a "sense of the meeting," which draws together the thinking of the meeting to the point of action.

Though Friends have no ordination, they have always given a special place to Recorded Ministers (or Public Friends). Recorded Ministers are those whose testimony in local meetings has been officially recognized; they are free to "travel in the ministry" by visiting other meetings, should they be led to do so. Pastoral meetings maintain their Recorded Ministers, who also do much of the work

Recorded Ministers

of seeing to the relief of the poor, care of properties, and discipline of erring members. Ministers have usually had their own meetings together, and in most yearly meetings executive responsibility had been taken by a meeting like the Meeting for Sufferings in London (these are also called Representative meetings or committees or Permanent boards). London Meeting for Sufferings in the 17th century served as a political pressure group, lobbying Parliament for relief from persecution, coordinating legal strategy, and using the press for public appeals; in the 19th century they broadened their concerns to respond to sufferings everywhere.

Quakerism and world Christianity. The cause of schisms in the past—the tension between entire reliance on the Inward Light and the profession of orthodox Christian doctrines—remains unresolved. As it has divided Friends among themselves, it has also tended to separate them from other Christians. The London Yearly Meeting in 1940 declined to join the World Council of Churches out of uneasiness with its creedal basis, though some U.S. groups of Friends sent delegates to the first meeting of the council in 1948. Looked at in the context of Christendom as a whole, Friends offer a distinctive opportunity for spontaneity of worship, fellowship in mysticism, and proving mystical insight in labour for a suffering world. Many alienated from institutional Christianity have found this combination attractive; they may well feel more comfortable identifying themselves as Friends than as Protestants or even as Christians. This may make it more difficult for Quakerism to be subsumed into a reunited Christian church; but the faith of most Friends has always been that of Schweitzer in *The Quest of the Historical Jesus:* as "we do the work of Christ we shall come to know who he is." (R.T.V.)

Methodists

Methodism began in the 18th century as a religious society that wished to reform the Church of England from within; by force of circumstance it became separate from its parent body and took on the characteristics of an autonomous church.

HISTORY

Origins. John Wesley, the founder of Methodism, was born in 1703. After ordination in the Church of England, he was elected a Fellow of Lincoln College at Oxford in 1726. In the following year he left Oxford temporarily to act as curate to his father, the rector of Epworth. Back in Oxford, to which his younger brother Charles had now come, he found himself a member and soon the leader of a group of earnest students pledged to frequent attendance at Holy Communion, serious study of the Bible, and regular visitation of the filthy Oxford prisons. The members of this group received the sobriquet of Methodists.

In 1735 both John and Charles Wesley set out for Georgia to be pastors to the colonists and missionaries (it was hoped) to the Indians, at the invitation of the founder of the colony, James Edward Oglethorpe. They were unsuccessful in their pastoral work and did no missionary work. The brothers returned to England conscious of their lack of genuine Christian faith. They looked for help from members of the Church of the Brethren, who were staying in England for a while before joining Moravian settlements in the American colonies; among these Peter Böhler was especially important. On May 24, 1738, John Wesley's *Journal* narrates that he "felt" his "heart strangely warmed" and continues, "I felt I did trust in Christ, Christ alone, for salvation; and an assurance was given me that He had taken away *my* sins, even *mine,* and saved *me* from the law of sin and death." Charles Wesley had reported a similar experience a few days previously.

John Wesley's conversion experience

Some months later John Wesley was invited by his friend George Whitefield, also an Anglican clergyman who had undergone a "conversion experience," to come to the city of Bristol and help to preach to the colliers of Kingswood Chase, just outside the city, where human conditions were at their lowest. Wesley came and found himself, much against his will, preaching in the open air. This enterprise was the beginning of the Methodist Revival. Whitefield and Wesley at first worked together but later separated on doctrinal grounds: Whitefield believed in double predestination; Wesley regarded this as an erroneous doctrine and insisted that the love of God was universal.

Under the leadership at first of Whitefield and afterward of Wesley the movement rapidly gained ground among those who felt themselves neglected by the Church of England. Wesley differed from contemporary Anglicans not in doctrines but in emphases: he claimed to have reinstated the biblical doctrines that a man may be assured of his salvation and that, by the power of the Holy Spirit, he is capable of attaining perfect love for God and his fellows in this life. Wesley's helpers included only a few ordained clergymen and notably his brother Charles, who wrote more than 6,000 hymns to express the message of the Revival. In spite of Wesley's wish that the Methodist Society would never leave the Church of England, relations with Anglicans were often strained.

In 1784, when there was a shortage of ordained ministers in America after the Revolution, the Bishop of London refused to ordain a Methodist for the United States. Wesley, acting in an emergency and on biblical principles that allow (as he thought) a presbyter to ordain, ordained Thomas Coke as superintendent and two others as presbyters. In the same year, by a Deed of Declaration, he appointed a Conference of 100 men to govern the Society of Methodists after his death.

Break with the Church of England

The definite break with the Church of England came in 1795, four years after Wesley's death. After this, English Methodism, with vigorous outposts in Ireland, Scotland, and Wales, rapidly developed as a church. But in order not to perpetuate the split from the Church of England, it was reluctant at first to ordain with the laying on of hands. Its system centred in the annual Conference (at first of ministers only, later thrown open to laypeople), which controlled all its affairs. The country was divided into districts and the districts into circuits, or groups of congregations. The ministers were appointed to the circuits, and each circuit was led by a superintendent, though much power remained in the hands of the local trustees.

This tightly knit system enabled the Wesleyan Methodist Church to grow rapidly throughout the 19th century, at the end of which it counted 450,000 members. The growth was largest in the expanding industrial areas. There their faith enabled Methodist workers, men and women, to endure economic hardship, while at the same time working for the alleviation of poverty. Because their faith encouraged them to live simply, their economic status tended to rise, with the unintended result that Wesleyan Methodism became a middle-class church that was not immune to the excessive stress on the individual in material and spiritual matters that marked the Victorian age.

At the same time the autocratic habits of some ministers in authority, notably Jabez Bunting, an outstanding but sometimes ruthless leader, alienated many of the more ardent and democratic spirits, and there was a series of schisms. The Methodist New Connexion broke off in 1797, the Primitive Methodists in 1811, the Bible Christians in 1815, and the United Methodist Free Churches in 1857.

The smaller Methodist groups were in closer contact with the working classes than the Wesleyans and provided the leadership in early trade unionism to an extent disproportionate to their size. The Wesleyans were at first conservative in politics but in the second half of the 19th century identified themselves more and more with the liberalism of William Gladstone.

A movement to reunite the Methodist groups began about the turn of the century and reached success in two stages. In 1907 the Methodist New Connexion, the Bible Christians, and the United Methodist Free Churches joined to form the United Methodist Church; and in 1932 the Wesleyan Methodist Church, the Primitive Methodist Church, and the United Methodist Church came together to form the Methodist Church.

The Methodist Church has shared with the other English churches in the numerical decline that began about 1910. This decline, together with changes in modern life and thought, roused it out of its Victorian complacency and

filled it with a desire to express Wesley's original ideals in a contemporary form. It continued to plan new attempts at evangelism. Its concern for education, shown especially in the development of Kingswood School (Wesley's foundation) and other boarding schools, as well as in the training of Christian teachers at Westminster and Southlands colleges, has not abated. Its strong social interest has expanded from preoccupation with total abstinence to a wide range of national and international issues, especially those connected with race, poverty, and peace.

Ecumenical efforts

The Methodist Church involved itself in the ecumenical movement when it began in 1910. Thereafter the church shared in all negotiations for church union. Relations with the Church of England improved so much by the 1960s that a plan for the reunion of the two churches (in two stages) was approved in principle by both in 1965. The final form of the plan was approved by the Methodist Church with a very large majority in 1969, but the Church of England did not muster a large enough majority to bring the plan into effect. The same thing happened in 1972.

Proposals for a "Covenant for Visible Unity," to include the United Reformed Church and the Moravian Church as well as the Methodists and the Anglicans, were put to the churches in 1982; once again the Anglican vote fell short, while the other churches were in favour. Most Methodists were grievously disappointed, but many threw themselves into projects in their own neighbourhood intended to realize locally the unity that was not possible nationally. In these projects Anglicans, United Reformed Church people, and sometimes Baptists and Roman Catholics, are taking part.

As a founder member of the British Council of Churches and the World Council of Churches, the Methodist Church has shared fully in the activities of these councils and provided many leaders. Official discussion with Roman Catholics on national and world levels has revealed a surprising degree of agreement while promoting tolerance and understanding of previously contentious issues.

Ordination of women

The first woman was ordained to "The Ministry of Word and Sacraments" in 1974. This was the climax of many years of discussion and controversy. It indicated a growing appreciation of the place of women in the life of the church. The theological objections had been carefully considered and rejected before the final step was taken.

Methodism in America. Methodism was taken to America by immigrants from Ireland who had been converted by John Wesley. Wesley also sent preachers, and by far the most successful of these was Francis Asbury, a blacksmith, who arrived in 1771 and covered vast distances. He adapted Wesley's principles to the needs both of the settled communities and of the frontier. Wesley took the side of the English government at the time of the Revolution, but Asbury aligned himself with the new American republic. Wesley sent the men whom he had ordained as presbyters, with Thomas Coke as superintendent, to help Asbury. The Methodist Episcopal Church was constituted in 1784 and regarded itself as autonomous. Asbury and Coke allowed themselves to be called bishops.

The next 50 years saw a remarkable advance led by the circuit riders who preached to the frontiersmen in simple terms. The slavery issue split the Methodist Church into two: the Methodist Episcopal Church and the Methodist Episcopal Church, South (organized in 1845). After the Civil War both churches increased rapidly and became gradually assimilated to the general pattern of American Protestantism. When it was clear that the old issues no longer divided them, they began to move together. But it was not until 1939 that they came together to form The Methodist Church. The Methodist Protestant Church, a smaller group, joined in the same union.

The Central Jurisdiction

The church in the South lost its black members before and during the Civil War. At the time of the union the Central Jurisdiction was formed for all the black members wherever they lived; it existed alongside the other jurisdictions that were determined by geography. The Central Jurisdiction was abolished in 1968; and black Methodists are now integrated in the church.

The originally German-speaking Evangelical United Brethren Church, itself a union of the Church of the United Brethren in Christ and the Evangelical Church, was united with The Methodist Church in 1968 to form the United Methodist Church. Women were given limited clergy rights in 1924 and were accepted for full ordination in 1956.

Methodism in Canada. Methodism was extended to Canada by preachers from the United States and later reinforced by British Methodists. In 1874 The Methodist Church of Canada became autonomous; it went on to negotiate a union with other Canadian nonepiscopal churches to form The United Church of Canada in 1925.

TEACHINGS

Methodism is marked by an acceptance of the doctrines of historic Christianity; by an emphasis on those doctrines that indicate the power of the Holy Spirit to confirm the faith of the believer and transform his personal life; by insistence that the heart of religion lies in personal relationship with God; by simplicity of worship; by the partnership of ordained ministers and laity in the worship and administration of the church; by a concern for the underprivileged and the betterment of social conditions; and (at least in its British form) by the formation of small groups for mutual encouragement and edification.

All Methodist churches profess allegiance to the Scriptures as the supreme guide to faith and practice. They welcome the findings of modern biblical scholarship (except for the fundamentalist groups to be found within them). They accept the historic creeds and hold themselves to be in the tradition of the Protestant Reformation. Arguments about the virgin birth and the physical resurrection of Jesus do not greatly concern Methodists; they allow for differences of conviction on these points within the historic faith. They emphasize the teaching about Christian perfection, interpreted as "perfect love," which is associated with John Wesley, who held that every Christian should aspire to this by the help of the Holy Spirit.

Methodist churches assert the value of infant baptism and the need to receive regularly the sacrament of the Lord's Supper, in which they believe Christ to be truly present, though they have no precise definition of the manner of his presence. They believe themselves to be integral parts of the one, holy, catholic, and apostolic church, and their ministers to be true ministers of Word and sacrament in the church of God.

WORSHIP AND ORGANIZATION

Patterns of service. Methodist worship everywhere is partly liturgical, partly spontaneous. John Wesley regularly used the Anglican *Book of Common Prayer* and adapted it for use in the United States. He also conducted services that included extemporaneous prayer. His custom was continued in Britain. In the 20th century Anglican Morning Prayer gradually dropped out of Methodism, but Anglican Holy Communion continued until the Liturgical Movement impelled all churches, Roman Catholic and Protestant alike, to revise their liturgies. The *Methodist Service Book* (1975), written in a modern language, offers much opportunity for congregational participation. The Sunday Service, or Holy Communion, restores the traditional fourfold pattern—the offering of bread and wine, the thanksgiving, the breaking of the bread, and the sharing of the elements. Non-liturgical services, which constitute the majority, tend to be stereotyped although they claim to be spontaneous. Far more services are conducted by lay preachers than by ordained ministers.

Lay participation

In American Methodism services are rarely conducted by laypeople. The Liturgical Movement affected the *Book of Worship* (1965), the *Ordinal* (1980), and the *United Methodist Hymnal*, subtitled *The Book of United Methodist Worship* (1988), which is arranged to eliminate all traces of sexism.

Hymns are important in all branches of Methodism. Those of Charles Wesley are still dominant in British Methodism, but they are mingled with many contemporary hymns as well as hymns from other traditions. In *Hymns and Psalms* (1983) certain changes were made in order to eliminate sexist overtones. American books contain fewer hymns by Wesley.

Polity. In the churches of the British tradition the annual Conference is the supreme authority for doctrine, order, and practice. All ministers have parity of status, but special functions are exercised by the president and secretary of the Conference, the chairmen of districts, the secretaries of divisions, and superintendents. District affairs are regulated by Synods, Circuits by Circuit Meetings, local Societies by Church Councils. The American tradition is episcopal; the bishops are elected by the Jurisdictional Conferences, which, like the General Conference, meet every four years. Each diocese has an annual Conference and is divided into District Conferences, each with its superintendent. The dioceses are combined into five Jurisdictions that cover the nation. The circuit system is not developed. A minister is ordained first deacon, then elder.

In the United States the African Methodist Episcopal Zion Church and the African Methodist Episcopal Church antedate the explosion of the slavery question; the Colored (now "Christian") Methodist Episcopal Church was founded as a result of it. All three are exclusively black but follow the doctrine and organization of the United Methodist Church.

There are minority Methodist churches in most European countries. Those in Italy and Portugal are of English origin, that in Germany is of mixed English and American origin; the rest are all derived from American Methodism, though they exhibit many similarities in spirituality to the English type.

Missions. Thomas Coke began the missionary activities of British Methodism by his eloquence and ceaseless travels. The first area where missions took root was the West Indies; then came Sierra Leone and southern Africa. The Gold Coast, French West Africa, and Nigeria received missionaries not much later, though the climate in many parts of Africa took a toll of missionary lives.

Methodism in India, China, and Australia

In India converts were very few until about 1880, when a mass movement swept many thousand low-caste Indians in the south into the Methodist and other churches. In China missionary work had a checkered career, though there were mass movements there also. The last missionary left China in 1949. In Australia the Methodist Church began in 1815 and, like the Methodist Church in South Africa, became independent before the end of the 19th century. The movement toward autonomy became a flood after World War II; only a few small churches remain under the control of the Overseas Division of the church. Most of the autonomous churches negotiate for united churches in their countries; and the Church of South India, including Anglicans, Methodists, Congregationalists, and Presbyterians, has been in existence since 1947, and the Church of North India since 1970.

American Methodists have been equally enthusiastic for missionary activity, and their greater resources have carried them over still larger areas of the Earth's surface. North India, Mexico, and most of the other countries of Latin America, Cuba, Korea, Japan, Taiwan, and many parts of Africa possess Methodist churches of the American tradition. The movement toward autonomy took place more slowly in these areas than in the British sphere of influence. The General Conference of the United Methodist Church makes plans for combining fraternal relations among them with their newly found independence.

World Methodism. The two Methodist traditions diverged considerably for most of the 19th century but toward its end began to converge again. Ecumenical Methodist (since 1951 World Methodist) Conferences have been held regularly since 1888. The World Methodist Council represents some 80 churches.

Methodism in the world church. In Britain the Methodist Church is the largest of the Free Churches; it is not a nonconformist church but stands between nonconformity and Anglicanism, with affinities to both. In the United States it is closely aligned with the other non-Anglican Protestant denominations. (R.E.D.)

Disciples of Christ

There are three major bodies of the Disciples of Christ, all of which stem from a common source.

The Churches of Christ emphasize rigorous adherence to the New Testament as the model for Christian faith, practice, and fellowship. They reject ecclesiastical institutions other than the congregation, practice a dynamic evangelism based on a literal view of the Bible, and remain aloof from interdenominational activities.

The Christian Church (Disciples of Christ) affirms a free and voluntary covenantal relationship binding members, congregations, regions, and general units in one ecclesiastical body committed to a mission of witness and service. Recognizing its status as a denomination, it acknowledges the right of "dissent in love" and engages fully in the ecumenical venture.

The congregations loosely related in the Undenominational Fellowship of Christian Churches and Churches of Christ refused to enter such a "Christian Church." They earlier had refused to follow the Churches of Christ in rejecting musical instruments in worship and missionary organizations as a matter of biblical principle; they later repudiated the openness of their fellow Disciples toward biblical criticism, theological liberalism, ecumenical involvement through "official" channels, and development of denominational institutions.

In a larger sense Disciples of Christ includes sister churches in Australia and New Zealand, known locally as Churches of Christ, with origins largely independent of the United States. It also denotes churches in other lands resulting from the missionary efforts of all these bodies; most of these younger churches, as well as Churches of Christ in Great Britain, have entered united churches.

Originally Disciples blended the independence and pragmatism of the American frontier with an uncomplicated biblical faith that demanded restoration of the "ancient order" in the church. They repudiated "human creeds" and traditions as requirements for Christian fellowship, understood baptism as the immersion of believers only, and recognized no churchly authority beyond the congregation. This simple formula's typical "sectarianism" was combined with a strong catholic impulse: a plea for the union of all Christians, the regular celebration of the Lord's Supper in weekly worship, and the use of inclusive biblical names.

HISTORY

Origins. The movement emerged on the American frontier through various efforts to cut through the complexities of sectarian dogma and find a basis for Christian unity. Out of the Great Western Revival (1801) in Kentucky arose the short-lived Springfield Presbytery, which dissolved in 1804 so that its members might "go free" simply as Christians. Their leader, Barton W. Stone, championed revivalism, a simple biblical and non-creedal faith, and Christian union. In the upper Ohio Valley Presbyterian Thomas Campbell organized the Christian Association of Washington (Pennsylvania) in 1809 to plead for the "unity, peace, and purity" of the church. Soon its members formed the Brush Run Church and ordained his son Alexander, under whose leadership they accepted immersion of believers as the only scriptural form of baptism and entered the Redstone Baptist Association. Alexander Campbell rapidly gained influence as a reformer, winning fame as preacher, debater, editor (*Christian Baptist*), and champion of the new popular democracy. His colleague Walter Scott developed a reasonable, scriptural "plan of salvation." Its "positive," or objective, steps into the church (faith, repentance, baptism, remission of sins, gift of the Holy Spirit) attracted thousands who longed for religious security but had not experienced the emotional crisis and subjective assurance that characterized the prevailing revivalism.

Stone and the Campbells

By 1830 the regular Baptists and the reformers parted company, the latter terming themselves Disciples. Two years later Stone and many of his followers joined with them, though continuing to use the name Christians.

Alexander Campbell from 1830 on turned to constructive church craft. He founded *The Millennial Harbinger*, established Bethany College, then in Virginia (1840), and agitated unsuccessfully for a general church organization based on congregational representation. The first general

convention met at Cincinnati, Ohio, in 1849 and launched the American Christian Missionary Society as a "society of individuals" and not an ecclesiastical body. Similar cooperative organizations emerged in various states to support evangelists and to establish new churches. The Christian Woman's Board of Missions (1874) and the Foreign Christian Missionary Society (1875) initiated successful programs overseas, and other boards were soon founded to promote building loans for new churches, care for aged ministers, homes for orphans and the aged, temperance, and other causes. The Centennial Convention at Pittsburgh in 1909 claimed an attendance of 30,000; they had come to celebrate a century of triumph for the New Reformation, or Restoration Movement.

Controversy and separation. Meanwhile, schism had begun to sunder the ranks, yet without shaking the confidence of the Disciples in their plea for union. They had held together during the controversy over slavery and through the Civil War, when major American denominations had divided. In the succeeding era of bitterness, however, the Disciples also suffered schism. New developments in response to growing urbanization and sophistication brought two sharply divergent responses. The conservatives regarded such developments as unauthorized "innovations," while the progressives (pejoratively termed digressives) looked on them as permissible "expedients."

Plea for union

Discord first arose over the "society principle" involving general missionary work. Alexander Campbell's biblical view of the church had kept pushing him toward a general church organization, but he could never find a convincing biblical text to support his proposals. Frontier independence and pragmatic popular biblicism prevailed. The "society principle" seemed to its advocates a legitimate solution: entertaining no ecclesiastical pretensions as a secular corporation, the missionary society provided a means by which individual Disciples could work in voluntary cooperation. But the opponents saw in it a repudiation of the Bible as the determining rule of practice.

The introduction of musical instruments (reed organs) into Christian worship led to many local disputes. Other innovations added occasion for controversy—the infringement of the "one-man pastoral system" on the local ministry of elders, introduction of selected choirs, use of the title Reverend, and lesser issues.

In 1889 several rural churches in Illinois issued the Sand Creek Declaration, withdrawing fellowship from those practicing "innovations and corruptions." In 1904 a separate "preacher list" issued unofficially by some conservative leaders certified their preachers for discounts on railway tickets. The Federal Religious Census of 1906 acknowledged the separation between Churches of Christ and Disciples of Christ (who commonly used the name Christian Churches) even though many congregations did not decide which they were for some years.

The Sand Creek Declaration

The crucial issue centred on the manner of understanding biblical authority. Both conservatives and progressives accepted the New Testament as the only rule for the church. The conservatives, heavily concentrated in the South, applied a strict construction to Scripture; this required a specific New Testament precept to authorize any practice. The progressives tended toward a broader construction, accepting as expedient such measures as they found harmonious with Scripture or not in conflict with it.

Disciples in the 20th century. Disciples had experienced their most rapid growth in rural America. Their leaders responded to the passing of the frontier, the growth of cities, and the emergence of urban expectations. Whereas the Churches of Christ had opted for the practices established in the rural past, regarding them as biblical, the Disciples of Christ (progressives) were able to find some flexibility in the biblical rule. Nevertheless, rural and small-town Christian Churches predominated in numbers and membership even past mid-century, and the newer social and cultural influences did not affect all of them simultaneously.

Development of a professional ministry

Urban churches demanded full-time leadership, and Disciples gradually developed a professional ministry. In the first half of the century they worked hard to establish collegiate education as standard for ministers. As late as 1930 only 11 percent had graduate education, and the rapid growth of theological seminaries did not come until after World War II. The expanding corps of educated leadership reworked the inherited formulas, introducing both ideas and practices that troubled the more traditional.

The cooperative organizations underwent notable changes. In 1917 the old general convention, a week-long series of annual meetings of the various societies, gave way to the International Convention (U.S. and Canada), to which all cooperative agencies were expected to submit reports for review and advice.

Meanwhile, a number of the agencies had combined in 1920 to form the United Christian Missionary Society. Ten years later most state and national agencies entered Unified Promotion, a cooperative program of fund raising, with voluntarily accepted restraints on independent campaigns, and with distribution on the basis of agreed allocations. Thus they gradually evolved, in effect, one general budget. From the start the United Society drew intense criticism for ecclesiasticism and theological liberalism. Opposition centred on reports of "Open Membership" in the China mission. (Open membership, increasingly practiced in the United States, meant reception of unimmersed Christians from other denominations.)

In 1927 traditional forces established the North American Christian Convention. Many churches gave their support to "independent" missionaries in large numbers, as well as to "independent" Bible colleges, youth camps, district meetings, Bible school curricula, various publications, and a directory of ministers—all of them explicitly denying official status—more or less parallel to the "cooperative" agencies. The power struggle focused on the placement of ministers and resulted, on the cooperative side, in enhancing the leadership of the state secretaries and creating the pressure for delegate conventions in the states.

The cooperative conventions (state and international) also became instruments of ecumenical participation, electing representatives to the old Federal Council of Churches (and to the succeeding National Council and the World Council of Churches) as well as to the state councils. Thus, for the sake of their original catholic commitment, the "cooperatives" accepted status as a denomination, a compromise that the independents rejected.

Ecumenical participation

A growing sense of moral obligation toward the common cause led in 1950 to the formation of the Council of Agencies, which included all organizations reporting to the International Convention. Legally independent, they sought by consultation to avoid overlapping and to develop a common mind. From the council came a proposal for a Commission on Restructure, appointed by the convention in 1960. In 1967 the convention approved the commission's Provisional Design for the Christian Church (Disciples of Christ), ratified in the ensuing year by all 40 area conventions and 15 national agencies.

Beginning in 1968 the International Convention was replaced by the General Assembly, the state conventions by regional assemblies, and the old cooperative agencies by "general units" of the church. State secretaries became regional ministers, and the chief executive officer was named general minister and president. In 1977 the General Assembly removed the word Provisional from the title of the Design. Congregations retained full legal independence, but the system provided for corporate unity through decisions by representatives from congregations and regions.

Fear of infringement on congregational freedom and theological opposition to the doctrine of the church underlying restructure led to active opposition. Many independent congregations formally requested withdrawal of their names from the Yearbook of Christian Churches (Disciples of Christ), and a campaign led some cooperative churches to follow suit. From 1967 to 1969 the number of congregations listed dropped from 8,046 to 5,278.

Schism

Meanwhile, a self-appointed Chaplaincy Endorsement Commission for the Undenominational Fellowship of Christian Churches and Churches of Christ asked recognition by the U.S. government to represent those congregations that had elected "to continue as free, independent, and completely autonomous local churches" apart from the restructured Christian Church.

The World Convention of Churches of Christ since 1930

has sponsored mass meetings for fellowship and inspiration at five-year intervals. It attracts both cooperative and independent Disciples from America and from many nations but few from American Churches of Christ.

Churches of Christ in the 20th century. In 1906 the membership and leadership of the Churches of Christ were located mainly in the South, with heaviest concentrations in Tennessee and Texas. The reported membership of 159,658 apparently did not include all who accepted the general position of the Churches of Christ. In the ensuing half-century they grew into the largest of the three Disciples groups. The migration from the rural South to urban centres brought impressive membership gains in the North and the West—aided by a vigorous evangelism making intensive use of radio. Missionaries established churches in Asia, Africa, Latin America, and Europe, winning converts especially from Roman Catholicism. Many churches now forward their missionary funds to an agent for disbursal, while making certain that the actual appointment of missionaries remains the prerogative of congregational elders.

Missionaries

The churches' doctrine permits individual initiative in certain types of religious (not ecclesiastical) enterprises. A vigorous journalism has flourished for more than a century, the most influential papers being the *Gospel Advocate* (Nashville, Tenn.) and *Firm Foundation* (Austin, Texas). Benevolent homes provide care for children and the aged. A number of churches conduct Christian day schools, while private colleges offer Christian higher education and receive support from churches. A graduate school of religion at Harding College in Memphis, Tenn., offers a three-year Master of Theology degree.

Variations of conviction about specific practices (whether a single, "common" cup or many cups are to be used in communion) and doctrines (especially millennial ones about the perfect age of Christ's reign on earth) have produced sharp controversies and withdrawal of fellowship.

In the 1960s some leaders in the Churches of Christ set up informal forums or conferences on unity with members of the Christian Churches, both cooperative and independent. Although having no official status, these meetings provided opportunity for a limited but continuing ecumenical dialogue. Their doctrinal stance, in repudiation of ecclesiastical organization, prevents members of both the Churches of Christ and the Undenominational Fellowship of Christian Churches and Churches of Christ from official participation in general ecumenical gatherings.

TEACHINGS

Alexander Campbell summarized his theology in *The Christian System* (1835), the most influential book in shaping Disciples thought. In it he outlined a commonsense biblical doctrine against the complex theories of the schools and the sects. He emphasized reliance on the Bible and insisted on going to the sources. Relying on John Locke, "The Christian philosopher," Campbell perceived the grounds for Christian faith in historical events and objective evidence (recorded in Scripture) rather than in mysticism or subjective religious "experience." He therefore repudiated the Calvinist (and revivalist) concept of miraculous conversion and the similar concept of miraculous call to the ministry. Debates on these issues, as well as on the damnation of unbaptized infants, which Disciples denied, led them to think of themselves as anti-Calvinist.

The general framework of their thought nevertheless followed Reformed (Calvinist) lines, modified by the influence of British Independents (the originally Scottish Glasites—or Sandemanians—in practice a strictly New Testament sect, and the Congregationalists). Disciples shared the orthodox Protestant emphasis on the authority of Scripture. Their classic biblical position differs from that of other Protestants in being a product of the early 19th rather than of the 16th or 17th century.

Early Disciples understood their uniqueness to lie in the rigour, precision, and simplicity with which they set forth the biblical basis for the unity of all Christians. Campbell distinguished sharply between Old and New Covenants (Testaments), limiting to the latter any authority for "the original faith and order" of the church. Only explicit apostolic teaching or precedent belonged in the realm of faith, of the essential; all else, however logical or helpful, fell in the area of opinion and consequently of Christian liberty. Thus they rejected creeds as tests of fellowship; they believed such tests usurped the sole authority of the New Testament and set forth demands not found there. The popular Disciples' bias against theology as a divisive preoccupation with human opinions—as well as Alexander Campbell's early protest against ecclesiastical institutions as unwarranted by Scripture and threatening to freedom—also was inferred from the New Testament.

Standards of apostolic teaching

WORSHIP AND ORGANIZATION

After fruitless attempts to derive a stated order of worship from the New Testament, Disciples settled into an informal but relatively stable pattern composed of hymns, extemporaneous prayers, Scripture, sermon, and breaking of bread. Except for its omission of the Decalogue, the public confession of sin, and the creed, it resembled classic Reformed (or Presbyterian) worship, especially in its austerity of spirit. In the second half of the 19th century it took over more of the mood of popular revivalism, which still prevails among Churches of Christ and the independent Christian Churches.

Because many churches in the 19th century had the services of a preacher only occasionally but regularly observed the Lord's Supper (communion) after the Bible School (Sunday School) hour, the breaking of bread came to precede the sermon, which was simply added on when a preacher was present. At the table two local elders presided, one offering a prayer of thanksgiving for the bread and the other for the cup. The minister now commonly presides, but the elders ordinarily offer the prayers.

Christian Worship: A Service Book (1953), a semiofficial manual for voluntary use, exerted wide influence in restoring and stabilizing the typical pattern, with an emphasis on use of scriptural sentences throughout. The influence of the Liturgical Movement brought greater use of responsive readings, litanies, and affirmations of faith, as well as closer accommodation to the historical pattern of the liturgy—all demonstrated in the 1987 "resource for Christian worship," *Thankful Praise.*

Campbell regarded immersion and "the breaking of bread" (*i.e.,* baptism and communion) as ordinances of Christ. While the insistence on believer's baptism alone separated Disciples from the "paedobaptists" (those advocating baptism of children), weekly communion served as a universal element in their worship and tempered their rationalist bent. Despite their memorialist doctrine (that communion is a commemoration of Christ's Last Supper involving no miracle of transubstantiation), they understood the service as present communion with their Lord.

Campbell saw the biblically authorized ministry as that of elders and deacons, ordained by the congregations, and of evangelists, who served the church at large. Since the 1950s congregations have commonly elected women to diaconate and eldership, and Disciples have long ordained women as ministers. By the 1980s fully one-third of their seminarians were women.

The Design recognizes "the order of the ministry," consisting of ordained ministers and licensed ministers. Since restructure, the General Assembly has established policies and criteria for the order of ministry, which are interpreted and applied by regional commissions.

Internal differences. The divisions in the movement expressed varying attitudes toward Scripture as the norm of faith and practice: Churches of Christ construing it strictly, Disciples more loosely. Many who introduced organs in worship held the same view of biblical authority as those who refused to do so; their interpretation simply led to a different conclusion about the use of musical instruments in apostolic times. They provided the constituency for the "independent" Christian Churches, whereas Disciples tended to find more and more flexibility in the principle of expediency.

Controversies in biblical interpretation

Beginning in the early 19th century as a revolution occurred in the scholarly understanding of the biblical documents and the nature of their authority, the Churches of Christ generally held steadfastly to older views of Scripture, as the independents also tended to do, while Disciples

accepted the approach of critical scholarship. At the beginning of the 20th century, the most influential Disciples scholar was J.W. McGarvey, a champion of the traditional doctrines and view of the Bible and an opponent of the musical instrument in worship. Early in the century Herbert L. Willett, E.S. Ames, and C.C. Morrison led in a liberal reformulation of the plea, emphasizing a pragmatic and reasonable approach to faith, the repudiation of creeds, an openness to the scientific world view, and a commitment to Christian unity. Neoorthodoxy held less appeal for most Disciples, but William Robinson gained attention for his emphasis on biblical doctrine.

Recent trends. With the rapid growth of seminaries and religion faculties and extensive ecumenical involvement, Disciples enjoyed a theological renaissance in the 1950s. During the heyday of biblical theology some of them worked out a contemporary formulation of the tradition within the ecumenical context. A Panel of Scholars, appointed by two of the national agencies, published three volumes of papers in 1963 reflecting the new mood.

The institutional developments leading to restructuring were accompanied by a reformulation of the doctrine of the church. The founders had spoken of the Church of Christ as a local congregation; they recognized no other organization as a church. The new generation of Disciples could no longer deny the churchly character of the institutions that had been developed. The Design speaks of three manifestations of the Christian Church—congregational, regional, general (United States and Canada). The name that they adopted—the Christian Church (Disciples of Christ)—they found to have been dictated by their history. They saw that church manifesting itself organizationally "within the universal body of Christ" and committed to "responsible ecumenical relationships." In 1962 Disciples entered the Consultation on Church Union and in 1985 an ecumenical partnership with the United Church of Christ. They gave a cordial reception to the World Council of Churches document *Baptism, Eucharist and Ministry* (1982), even while recognizing problems posed by their eldership for the emerging consensus.

Ecumenical efforts

In the immediate decades after restructure no major theological controversy arose. Resurgent Fundamentalism and Evangelicalism on the larger scene had little impact.

Social issues. On social questions Disciples have held positions characteristic of the American denominations of English background. With regard to the issue of slavery Campbell prevented schism by admitting that Scripture and civil law permitted slavery, though, as a matter of personal opinion, he favoured emancipation. During the Civil War a number of leading Disciples, especially in the Border States, espoused pacifism on biblical grounds.

Disciples representatives to the National Council of Churches and the World Council of Churches have supported those organizations' general stand on social issues. In the second half of the 20th century, though a moderate conservatism obtained at the grassroots, ministers, seminaries, general units, and General Assembly placed social issues high on their agenda, with vocal sympathy for liberation theology. In 1969 the General Assembly called for a 20 percent presence of ethnic minorities on church policy-making bodies, even though the combined number of Native American, black, Hispanic, and Asian-American Disciples fell well below that figure. (R.E.O.)

Unitarians and Universalists

Unitarians and Universalists, who have merged in the United States, are groups of religious liberals. In previous centuries they appealed for their views to Scripture interpreted by reason, but most contemporary Unitarians and Universalists base their religious beliefs on reason and experience.

Unitarianism as an organized religious movement emerged during the Reformation period in Poland, Transylvania, and England, and later in North America from the original New England Puritan churches. In each country Unitarian leaders sought to achieve a reformation that was completely in accordance with the Hebrew Scriptures and the New Testament; in particular, they found no warrant for the doctrine of the Trinity accepted by other Christian churches.

Universalism as a religious movement developed from the influences of radical Pietism in the 18th century and dissent in the Baptist and Congregational churches from predestinarian views that only a small number, the elect, will be saved. Universalists argued that Scripture does not teach eternal torment in hell and with Origen, the 3rd century Alexandrian theologian, they affirmed a universal restoration of all to God.

HISTORY

Servetus and Socinus. In *De Trinitatis erroribus* (1531; "On the Errors of the Trinity") and *Christianismi restitutio* (1553; "The Restitution of Christianity") the Spanish physician and theologian Michael Servetus provided important stimulus for the emergence of Unitarianism. Servetus' execution for heresy in 1553 led Sebastian Castellio, a liberal humanist, to advocate religious toleration in *De haereticis . . .* (1554; Concerning Heretics") and caused some Italian religious exiles, who were then in Switzerland, to move to Poland.

One of the most important of these Italian exiles was Faustus Socinus (1539–1604). His acquisition in 1562 of the papers of his uncle Laelius Socinus (1525–62), a theologian suspected of heterodox views, led him to adopt some of Laelius' proposals for the reformation of Christian doctrines and to become an anti-Trinitarian theologian. Laelius' commentary on the prologue to the Gospel According to John presented Christ as the revealer of God's new creation and denied Christ's preexistence. Faustus' own *Explicatio primae partis primi capitis Ioannis* (first edition published in Transylvania in 1567–68; "Explanation of the First Part of the First Chapter of John's Gospel") and his manuscripts of 1578, *De Jesu Christo Servatore* (first published 1594; "On Jesus Christ, the Saviour") and *De statu primi hominis ante lapsum* (1578; "On the State of the First Man Before the Fall"), were of subsequent influence, the first, particularly, in Transylvania and all three in Poland.

Laelius Socinus

Unitarianism in Poland. Unitarianism appeared in Poland in incipient form in 1555 when Peter Gonesius, a Polish student, proclaimed views derived from Servetus at a Polish Reformed Church synod. Controversies that ensued with tritheists, ditheists, and those who affirmed the unity of God resulted in a schism in 1565 and the formation of the Minor Reformed Church of Poland (Polish Brethren). Gregory Paul, Marcin Czechowic, and Georg Schomann soon emerged as leaders of the new church. They were encouraged by Georgius Blandrata (1515–88), an Italian physician to the Polish-Italian bride of King John Sigismund, who aided the development of anti-Trinitarianism in Poland and Transylvania. In 1569 Racow was founded as the Polish Brethren's central community.

Faustus Socinus went to Poland in 1579. He rejected Anabaptist insistence on immersionist adult baptism and affirmed that Jesus Christ was a man whom God had resurrected and to whom he had given all power in heaven and earth over the church. Socinus emphasized the validity of prayer to Christ as an expression of honour and as a request for aid. Through his ability in theological debate he soon became the leader of the Polish Brethren, whose adherents were frequently referred to as Socinians.

After Socinus' death his followers published the *Racovian Catechism* (1605). The hostility of their opponents, however, caused the destruction of the Socinians' famous printing press and school at Racow (1632). In 1658 a legislative decree was enacted stating that by 1660 the Socinians must either become Roman Catholics, go into exile, or face execution. A few of these Polish exiles reached Kolozsvár, centre of the Transylvanian Unitarian movement, and some of their leaders moved to the Netherlands, where they continued the publication of Socinian books.

Transylvanian Unitarianism. Blandrata encouraged Ferenc Dávid (1510–79), a Transylvanian theologian, to deliver anti-Trinitarian sermons. Study at Wittenberg had led Dávid to convert from Roman Catholicism to Lutheranism. As superintendent of Transylvanian Lutheran churches Dávid had engaged in debates with Pe-

Ferenc Dávid

ter Melius, leader of the Transylvanian Reformed Church, with the result that Dávid had joined the Reformed Church, of which he soon became superintendent. Cooperation between Dávid and Blandrata led to the publication of two Unitarian books, *De falsa et vera unius Dei Patris* (1567; "On the False and True Unity of God the Father, Son, and Holy Spirit") and *De regno Christi . . .* (1569; "On the Reign of Christ"), which showed the influences of Servetus and Laelius Socinus.

Biblical study and discussions with colleagues (*e.g.*, with Jacobus Palaeologus) led Dávid to nonadorantism (denial that prayer should be addressed to Christ), which caused a serious crisis. In 1568 John Sigismund, Unitarian king of Transylvania, granted religious freedom to Catholics, Lutherans, the Reformed Church, and those who were soon to be called Unitarians, and in 1571 the Transylvanian Diet gave constitutional recognition to all four received religions. But Sigismund's successor, Stephen Báthory, forbade further innovations (changes in doctrine from beliefs held during Sigismund's reign). Dávid's nonadorantist innovation thus endangered the Unitarians' legal status. Blandrata sought to protect them by the arrest and trial of Dávid, who died in prison in 1579. This oldest Unitarian Church survives in Hungary and Romania.

English Unitarianism. John Biddle (1615–62), an English Socinian, whose knowledge of the Greek text of the New Testament convinced him that the doctrine of the Trinity was not of scriptural origin, published his Unitarian convictions in *Twelve Arguments Drawn out of Scripture . . .* (1647) and other works; English readers, moreover, were exposed to Unitarian views through Socinian books published in the Netherlands. Although the Toleration Act of 1689 excluded Unitarians, advocates of an Arian Christology (belief in Christ's preexistence as a subordinate, divine, created being) soon appeared within the Church of England and among Dissenters. This led some Anglicans to seek, without success, the rescinding of the requirement of subscription to the Anglican Thirty-nine Articles. Dissenting ministers, meeting in the Salters' Hall in London in 1719, separated into two groups, one insisting on adherence to confessional documents, the other requiring only agreement with Scripture. Of those in the second group, Presbyterians, General Baptists, and a few independents gradually moved during the 18th century with their congregations toward Unitarian views.

The first English Unitarian congregation, Essex Street Chapel, was founded in London in 1774 by Theophilus Lindsey, who previously had been an Anglican clergyman. The scientist and dissenting minister Joseph Priestley (1733–1804) influenced Unitarian ministers by his scriptural rationalism, materialist determinism, and emphasis on a humanitarian Christology. The scholar and theologian Thomas Belsham supported Priestley's emphasis on a humanitarian Christology and opposition to Arian views. The British and Foreign Unitarian Association was founded in 1825.

In the 19th century Parliament was persuaded to repeal some of the laws against nonconformity, which freed the Unitarians for a more active church life. English Unitarians, moreover, were greatly influenced by James Martineau (1805–1900), who, after studies in Germany, was led to a religious epistemology emphasizing intuition. In 1928 a union of the British and Foreign Unitarian Association with the National Conference (which included other Free Christian Churches) resulted in the founding of the General Assembly of Unitarian and Free Christian Churches. Unitarianism is also present in Wales, Scotland, and the Non-Subscribing Presbyterian Church of Ireland.

American Unitarianism. In the American colonies Congregationalist ministers influenced by Arian Christology and by Arminian theology, gradually moved in the 18th century toward Unitarian views. Conflicts with supporters of Jonathan Edwards' theological heritage resulted in the election at Harvard College of a liberal, Henry Ware, as Hollis Professor of Divinity in 1805. When the liberal Congregationalists were accused of agreeing with Belsham's strictly humanitarian Christology, the Unitarian clergyman William Ellery Channing defended them as Arians. Channing's 1819 sermon "Unitarian Christianity," a manifesto, presented both a recognition that the liberals would have to separate from the Congregational Church and a coherent theology. In 1825 the American Unitarian Association (AUA), an association of individuals, was organized.

American Unitarian Association

Channing's Arian Christology as well as his affirmations of the divine unity, the authority of Scripture rationally interpreted, and an optimistic view of human nature were dominant among early American Unitarians. His Lockean epistemology (modified by views of Scottish commonsense philosophers and the English Unitarian Richard Price), however, was challenged by such Transcendentalists as Ralph Waldo Emerson, in his "Divinity School Address" (1838), and Theodore Parker, in his sermon "The Transient and Permanent in Christianity" (1841), both of whom emphasized intuition and moral idealism. Parker's leadership in addressing issues of social reform, such as issues relating to the anti-slavery movement, made a lasting impact on Unitarians.

Although Transcendentalism divided the Unitarians, Henry Whitney Bellows, a prominent figure in Unitarianism after the Civil War, succeeded in organizing the National Conference of Unitarian Churches in 1865. A separatist Free Religious Association (FRA) was organized in 1867 by persons who, although holding a variety of views, were agreed in their opposition to the preamble of the National Conference's constitution, which was virtually a Christian creed. A Western Unitarian Conference, organized in 1852, also experienced a controversy over whether Unitarianism was to include persons whose views were not theistic and Christian. In 1894 a revision in the constitution of the National Conference enabled members of the FRA to rejoin the Conference. Later renamed the General Conference, it merged with the AUA in 1925.

In the 20th century religious humanism, the endeavour to reformulate liberal theology on strictly non-theistic grounds, emerged within Unitarianism, leading to a theist-humanist controversy. After such Unitarian ministers as John Dietrich and Curtis Reese signed the Humanist Manifesto (1933), religious humanism became the view of many Unitarians. A Commission of Appraisal (1934–36) recommended modifications in the structure and program of the AUA. Frederick May Eliot, chairman of the commission, was persuaded to become president of the AUA, and while in office he prepared the denomination for future growth. In the 1930s a critical movement emerged, largely in response to a general crisis of faith in liberal thought; its leader was James Luther Adams, whose writings contributed significantly to Unitarian theology and social thought. Of particular importance for Unitarianism today are his studies of voluntary associations and their implications (*On Being Human—Religiously*, 1976).

Early Universalism. Radical Pietism emerged in Germany under the leadership of Johann Wilhelm Petersen, who led groups of Philadelphian Pietists identifying themselves with the sixth church referred to in Revelation 3:7–13. A Philadelphian Society was organized in London in 1681 under Jane Leade, whose religious views were based on the thought of the German mystic Jakob Böhme and on her own visions and dreams. Convinced that Leade was correct in affirming a universal restoration (the ultimate reconciliation to God of all human beings, the devil, and his angels), Petersen gave her views scriptural foundations in his *Mystery of the Restitution of All Things* (1700–10), which included *The Everlasting Gospel*, a restorationist treatise by George Klein-Nicolai published under the pseudonym Paul Siegvolck. German Philadelphian Pietists took these and other works to Pennsylvania in the early 18th century, where George de Benneville (1703–93), a French Universalist who had gone to Pennsylvania in 1741, brought them into contact with other groups that affirmed universal salvation.

A different view of Universalism appeared in the work of the Welsh revivalist preacher James Relly (1720–78). In his *Union, or A Treatise of the Consanguinity and Affinity Between Christ and His Church* (1759) he presented scriptural texts for the view that universal salvation is assured. Christ's unity with all human beings and his acceptance of the guilt and endurance of the punishment

James Relly

for the sins of mankind ensured that among the elect for whom Christ had suffered was the entire human race. The English Methodist John Murray (1741–1815) unsuccessfully sought to refute Relly's views; instead he became convinced of their truth and took this theology to New England in 1770. His church at Gloucester, Mass. (1780), was the first American Universalist congregation.

Urged by George de Benneville to read *The Everlasting Gospel* and other Universalist works, Elhanan Winchester (1751–97), a Baptist minister, became converted to restorationist Universalism. He traveled to England, where he founded a Universalist Church in London in 1793 and wrote *The Universal Restoration* . . . (1794). He emphasized scriptural texts that affirmed the finite and remedial nature of punishment after death. Winchester subsequently continued his ministry in the United States.

American Universalism. Hosea Ballou (1771–1852) was the greatest 19th-century American Universalist leader. His *A Treatise on Atonement* . . . (1805) converted most Universalist ministers to a Unitarian view of God, an Arian Christology, and the view that, because sin is finite in nature and all of its effects will be experienced in this life, all of mankind will be saved after death. Ballou later abandoned his Arian belief in Christ's preexistence.

The Winchester Profession

The Winchester Profession (1803), adopted by the General Convention of Universalists in the New England States at Winchester, N.H., was phrased in general terms to embrace differing Universalist views. In 1870, however, a resolution adopted by the General Convention required that the Winchester Profession be interpreted as requiring belief in the authority of Scripture and the lordship of Jesus Christ. This restriction was rescinded in 1899.

Ballou's theology was dominant during the first half of the 19th century, when Universalist ministers founded congregations in many states. Opposed to Ballou's theology, however, was a small group of ministers and laypersons, who left the denomination to form the Massachusetts Association of Universal Restorationists, which existed from 1831 to 1841. Although both factions believed that there would be no eternal punishment for sinners after death, the Massachusetts restorationists embraced the position that there would be a limited punishment followed by a general restoration to God. Adin Ballou (1803–90), a leading restorationist, was an outstanding advocate of the application of New Testament ethics to social issues. By the end of the 19th century most Universalists held restorationist views.

Clarence Skinner (1881–1949), dean of Crane Theological School, greatly influenced American Universalists by his emphasis on social issues and his reinterpretation of Universalism as referring not to salvation after death but to the unities and universals in human life (*A Religion for Greatness*, 1945). In 1935 the Universalists adopted a noncreedal Bond of Fellowship, which they revised in 1953. Clinton Lee Scott and Kenneth Patton affirmed religious humanism and emphasized drawing religious sustenance from the traditions of the world's great religions.

TEACHINGS

The Unitarian theologian Earl Morse Wilbur (1866–1956) advanced the thesis, now widely accepted, that the history of Unitarianism in Poland, Transylvania, England, and America gains unity from certain common themes. These themes are freedom of religious thought rather than required agreement with creeds or confessions, reliance not on tradition or external authority but on the use of reason in formulating religious beliefs, and tolerance of differing religious views and customs in worship and polity.

Unitarian Universalists are creedless and deny the authority of dogmas promulgated by church councils. Their teachings historically have included the unity of God, the humanity of Jesus, mankind's religious and ethical responsibility, and the possibility of attaining religious salvation through differing religious traditions. They emphasize the authority of the individual's religious conviction, the importance of religiously motivated action for social reform, democratic method in church governance, and reason and experience as appropriate bases for formulating religious beliefs. Their traditional concern for social issues has caused Unitarian Universalists to give active support to the demands for equality of blacks, feminists, and other groups. Gains in equality for women within the Unitarian Universalist Association were significant, but its predominantly white, middle-class membership remains an issue.

Although the nonadorantist Unitarians in Romania and Hungary are firmly Christian, in England, the United States, and Canada, the beliefs of Unitarians range from Unitarian Christianity to religious humanism; there are also aspirations toward becoming a universal religion. Universalist teachings have changed also; whereas the restorationist theology that was dominant among American Universalists toward the end of the 19th century emphasized the salvation of all after death, many 20th-century Universalists affirm a naturalistic worldview and regard salvation as an aspect of present human experience.

WORSHIP AND ORGANIZATION

English and American Unitarian Universalist worship is predominantly thematic in emphasis and sermon-centred in form. It makes use of hymnals that have been revised to reflect changing religious interests; for example, today's hymns express themes of religious humanism. There also is some liturgical experimentation. Whereas baptism and frequent observance of the Lord's Supper characterize Hungarian and Romanian Unitarian worship, in England and the United States infants may be dedicated and observance of the Lord's Supper is rare, except among Unitarian Christians.

Unitarian Universalist Association

The American Unitarian Association and the Universalist Church of America merged in 1961 to form the Unitarian Universalist Association (UUA). The UUA's churches and fellowships are located primarily in the United States and Canada. (Canadian congregations are also members of the Canadian Unitarian Council.) The UUA is a member of the International Association for Religious Freedom, which was founded in 1900 as the International Association for Liberal Christianity and Religious Freedom (its name being changed in 1969 to reflect the inclusion of member-groups from non-Christian religious traditions).

English Unitarians and American Unitarian Universalists have congregational polity and emphasize the democratic process. Ministerial and lay delegates from congregations constitute the annual General Assembly, a legislative body. In Hungary and Romania a bishop and a lay president in each country supervise the Unitarian churches, which are governed by annual synods. (J.C.G.)

BIBLIOGRAPHY

General: EINAR MOLLAND, *Christendom: The Christian Churches, Their Doctrines, Constitutional Forms, and Ways of Worship* (1959, originally published in Swedish, 1953); JOHN DILLENBERGER and CLAUDE WELCH, *Protestant Christianity Interpreted Through Its Development*, 2nd ed. (1988); JOHN S. WHALE, *The Protestant Tradition* (1955, reprinted 1962), a summary of the creedal positions of Protestant bodies; WILHELM PAUCK, *The Heritage of the Reformation*, rev. and enl. ed. (1961, reissued 1968), essays on the theological and practical impact of Protestantism; B.A. GERRISH, *The Old Reformation and the New: Essays on the Reformation Heritage* (1982), a study connecting the theology of the early years of Protestantism with recent developments; ROBERT MCAFEE BROWN, *The Spirit of Protestantism* (1961; reissued 1974), a summary of the main themes of Protestant life; MARTIN E. MARTY, *Protestantism* (1972; reissued 1974), with extensive bibliographic essays; JAMES HASTINGS NICHOLS, *Primer for Protestants* (1947; reissued 1971), a brief survey of Protestant history and theology for the layperson; JOHN B. COBB, JR., *Varieties of Protestantism* (1960), a theological analysis of alternatives in Protestantism; and *Living Options in Protestant Theology: A Survey of Methods* (1962, reissued 1986); ROGER MEHL, *The Sociology of Protestantism* (1970; originally published in French, 1965), an excellent survey of Protestant sociology; JAROSLAV PELIKAN, *The Christian Tradition: A History of the Development of Doctrine*, vol. 4, *Reformation of Church and Dogma (1300–1700)* (1984), a magisterial approach; GEORGE H. WILLIAMS, *The Radical Reformation* (1962), a comprehensive and authoritative work in English on this subject; FRANKLIN HAMLIN LITTELL, *The Origins of Sectarian Protestantism: A Study of the Anabaptist View of the Church* (1964, reprinted 1968), a historical analysis of the main themes in the radical Reformation; LOUIS BOUYER, *The Spirit and Forms of Protestantism* (1955, reprinted 1968; originally published in French, 1954); J. LESLIE DUNSTAN (ed.),

Protestantism (1961, reissued 1969), a study combining sources with narrative and interpretation; PAUL TILLICH, *The Protestant Era*, trans. from German (1948, reissued 1951), a collection of essays, one of which discusses the "end of the Protestant era"; CHARLES W. KEGLEY, *Protestantism in Transition* (1965), a theologian's survey of Protestant tendencies after the mid-20th century; ERNST TROELTSCH, *Protestantism and Progress: The Significance of Protestantism for the Rise of the Modern World* (1986; originally published in German, 2nd ed., 1911), a classic interpretation of Protestant contributions to modernity; MAX WEBER, *The Protestant Ethic and the Spirit of Capitalism* (1930, reissued 1985; originally published in German, 1904), a much debated study of the link between Protestantism and the rise of capitalism; JOHN A. HARDON, *The Protestant Churches of America*, rev. ed. (1969), a summary by a Roman Catholic of Protestant doctrinal positions; FREDERICK E. MAYER, *The Religious Bodies of America*, 4th ed. rev. by ARTHUR CARL PIEPKORN (1961), denomination-by-denomination study of doctrinal positions in American religious groups; ARTHUR CARL PIEPKORN, *Profiles in Belief: The Religious Bodies of the United States and Canada*, vol. 2, *Protestant Denominations* (1978), vol. 3, *Holiness and Pentecostal* (1979), and vol. 4, *Evangelical, Fundamentalist, and Other Christian Bodies* (1979), an extensive review of American Protestant bodies; WINTHROP S. HUDSON, *American Protestantism* (1961, reprinted 1972), a brief survey of Protestant history in America; JERALD C. BRAUER, *Protestantism in America: A Narrative History*, rev. ed. (1965, reprinted 1974), a presentation of the main themes of American Protestant history; ANDREW L. DRUMMOND, *Story of American Protestantism* (1949, reissued 1951), a British view of Protestant history in the United States; and MARTIN E. MARTY, *Righteous Empire: The Protestant Experience in America* (1970, reprinted 1977), written for the nation's bicentennial. Newsworthy developments in the Protestant Church are chronicled in *Christian Century* (weekly).

History: KENNETH SCOTT LATOURETTE, *A History of Christianity*, rev. ed., 2 vol. (1975), with useful bibliographies; and ÉMILE G. LÉONARD, *Histoire générale du protestantisme*, 3 vol. (1961–64)—vol. 1 has also appeared in English with the title, *A History of Protestantism* (1968). Additional references may be found in OWEN CHADWICK, *The History of the Church: A Select Bibliography*, 3rd ed. (1973).

For Puritanism, see WILLIAM HALLER, *The Rise of Puritanism: or, The Way to the New Jerusalem as Set Forth in Pulpit and Press from Thomas Cartwright to John Lilburne and John Milton, 1570–1643* (1938, reissued 1984); CHRISTOPHER HILL, *Society and Puritanism in Pre-Revolutionary England* (1964, reissued 1986); PATRICK COLLINSON, *The Elizabethan Puritan Movement* (1967); SAMUEL ELIOT MORISON, *The Intellectual Life of Colonial New England*, 2nd ed. (1956, reprinted 1980); and FRANCIS J. BREMER, *The Puritan Experiment: New England and Society from Bradford to Edwards* (1976). For Arminianism, see A.W. HARRISON, *The Beginnings of Arminianism to the Synod of Dort* (1926); and CARL BANGS, *Arminius: A Study in the Dutch Reformation*, 2nd ed. (1985). For Pietism, see KOPPEL S. PINSON, *Pietism as a Factor in the Rise of German Nationalism* (1934, reissued 1968); and F. ERNEST STOEFFLER, *The Rise of Evangelical Pietism* (1965, reprinted 1971), and *German Pietism During the Eighteenth Century* (1973). For Protestant missionary expansion, see KENNETH SCOTT LATOURETTE, *A History of the Expansion of Christianity*, vol. 3–7 (1940–45); and STEPHEN NEILL, *A History of Christian Missions*, 2nd ed. rev. by OWEN CHADWICK (1986). For the 19th and 20th centuries, see KENNETH SCOTT LATOURETTE, *Christianity in a Revolutionary Age*, 5 vol. (1958–62, reissued 1973); STEPHEN NEILL (ed.), *Twentieth Century Christianity: A Survey of Modern Religious Trends by Leading Churchmen*, rev. ed. (1963); and DAVID B. BARRETT (ed.), *World Christian Encyclopedia: A Comparative Study of Churches and Religions in the Modern World, AD 1900–2000* (1982).

For American Protestantism, see H. SHELTON SMITH, ROBERT T. HANDY, and LEFFERTS A. LOETSCHER, *American Christianity: An Historical Interpretation with Representative Documents*, 2 vol. (1960–63), a general guide; E.S. GAUSTAD, *A Documentary History of Religion in America*, 2 vol. (1982–83), a comprehensive overview; WILLIAM WARREN SWEET, *The Story of Religion in America*, 2nd rev. ed. (1950); WINTHROP S. HUDSON, *American Protestantism* (1961, reprinted 1972); and R.T. HANDY, *A Christian America: Protestant Hopes and Historical Realities*, 2nd ed. rev. and enl. (1984), on cultural intentions. For the social Gospel, see C.H. HOPKINS, *The Rise of the Social Gospel in American Protestantism, 1865–1915* (1940, reprinted 1982). For churches under the Nazis, see J.S. CONWAY, *The Nazi Persecution of the Churches, 1933–45* (1968); and ARTHUR C. COCHRANE, *The Church's Confession Under Hitler*, 2nd ed. (1976), which accents resistance documents. For the ecumenical movement, see RUTH ROUSE and STEPHEN NEILL (eds.), *A History of the Ecumenical Movement, 1517–1948*, 3rd ed. (1986); and HAROLD E. FEY (ed.), *A History of the Ecumenical Movement, 1948–1968: The Ecumenical Advance*, 2nd ed. (1986). Research findings related to primarily American Protestant church history are published in *Church History* (quarterly).

Lutheran churches: CONRAD BERGENDOFF, *The Church of the Lutheran Reformation* (1967), a survey of Lutheranism, but especially useful for information on Lutheranism in the Scandinavian countries, in a very readable narrative with bibliography; and JAROSLAV PELIKAN, *From Luther to Kierkegaard: A Study in the History of Theology* (1950, reprinted 1963), a brief history of developments in Lutheran theology to the mid-19th century. JULIUS BODENSIECK (ed.), *The Encyclopedia of the Lutheran Church*, 3 vol. (1965), is the standard English reference work on Lutheranism, although articles vary from the scholarly to the propagandistic.

On Lutheranism in the United States, see SYDNEY E. AHLSTROM, "Theology in America: A Historical Survey," in JAMES WARD SMITH and A. LELAND JAMISON, *Religion in American Life*, vol. 1, *The Shaping of American Religion* (1961), pp. 232–321, a survey that sets Lutheran theology in the context of other developments; and ABDEL R. WENTZ, *A Basic History of Lutheranism in America*, rev. ed. (1964), a standard work on the major developments. E. CLIFFORD NELSON, *The Rise of World Lutheranism: An American Perspective* (1982), is a discussion of the cooperation among 20th-century Lutheran churches. E. CLIFFORD NELSON (ed.), *Lutherans in North America*, rev. ed. (1980), is a study on the various periods of American history.

Works on Lutheran teachings include THEODORE G. TAPPERT (ed. and trans.), *The Book of Concord: The Confessions of the Evangelical Lutheran Church* (1959, reprinted 1987), official translation; EDMUND SCHLINK, *Theology of the Lutheran Confessions* (1961, reissued 1975; originally published in German, 1940), a dialectical approach to Lutheran theology; WILHELM MAURER, *Historical Commentary on the Augsburg Confession* (1986; originally published in German, 2 vol., 1976–78), a thorough work on the basic Lutheran document; PAUL TILLICH, *Systematic Theology*, 3 vol. (1951–63, reprinted 1973), a systematics by the most original Lutheran theologian of the 20th century; and KARL FERDINAND MÜLLER and WALTER BLANKENBURG, *Leiturgia: Handbuch des evangelischen Gottesdienstes*, 5 vol. (1954–70), a historical and theological examination of the Lutheran service.

Reformed and Presbyterian churches: WILLIAM J. BOUWSMA, *John Calvin: A Sixteenth-Century Portrait* (1988), which places Calvin in his contemporary context; JOHN T. MCNEILL, *The History and Character of Calvinism* (1954, reissued 1973), a comprehensive treatment of the rise and development of Presbyterian and Reformed churches, with an excellent bibliography; JAMES HASTINGS NICHOLS, *Corporate Worship in the Reformed Tradition* (1968), an overview of the variety of forms developed in the Reformed tradition of the public worship of God; HEINRICH HEPPE, *Reformed Dogmatics Set Out and Illustrated from the Sources*, rev. and ed. by ERNST BIZER (1950, reprinted 1978; originally published in German, 1861), a work enabling the reader to get beyond Calvin's *Institutes* to some acquaintance with other Reformed theologians of the 16th and 17th centuries; ROBERT MCAFEE BROWN, *Theology in a New Key: Responding to Liberation Themes* (1978), and *Unexpected News: Reading the Bible with Third World Eyes* (1981), interpretations of Third World theology; JOHN H. LEITH, *An Introduction to Reformed Tradition: A Way of Being the Christian Community*, rev. ed. (1981); ARTHUR C. COCHRANE (ed.), *Reformed Confessions of the 16th Century* (1966), 12 classic confessions of the 16th century, with historical introductions; and THOMAS F. TORRANCE (ed. and trans.), *The School of Faith: The Catechisms of the Reformed Church* (1959), 10 catechisms of the 16th and 17th centuries. Useful periodicals include *Reformed World* (quarterly), published by the World Alliance of Reformed Churches (Presbyterian and Congregational), which reports on the life and work of Reformed and Presbyterian churches throughout the world; and *American Presbyterians: Journal of Presbyterian History* (quarterly), on all aspects of American Presbyterian history.

Anglican Communion: STEPHEN C. NEILL, *Anglicanism*, rev. ed. (1977), the most comprehensive treatment of Anglican history; JOHN R.H. MOORMAN, *A History of the Church in England*, 3rd ed. (1973, reissued 1980), the basic facts about Anglicanism's mother church; MARION J. HATCHETT, *Commentary on the American Prayer Book* (1981), a detailed examination of Anglican worship and its rationale; JAMES T. ADDISON, *The Episcopal Church in the United States, 1789–1931* (1951, reissued 1969), the most thorough treatment of American Anglicanism; and PAUL A. WELSBY, *A History of the Church of England, 1945–1980* (1984, reissued 1986), which explains recent changes in the Church of England. See also RAYMOND W. ALBRIGHT, *A History of the Protestant Episcopal Church* (1964), the standard story of American Episcopalianism.

Baptists: ROBERT G. TORBET, *A History of the Baptists*, 3rd ed. (1963, reprinted 1973), the most complete account of the Baptists; H. LEON MCBETH, *The Baptist Heritage* (1987), a comprehensive history of four centuries of Baptist witness; ALFRED C. UNDERWOOD, *A History of the English Baptists* (1947), which gives major attention to Baptist beginnings; JAMES E. WOOD, JR., *Baptists and the American Experience* (1976), a collection of essays; WINTHROP S. HUDSON, *Baptists in Transition: Individualism and Christian Responsibility* (1979); and ALBERT W. WARDIN, JR., *Baptist Atlas* (1980), international in scope. See also JAMES MELVIN WASHINGTON, *Frustrated Fellowship: The Black Baptist Quest for Social Power* (1986); and WILLIAM H. BRACKNEY (ed.), *Baptist Life and Thought, 1600–1980: A Source Book* (1983). NORMAN H. MARING and WINTHROP S. HUDSON, *A Baptist Manual of Polity and Practice* (1963), gives details of ecclesiastical organization. *Baptist History and Heritage* (quarterly) deals primarily with Southern Baptists.

Congregationalists: WILLISTON WALKER, *The Creeds and Platforms of Congregationalism* (1893, reprinted 1960); GEOFFREY F. NUTTALL, *Visible Saints: The Congregational Way, 1640–1660* (1957); RAYMOND P. STEARNS, *Congregationalism in the Dutch Netherlands: The Rise and Fall of the English Congregational Classis, 1621–1635* (1940); GAIUS GLENN ATKINS and FREDERICK L. FAGLEY, *History of American Congregationalism* (1942); WILLIAM WARREN SWEET, *Religion in Colonial America* (1942, reissued 1965); DOUGLAS HORTON, *Congregationalism: A Study in Church Polity* (1952), and *The United Church of Christ: Its Origins, Organization, and Role in the World Today* (1962); LOUIS H. GUNNEMANN, *The Shaping of the United Church of Christ: An Essay in the History of American Christianity* (1977), which looks at the merger of Congregationalism with the Evangelical and Reformed Church; R. TUDUR JONES, *Congregationalism in England, 1662–1962* (1962); and BROR WALAN, *Församlingstanken i Svenska missionsförbundet: en studie i de nyevangeliska rörelsens sprängning och Svenska missionsförbundets utveckling* (1964), a study of the Covenant Church of Sweden, with an English summary. Modern interpretative essays include DANIEL T. JENKINS, *Congregationalism: A Restatement* (1954); ERIK ROUTLEY, *English Religious Dissent* (1960); and NORMAN GOODALL (ed.), *Der Kongregationalismus* (1973).

Friends: Good introductions to Quakerism may be found in FRIENDS WORLD COMMITTEE FOR CONSULTATION, *Handbook of the Religious Society of Friends*, 5th ed. (1967); and in the interpretation by D. ELTON TRUEBLOOD, *The People Called Quakers* (1966, reissued 1971). The standard histories are WILLIAM CHARLES BRAITHWAITE, *The Beginnings of Quakerism*, 2nd ed. rev. by HENRY J. CADBURY (1955, reissued 1970), and a companion volume, *The Second Period of Quakerism*, 2nd ed. rev. by HENRY J. CADBURY (1961, reissued 1979); and RUFUS M. JONES, *The Quakers in the American Colonies* (1911, reissued 1966), and *The Later Periods of Quakerism*, 2 vol. (1921, reprinted 1970). More specialized works include ELIZABETH ISICHEI, *Victorian Quakers* (1970); RICHARD T. VANN, *The Social Development of English Quakerism, 1655–1755* (1969); JACK D. MARIETTA, *The Reformation of American Quakerism, 1748–1783* (1984); BARRY REAY, *The Quakers and the English Revolution* (1985); and MARY MAPLES DUNN and RICHARD S. DUNN (eds.), *The Papers of William Penn* (1981–), with 4 vol. published by 1987. The masterpiece of Quaker theology is ROBERT BARCLAY, *Apology for the True Christian Divinity* (1678, reissued 1967; originally published in Latin, 1676). WILLIAM CHARLES BRAITHWAITE, *Spiritual Guidance in the Experience of the Society of Friends* (1909, reissued 1941), best explains how Friends' polity should work. See also GLADYS WILSON, *Quaker Worship: An Introductory Historical Study of the English Friends' Meeting* (1952); and CLARENCE E. PICKETT, *For More Than Bread: An Autobiographical Account of Twenty-Two Years Work with the American Friends Service Committee* (1953). Quaker social thought on contemporary issues may be found in STELLA ALEXANDER (comp.), *Quaker Testimony Against Slavery and Racial Discrimination: An Anthology* (1958); and *Towards a Quaker View of Sex: An Essay*, rev. ed. (1964, reprinted 1976), published by the Friends Home Service Committee.

Methodists: GORDON RUPP, *Religion in England, 1688–1791* (1986), on the background and rise of Methodism; RUPERT DAVIES, GORDON RUPP, and A. RAYMOND GEORGE (eds.), *A History of the Methodist Church in Great Britain*, 4 vol. (1966–88), authoritative for history, doctrine, and missions; EMORY STEVENS BUCKE (ed.), *The History of American Methodism*, 3 vol. (1964), comprehensive and authoritative; RUPERT DAVIES, *Methodism*, 2nd rev. ed. (1985), a general survey, chiefly from the British point of view; UMPHREY LEE and WILLIAM WARREN SWEET, *A Short History of Methodism* (1956), concerned mainly with American Methodism; V.H.H. GREEN, *The Young Mr. Wesley* (1961, reissued 1963), a critical study of Wesley's Oxford days; MARTIN SCHMIDT, *John Wesley: A Theological Biography*, 2 vol. in 3 (1962–73; originally published in German, 2 vol., 1953–66); RICHARD P. HEITZENRATER, *The Elusive Mr. Wesley*, 2 vol. (1984), with material by Wesley and by his contemporaries and biographers; A.C. OUTLER (ed.), *The Works of John Wesley*, vol. 1–4 (1984–87), a collection of his sermons; COLIN W. WILLIAMS, *John Wesley's Theology Today* (1960, reissued 1969), a systematic exposition; RUPERT DAVIES, *What Methodists Believe*, 2nd ed. (1988). Methodist beliefs in their ecumenical setting: GERALD F. MOEDE, *The Office of Bishop in Methodism* (1965), a history of constitutional developments in American Methodism; WADE C. BARCLAY, *History of Methodist Missions*, 4 vol. (1949–73), about American overseas missions; JOHN KENT, *The Age of Disunity* (1966), on division and reunion in British Methodism; BERNARD SEMMEL, *The Methodist Revolution* (1973), the social, economic, and political effect of Methodism on Britain; JOHN M. MOORE, *The Long Road to Methodist Union* (1943), on the reunion of American Methodism; and JOHN MUNSEY TURNER, *Conflict and Reconciliation: Studies in Methodism and Ecumenism in England, 1740–1982* (1985). See also METHODIST CHURCH (GREAT BRITAIN), *The Methodist Service Book* (1975), and *Hymns and Psalms* (1983), worship in British Methodism. Aspects of the church's history are discussed in *Methodist History* (quarterly).

Disciples of Christ: LEROY GARRETT, *The Stone-Campbell Movement: An Anecdotal History of Three Churches* (1981), which gives major attention to the 19th century; JAMES DEFOREST MURCH, *Christians Only: A History of the Restoration Movement* (1962), reflecting an independent viewpoint; and WILLIAM E. TUCKER and LESTER G. MCALLISTER, *Journey in Faith: A History of the Christian Church (Disciples of Christ)* (1975), centring on the cooperative stream. Biographical studies include WILLIAM GARRETT WEST, *Barton Warren Stone: Early American Advocate of Christian Unity* (1954); and LESTER G. MCALLISTER, *Thomas Campbell: Man of the Book* (1954). Beliefs, worship, and organization are treated in ALEXANDER CAMPBELL, *The Christian System, in Reference to the Union of Christians and a Restoration of Primitive Christianity as Plead in the Current Reformation*, 2nd ed. (1839, reprinted 1980), the classic summary of Campbell's theology; ROYAL HUMBERT (ed.), *A Compend of Alexander Campbell's Theology* (1961), with critical and historical commentary; KEITH WATKINS, *The Breaking of Bread: An Approach to Worship for the Christian Churches (Disciples of Christ)* (1966), a comprehensive, historical, and theological analysis; DAVID EDWIN HARRELL, JR., *A Social History of the Disciples of Christ*, 2 vol. (1966–73); and KENNETH LAWRENCE (ed.), *Classic Themes of Disciples Theology: Rethinking the Traditional Affirmations of the Christian Church (Disciples of Christ)* (1986). See also *Mid-Stream: An Ecumenical Journal* (quarterly).

Unitarians and Universalists: EARL MORSE WILBUR, *A History of Unitarianism*, 2 vol. (1945), which remains basic for understanding Unitarianism; C. GORDON BOLAM *et al.*, *The English Presbyterians, from Elizabethan Puritanism to Modern Unitarianism* (1968); CONRAD WRIGHT, *The Beginnings of Unitarianism in America* (1955, reissued 1976), which portrays the 18th century; and CONRAD WRIGHT (ed.), *A Stream of Light: A Sesquicentennial History of American Unitarianism* (1975). RICHARD EDDY, *Universalism in America*, 3rd ed., 2 vol. (1891–94), still useful, has been followed by RUSSELL E. MILLER, *The Larger Hope*, vol. 1, *The First Century of the Universalist Church in America, 1770–1870* (1979), and vol. 2, *The Second Century of the Universalist Church in America, 1870–1970* (1985). ERNEST CASSARA (ed.), *Universalism in America: A Documentary History*, 2nd ed. (1984); and GEORGE HUNTSTON WILLIAMS, *American Universalism: A Bicentennial Historical Essay*, 2nd ed. (1976), are important supplements to Miller. DAVID ROBINSON, *The Unitarians and the Universalists* (1985), is a study of the merged denominations.

Evangelicals, Fundamentalists, and Pentecostals: LEONARD I. SWEET, *The Evangelical Tradition in America* (1984), which includes an extensive superior bibliography on all phases of the movements in the United States; WILLIAM G. MCLOUGHLIN (ed.), *The American Evangelicals, 1800–1900* (1968, reissued 1976), an anthology with a helpful introduction; GEORGE M. MARSDEN, *Fundamentalism and American Culture: The Shaping of Twentieth Century Evangelicalism, 1870–1925* (1980, reprinted 1982), the most important work on the subject; DEAN M. KELLEY, *Why Conservative Churches Are Growing: A Study in Sociology of Religion*, new ed. (1977, reprinted 1986), a landmark work signaling a power shift in the United States; CHARLES EDWIN JONES, *A Guide to the Study of the Holiness Movement* (1974), and *A Guide to the Study of the Pentecostal Movement* (1983), which provide access to important materials; and ROBERT MAPES ANDERSON, *Vision of the Disinherited: The Making of American Pentecostalism* (1979), dealing with the early years of the movement.

(W.O.C./R.H.B./J.C.S./E.C.N./M.E.M./R.P.S./R.S.De./W.L.Sa./W.S.H./D.T.J./R.T.V./R.E.D./R.E.O./J.C.G.)

Protists

Protists are eukaryotic, predominantly microscopic organisms. They may share certain morphological and physiological characteristics with animals or plants, or both. The protists comprise what have traditionally been called protozoa, algae, and lower fungi.

From the time of Aristotle, near the end of the 4th century BC, until well after the middle of the 20th century, the entire biotic world was generally considered divisible into just two great kingdoms, the plants and the animals. The separation was based on the assumption that plants are pigmented (basically green), nonmotile (most commonly from being rooted in the soil), photosynthetic and therefore capable solely of self-contained (autotrophic) nutrition, and unique in possessing cellulosic walls around their cells. By contrast, animals are without photosynthetic pigments (colourless), actively motile, nutritionally phagotrophic (and therefore required to capture or absorb important nutrients), and without walls around their cells.

When microscopy arose as a science in its own right, botanists and zoologists discovered evidence of the vast diversity of life mostly invisible to the unaided eye. With rare exception, authorities of the time classified such microscopic forms as minute plants (called algae) and minute animals (called "first animals," or protozoa). Such taxonomic assignments went essentially unchallenged for many years, despite the fact that the great majority of these minute forms of life—not to mention certain macroscopic ones, various parasitic forms, and the entire group known as the fungi—did not possess the cardinal characteristics on which the "plants" and "animals" had been differentiated and thus had to be forced to fit into those kingdom categories.

An authority who took exception to the imposition of the plant and animal categories on the protists was the German zoologist Ernst Haeckel. In 1866 he proposed a third kingdom, the Protista, to embrace such "lower" organisms, but his conception failed to gain widespread support during his lifetime. Some 80–90 years later, Herbert F. Copeland, an American botanist, attempted a revival of the protist concept, but again without much success.

The basis for a major change in the systematics of these lower forms came through an advancement in the concept of the composition of the biotic world. About 1960, resurrecting and embellishing an idea originally conceived 20 years earlier by the French marine biologist Edouard Chatton but universally overlooked, R.Y. Stanier, C.B. Van Niel, and their colleagues formally proposed the division of all living things into two great groups, the prokaryotes and the eukaryotes. (Prokaryotes—bacteria and other Monera—are unicellular organisms that differ from eukaryotes in nuclear and morphological characteristics and are typically of much smaller size.) This organization was based on characteristics—such as the presence or absence of a true nucleus, the simplicity or complexity of the DNA (deoxyribonucleic acid) molecules constituting the chromosomes, and the presence or absence of intracellular membranes (and of specialized organelles apart from ribosomes) in the cytoplasm—that revealed a long phylogenetic separation of the two assemblages. The concept of "protists" originally embraced all the microorganisms in the biotic world. The entire assemblage thus included the protists as defined below plus the bacteria, the latter considered at that time to be lower protists. The great evolutionary boundary between the prokaryotes and the eukaryotes, however, has meant a major taxonomic boundary restricting the protists to eukaryotic microorganisms (but occasionally including relatively macroscopic organisms) and the bacteria to prokaryotic microorganisms.

During the 1970s and '80s, attention was redirected to the problem of possible high-level systematic subdivisions within the eukaryotes. The American biologists R.H. Whittaker and Lynn Margulis, as well as others, became involved in such challenging questions. A major outcome was widespread support among botanists and zoologists for considering living organisms as constituting five separate kingdoms, four of which are placed in what may be thought of as the superkingdom Eukaryota (Protista, Plantae, Animalia, and Fungi); the fifth kingdom, Monera, constitutes the superkingdom Prokaryota.

This article discusses the kingdom Protista in general terms. For discussion of the differences and similarities among the four kingdoms of the superkingdom Eukaryota, as well as the Prokaryota, see BIOLOGICAL SCIENCES. For a generally more detailed treatment of the members of the Protista, see PROTOZOA and ALGAE.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, section 313, and the *Index*.

The article is divided into the following sections:

General features 268
Form and function 270
  Locomotion
  Respiration and nutrition
  Reproduction and life cycles
Ecology 273
Evolution and paleoprotistology 274
Classification 275
  Macrosystems of protist classification
  Diagnostic characterization
  Annotated classification
Bibliography 277

GENERAL FEATURES

The protists include all unicellular organisms not included with the prokaryotes. Protists also embrace a number of forms of syncytial (coenocytic) or multicellular composition, generally manifest as filaments, colonies, coenobia (a type of colony with a fixed number of interconnected cells embedded in a common matrix before release from the parental colony), or thalli (a leaflike, multicellular structure or body composed primarily of a single undifferentiated tissue). Not all protists are microscopic. Some groups have large species indeed; for example, among the brown algal protists some forms may reach a length of 60 metres or more. A common range in body length, however, is 5 micrometres (0.002 inch) to 2 or 3 millimetres (0.07 or 0.1 inch); some parasitic forms (*e.g.*, the malarial organisms) and a few free-living algal protists may have a diameter, or length, of only 1 micrometre. While members of many protistan groups are capable of motility, primarily by means of flagella, cilia, or pseudopodia, other groups (or certain members of the groups) may be nonmotile for most or part of the life cycle. Resting stages (spores or cysts) are common among many taxa, and modes of nutrition include photosynthesis, absorption, and ingestion. Some species exhibit both autotrophic and heterotrophic nutrition (see below *Form and function: Respiration and nutrition*). The great diversity shown in protist characteristics (Figures 1 and 2) supports the theories about the antiquity of the protists and of the ancestral role they play with respect to the other eukaryotic groups.

The architectural complexity of most protist cells is what sets them apart from the cells of plant and animal tissues. Not only are protists cells, they are also whole, complete, independent organisms, and they must compete and survive as such in the environments in which they live. Adaptations to particular habitats over eons of time have resulted in both intracellular and extracellular elaborations seldom, if ever, found at the cellular level in higher eukaryotic species. Internally, for example, complex rootlet

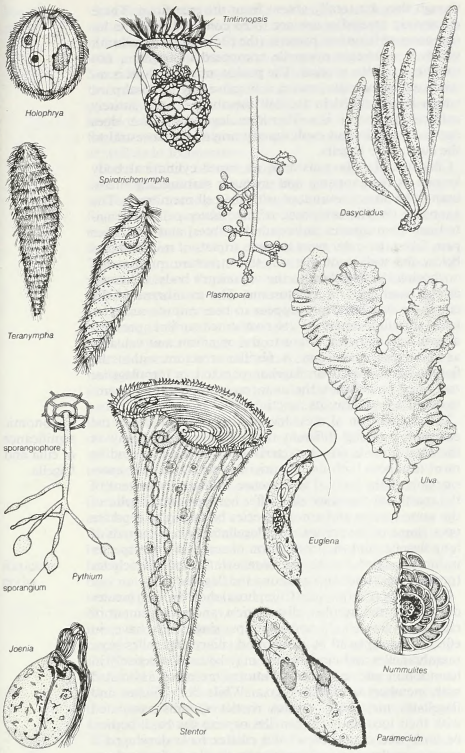


Figure 1: Representative protists.

From (*Joenia, Euglena, Nummulites, Tintinnopsis*) K.G. Grell, *Protozoology* (1973), Springer-Verlag, New York; (*Joenia*) after Hollande and Valentin, 1969, (*Euglena*) after Leedale, Meeuse, and Pringsheim, (*Nummulites*) after Brady from Kühn, (*Tintinnopsis*) after Corliss, 1961; (*Dasycladus, Ulva*) L. Margulis *et al., Handbook of Protoctista* (1990), Jones and Bartlette, Boston, (*Dasycladus*) after Taylor, (*Ulva*) after Abbot and Hollenberg; (*Plasmopara*) R.F. Scagel, *et al., Nonvascular Plants: An Evolutionary Survey* (1982), Wadsworth Publishing Co., Belmont, California; (*Haliphrya*) J.O. Corliss, "The Changing World of Ciliate Systematics," in *Systematic Zoology*, vol. 23, no. 1 (1974); (*Teranympha, Spirotrichonympha*) S.P. Parker, ed., *Synopsis and Classification of Living Organisms*, vol. 1 (1982), McGraw-Hill, Inc., New York; (*Pythium*) L. Margulis and K.V. Schwartz, *Five Kingdoms: An Illustrated Guide to the Phyla of Life on Earth* (1988), W.H. Freeman and Co., New York; (*Stentor, Paramecium*) K. Vickerman and F.E.G. Cox, *The Protozoa* (1967), John Murray, London; (*Stentor*) after Tartar

systems have evolved in association with the basal bodies, or kinetosomes (see below *Locomotion*), of many ciliates and flagellates, and nonhomologous endoskeletal and exoskeletal structures have developed in many protist taxa. Conspicuous food-storage bodies are often present, and pigment bodies apart from, or in addition to, chloroplasts are found in some species. In the cortex, just under the pellicle of some protists, extrusible bodies (extrusomes) of various types (*e.g.,* trichocysts, haptocysts, toxicysts, and mucocysts) have evolved, with presumably nonhomologous functions, some of which are still unknown. Scales may appear on the outside of the body, and, in some groups, tentacles, suckers, hooks, spines, hairs, or other anchoring devices have evolved. Many species have an external covering sheath, which is a glycopolysaccharide surface coat sometimes known as the glycocalyx. Cyst or spore walls, stalks, loricae, and shells (or tests) are also common external features.

In terms of conventional classifications of the lower eukaryotes but considering the system used in this article, the major taxa treated here as protists include the algae, the protozoa, and the so-called lower, or zoosporic (motile), fungi.

"Conventional" classification

Included are members from among the major conventional algal classes or divisions. Some are organized as phyla under the section heading chlorobionts—the chlorophytes sensu lato (chlorophyceans, charophyceans, micromonadophyceans, pleurastrophyceans, ulvophyceans) and the glaucophytes. Many of the algal heterokonts (chromophytes sensu lato) are organized under the heading chromobionts; these include the chrysophyceans, synurophyceans, haptophyceans, xanthophyceans, pedinellophyceans, bacillariophyceans (diatoms), and phaeophyceans (brown algae), as well as several other smaller algal groups. There are, in addition, four nonalgal protist taxa in the section chromobionts. Some members of the section euglenozoa—the euglenophytes (the remaining phyla being represented by the protozoan phylum Kinetoplastidea [the trypanosomatid/bodonid protozoans] and the two small protozoan phyla Pseudociliatea and Hemimastigophorea)—and all members of the sections cryptomonads (Cryptophyta), rhodophytes (Rhodophyta, or red algae), and dinozoa (Pyrrhophyta, mostly dinophyceans) are also conventionally classified as algae.

The protozoa included here are the members of the former conventional phyla (or subphyla) Sarcomastigophora, Ciliophora, Sporozoa (or Apicomplexa), Microsporidia, and Myxosporidia. The zooflagellates sensu lato and the phytoflagellates, the latter embracing the mostly motile and/or phagotrophic algal groups among those listed in the preceding paragraph, were formerly, in more conventional classifications, placed at lower taxonomic levels. The zoomastigophorans comprised not only the so-called higher zooflagellates (mostly the symbiotic forms) but also the opalinids and the proteromonadeans of the section chromobionts, the taxonomically enigmatic choanoflagellates (phylum Choanomonadea), and the lower kinetoplastideans (the trypanosomes, class Trypanosomatea, and relatives). Also included under the broad umbrella name of Sarcomastigophora were the so-called rhizopod and actinopod Sarcodina taxa of sections X and XI—although the classification presented here also includes in those sections several groups formerly considered to be fungi—*i.e.,* slime molds (Mycetozoa) and their presumed relatives.

Fungal groups that appear in the classification used in this article include the motile zoosporic groups, sometimes called the Oomycota and Hyphochytridiomycota (section chromobionts); the Chytridiomycetes, of the section chytrids (the latter group is not closely related to the former two); and Mycetozoa, or Myxomycetes, and its alleged relatives, which are found in the rhizopod sarcodine section.

The many groups of mostly microorganisms listed above are all interrelated to a degree. Not only are they not plants, animals, or fungi, but, despite the diverse characteristics they exhibit among themselves, they are evolutionarily, systematically, and taxonomically related by the common condition of being constructed solely on a cellular basis. They may show multicellularity as well as unicellularity, but never are they multitissued.

Problems in classification

It should be emphasized that the protists cannot be divided perfectly into algae, protozoa, and fungi. This is principally because certain groups have been assigned historically to more than a single one of these three categories by zoologists, botanists, and mycologists. The rationale for past taxonomic decisions at such levels has not always been based solely on the presence or absence of chloroplasts; the situation is further complicated by the reliance of some pigmented species on the consumption of nutrients from the surrounding milieu (facultative phagotrophy) rather than on photosynthesis.

The following groups considered as phyla (or divisions) in this article have been treated—although not always uniformly and under a variety of names—as both algae and protozoa (usually as phytoflagellates) in many conventional taxonomic systems: some members of Chrysophyta, Haptophyta, Xanthophyta, Pedinellophyta, Eustigmatophyta, some members of Chlorophyta, some members of Micromonadophyta, some members of Pleurastrophyta, Glaucophyta, Euglenophyta, Cryptophyta, Dinoflagellata, Choanomonadea, and some members of Proteromonadea sensu lato.

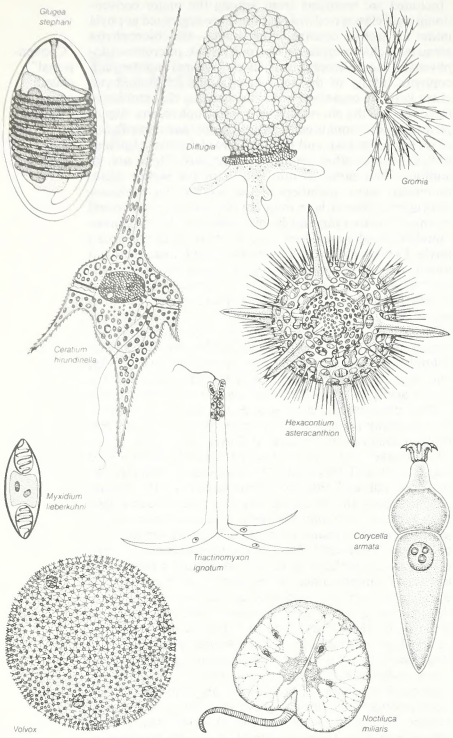


Figure 2: Representative protists.

From (*Glugea stephani*) L. Margulis and K.V. Schwartz, *Five Kingdoms: An Illustrated Guide to the Phyla of Life on Earth*, 2nd ed. (1988), W.H. Freeman and Co., New York. (*Diffugia*) K. Vickerman and F.E.G. Cox, *The Protozoa* (1967), John Murray, London. (*Gromia, Noctiluca miliaris*) J.J. Lee, S.H. Hutner, and E.C. Bovee (eds.), *An Illustrated Guide to the Protozoa* (1985), Society of Protozoologists, Lawrence, Kansas. (*Ceratium hirundinella, Hexacontium asteracanthion, Corycella armata*) K.G. Grell, *Protozoology* (1973), Springer-Verlag, New York. (*C. hirundinella*) after Lauterborn. (*H. asteracanthion*) after Haeckel from Kühn, (*C. armata*) after Léger. (*Myxidium leberkuhni, Tractnomyxon ignotum*) M.A. Sleigh, *Protozoa and Other Protists* (1989), Hodder and Stoughton, Ltd., London. (*Volvox*) G.M. Smith, *The Fresh-Water Algae of the United States*, 2nd ed. (1950), McGraw-Hill, Inc., New York.

The phyla Mycetozoa, Dictyosteliidea, Acrasidea, Plasmodiophorea, and Labyrinthomorpha have been claimed as fungi by many mycologists and as protozoa by most zoologists. The three phyla of the zoosporic "lower fungi"—the Oomycota, Hyphochytridiomycota, and Chytridiomycetes—formerly embraced only by mycologists, are now widely considered to be true protists and not fungi.

FORM AND FUNCTION

Locomotion. One of the most striking features of many protist species is the presence of some type of locomotory organelle, easily visible under the light microscope. A few forms can move by gliding or floating, although the vast majority move by means of "whips" or small "hairs" known as flagella or cilia, respectively. (These organelles give their names to informal groups—flagellates and ciliates—of protists.) A lesser number of protists employ pseudopodia. These same organelles may be used in feeding as well.

Cilia and flagella

Cilia and flagella are basically identical in structure and perhaps fundamentally in function as well. They are far more complex at the molecular level than they may seem to be when viewed solely by light microscopy. Cilia and flagella are also known among plants and animals, although they are totally absent from the true fungi. These eukaryotic organelles are not to be confused with the locomotory structure of bacteria (the prokaryotic flagellum), which is a minute organelle composed of flagellin, not tubulin, as in the protists. The prokaryotic flagellum is intrinsically nonmotile (rather, it is moved by its basal part, which is embedded in the cell membrane); it is entirely extracellular, and it is neither homologous with (*i.e.*, does not have a common evolutionary origin) nor ancestral to the eukaryotic flagella.

Cilia and flagella consist of an inner cylindrical body known as the axoneme and an outer surrounding membrane, the latter continuous with the cell membrane. The axoneme itself is composed of nine outer pairs of longitudinal microtubules (microtubular fibres) and one inner pair. The nine outer pairs become triplets of microtubules below the surface of the cell; this structure, presumably anchoring the flagellum to the organism's body, is known as the basal body or kinetosome. The membrane of the cilium or flagellum may appear to bear minute scales or hairs (mastigonemes) on its own outer surface, presumably functionally important to the organism and valuable as taxonomic characters. A fibrillar structure within the flagella, known as a paraflagellar, paraxial, or intraflagellar rod, may lie between the axoneme and the outer membrane of a flagellum; its function is not clear.

Taxonomic significance of cilia and flagella

The distribution of these locomotory organelles over the cell varies among different taxonomic groups. Many of the algal protists are characteristically biflagellate, and in most instances both flagella originate near or at the anterior pole of the body. The presence, absence, or pattern of the mastigonemes may also differ between two flagella of the same species and among species belonging to separate taxa. Some of the parasitic zooflagellates have hundreds of long flagella, and the locomotion of some of these species is further aided by the presence of attached spirochetes (prokaryotes) undulating among the flagella.

Ciliated protists (phylum Ciliophora) show an even greater diversity in the number, distribution, and arrangement of cilia over the cell. In some groups, single cilia have, in effect, been replaced by compound ciliary organelles (*e.g.*, membranelles and cirri), which may be used effectively in locomotion and in feeding. Patterns are again associated with members of different taxa. While both ciliates and flagellates may have various rootlet systems associated with their locomotory organelles or with the basal bodies, or both, the organelles in the ciliates have developed a more complex and elaborate subpellicular infrastructure. Called the infraciliature, or kinetidal system, it lies principally in the outer, or cortical, layer of the ciliate's body (only the outermost layer is called the pellicle) and serves primarily as a skeletal system for the organism. The system is composed of an array of single or paired kinetosomes with associated microtubules and microfibrils plus other specialized organelles (such as parasomal sacs, alveoli, contractile vacuole pores, and the cytoproct, or cell anus), which is unique among the protists. Variations are of great importance in the taxonomy and evolution of protists.

Typically, flagellates move through an aqueous medium by the undulatory motions of the flagella. The waves of movement are generated at the base of the flagellum. The direction and speed of propulsion and other elements of movement depend on a number of factors, including the viscosity of the medium, the size of the organism, the amplitude and length of the waves, the length and exact position of the flagella, and the kind and presence or absence of flagellar hairs. Some ciliates can move much more rapidly by virtue of having many though shorter, cilia beating in coordination with each other. The synchronized beat along the longitudinal ciliary rows produces what is known as a metachronal wave. Differences in details attest to the complexity of the overall process.

Flagella and cilia are also involved in sensory functioning, probably by means of their outer membranes which are known to contain, at the molecular level, as many as seven kinds of receptors. A variety of chemoreceptors can recognize minute changes in the medium surrounding the organism as well as cues from presumed mating partners that lead to sexual behaviour.

Pseudopodia

In comparison with flagella and cilia, pseudopodia seem rather simple. Pseudopodia are responsible for amoeboid movement, a type of locomotion particularly associated with members of the protist group traditionally called the Sarcodina. Such movement, however, is not exclusive to the amoebas; some flagellates, some sporozoa (apicomplexans), and even some cells of the other eukaryotic kingdoms demonstrate it. Pseudopodia, even more so than flagella and cilia, are widely used in phagotrophic feeding as well as in locomotion.

Three kinds of pseudopods (lobopodia, filopodia, and reticulopodia) are basically similar and are quite widespread among the rhizopod sarcodines, while the fourth type (axopodia) is totally different, more complex, and characteristic of certain specialized high-level taxa of the sarcodines under the designation actinopod sarcodines. The types, numbers, shapes, distribution, and actions of pseudopodia are important taxonomic considerations.

The lobopodium may be flattened or cylindrical (tubular). *Amoeba proteus* is probably the best-known protist possessing lobopodia. Although the mechanisms of amoeboid movement have long been a controversial topic, there is general agreement that contraction of the outer, nongranular layer of cytoplasm (the ectoplasm) causes the forward flow of the inner, granular layer of cytoplasm (the endoplasm) into the tip of a pseudopod, thus advancing the whole body of the organism. Actin and myosin microfilaments, adenosine triphosphate (ATP), calcium ions, and other factors are involved in various stages of this complex process (see PROTOZOA).

Other pseudopodia found among the rhizopod amoebas include the filopodia and the reticulopodia. The filopodia are hyaline, slender, and often branching structures in which contraction of microfilaments moves the organism's body along the substrate, even if it is bearing a relatively heavy test or shell.

Reticulopodia

Reticulopodia are fine threads that may not only branch but also anastomose to form a dense network, which is particularly useful in entrapping prey. Microtubules are involved in the mechanism of movement, and the continued migration of an entire reticulum carries the cell in the same direction. The testaceous, or shell-bearing, amoebas possess either lobopodia or filopodia, and the often economically important foraminiferans bear reticulopodia—in fact, granuloreticulopodia, giving the name to the taxonomic class in which these rhizopod amoeboid protists are placed.

The actinopod sarcodines are characterized in large measure by the axopodium, the fourth and most distinct type of pseudopodium. Axopodia are composed of an outer layer of flowing cytoplasm that surrounds a central core containing a bundle of microtubules, which are cross-linked in specific patterns among different species. The outer cytoplasm may bear extrusible organelles used in capturing prey. Retraction of an axopod is quite rapid in some forms, although not in others; reextension is generally slow in all actinopods. The modes of movement of the axopodia often differ; for example, the marine pelagic taxopod *Sticholonche* (formerly considered to be a heliozoan) have axopodia that move like oars, even rotating in basal sockets reminiscent of oarlocks.

Metabolism

Respiration and nutrition. At the cellular level, the metabolic pathways known for protists are essentially no different from those found among cells and tissues of other eukaryotes. Thus, the plastids of algal protists function like the chloroplasts of plants with respect to photosynthesis, and, when present, the mitochondria function as the site where molecules are broken down to release chemical energy, carbon dioxide, and water. The basic difference between the unicellular protists and the tissue- and organ-dependent cells of other eukaryotes lies in the fact that the former are simultaneously cells and complete organisms. Such microorganisms, then, must carry out the life-sustaining functions that are generally served by organ systems within the complex multicellular or multitissued bodies of the other eukaryotes. Many such functions in the protists are dependent on relatively elaborate architectural adaptations in the cell. Phagotrophic feeding, for example, requires more complicated processes at the protist's cellular level, where no combination of tissues and cells is available to carry out the ingestion, digestion, and egestion of particulate food matter. On the other hand, obtaining oxygen in the case of free-living, free-swimming protozoan protists is simpler than for multicellular eukaryotes because the process requires only the direct diffusion of oxygen from the surrounding medium.

Although most protists require oxygen (obligate aerobes), there are two main groups that may or must exhibit anaerobic metabolism: parasitic forms inhabiting sites without free oxygen and some bottom-dwelling (benthic) ciliates that live in the sulfide zone of certain marine and freshwater sediments. Mitochondria are not found in the cytoplasm of these anaerobes; rather, microbodies called hydrogenosomes or specialized symbiotic bacteria act as respiratory organelles.

Modes of nutrition

The major modes of nutrition are autotrophy (involving plastids, photosynthesis, and the organism's manufacture of its own nutrients from the milieu) and heterotrophy (the taking in of nutrients). Obligate autotrophy, which requires only a few inorganic materials and light energy for survival and growth, is characteristic of algal protists (*e.g., Chlamydomonas*). Heterotrophy may occur as one of at least two types: phagotrophy, which is essentially the engulfment of particulate food, and osmotrophy, the taking in of dissolved nutrients from the medium, often by the method of pinocytosis. Phagotrophic heterotrophy is seen in many ciliates that seem to require live prey as organic sources of energy, carbon, nitrogen, vitamins, and growth factors. The food of free-living phagotrophic protists ranges from other protists to bacteria to plant and animal material, living or dead. Scavengers are numerous, especially among the ciliophorans; indeed, species of some groups prefer moribund prey. Organisms that can utilize either or both autotrophy and heterotrophy are said to exhibit mixotrophy. Many dinoflagellates, for example, exhibit mixotrophy, one of the reasons they are claimed taxonomically by both botanists and zoologists.

Feeding mechanisms and their use are diverse among protists. They include the capture of living prey by the use of encircling pseudopodial extensions (in certain rhizopods), the trapping of particles of food in water currents by filters formed of specialized compound buccal organelles (in ciliates), and the simple diffusion of dissolved organic material through the cell membrane, as well as the sucking out of the cytoplasm of certain host cells (as in many parasitic protists). In the case of many symbiotic protists, methods for survival, such as the invasion of the host and transfer to fresh hosts, have developed through long associations and often the coevolution of both partners.

Cell division

Reproduction and life cycles. Cell division in protists, as in plant and animal cells, is not a simple process, although it may superficially appear to be so. The typical mode of reproduction in most of the major protistan taxa is asexual binary fission. The body of an individual protist is simply pinched into two parts or halves; the "parental" body disappears and is replaced by a pair of offspring or daughter nuclei, although the latter may need to mature somewhat to be recognizable as members of the parental species. The length of time for completion of the process of binary fission varies among groups of organisms and with environmental conditions, but it may be said to range from just a few hours in an optimal situation to many days under other circumstances. In some unicellular algal protists, reproduction occurs by fragmentation. Mitotic replications of the nuclear material presumably accompany or precede all divisions of the cytoplasm (cytokinesis) in protists.

Multiple fission also occurs among protists and is common in some parasitic species. The nucleus divides repeatedly to produce a number of daughter nuclei, which eventually become the nuclei of the progeny after repeated cellular divisions. There are several kinds of multiple fission, often correlated with phases or stages in the full life cycle of a given species. The number of offspring or filial products resulting from a multiple division (or very rapid succession of binary fissions) may vary from four to dozens or even hundreds, generally in a short period of time. Modes of such multiple fission range from budding, in which a daughter nucleus is produced and split

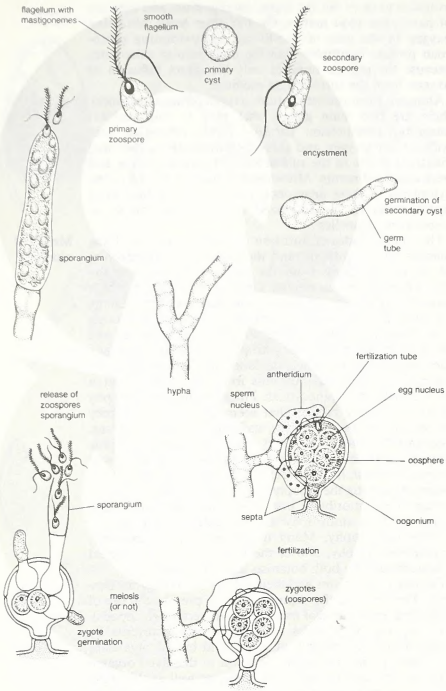


Figure 3: Life cycle of *Saprolegnia*, a lower fungal protist.
From L. Margulis and K.V. Schwartz, *Five Kingdoms: An Illustrated Guide to the Phyla of Life on Earth*, 2nd ed. (1988). W.H. Freeman and Co., New York

from the parent together with some of the surrounding cytoplasm, to sporogony (production of sporozoites by repeated divisions of a zygote) and schizogony (formation of multiple merozoites, as in malarial parasites). The latter two phenomena are characteristic of many sporozoan protists, which are obligate parasites of more advanced eukaryotes. Some multicellular algal protists reproduce via asexual spores, structures that are themselves often produced by a series of rapid fissions. The life cycles of two protist types as shown in Figures 3 and 4 illustrate many of these features of cell division.

Even under a light microscope, differences can be seen in the modes of division among diverse groups of protists. The flagellates, for example, exhibit a longitudinal, or mirror-image, type of fission (symmetrogeny). The ciliates, on the other hand, basically divide in a point-by-point correspondence of parts (homothetogeny), often seen as essentially transverse or perkinetal (across the kineties, or ciliary rows). Most amoebas sensu lato exhibit, in effect, no clear-cut body symmetry or polarity, and thus their fission is basically simpler and falls into neither of the categories described above.

Sexual phenomena are known among the protists. The erroneous view that practically all protists reproduce asexually is explained by the fact that certain well-known organisms, such as species belonging to the genera *Euglena* and *Amoeba*, do not demonstrate sexuality. Even many of the unicellular species can, under appropriate conditions, form gametes (male and female sex cells, although sometimes multiple sexes are formed, which makes the terms "male" and "female" inappropriate), which fuse and give rise to a new generation. In fact, sexual reproduction—that is, the union of one male and one female gamete (syngamy)—is the most common sexual phenomenon and occurs quite widely among the protists—for example, among various flagellate and sarcodine groups and among many parasitic phyla (*e.g.*, in *Plasmodium*, a malaria-causing organism).

Conjugation

Conjugation, the second major kind of sexual phenomenon and one occurring in the ciliated protists, has genetic and evolutionary results identical to those of syngamy. The process involves the fusion of gametic nuclei rather than independent gamete cells. A zygotic, or fusion, nucleus, not a true zygote, is produced and undergoes a series of meiotic divisions to produce a number of haploid pronuclei; all but one of these pronuclei in each organism will disintegrate. The remaining pronuclei divide mitotically; one pronucleus from each organism is exchanged, and the new micronuclei and macronuclei of the next generation are formed. Following the exchange of the pronuclei and the subsequent formation of new micronuclei and macronuclei in each organism, a series of asexual fissions, accompanied by mitotic divisions of the new diploid micronuclei, occurs in each exconjugant line. The new polyploid macronuclei are distributed passively in the first of these divisions; in subsequent fission, the macronuclei duplicate themselves through a form of mitosis. This last stage constitutes the only reproduction involved in the process.

Conjugation, as described here, is essentially limited to the ciliates, and there is considerable variation in the manner in which it is exhibited among them. For example, the two ciliates themselves may be of noticeably different size (called macroconjugants and microconjugants), or the number of predivisions of the micronuclei may vary, as may the number of nuclear divisions that take place after the zygotic nucleus is formed. Furthermore, chemical signals (gamones) are given or exchanged before a pair of protists unite in conjugation. It is not known if these gamones should be considered as sex pheromones, reminiscent of those known in many animals (for example, certain insects), but they seem to serve the similar purpose of attracting or bringing together different mating types.

While conjugation may be considered a process of reciprocal fertilization, a parallel sexual phenomenon in ciliates, which takes place in single, unpaired individuals, may be considered a process of self-fertilization. In this type of fertilization, called autogamy, complete homozygosity is obtained in the lines derived from the single parent, and the species that seem to prefer this process are known as intensive inbreeders.

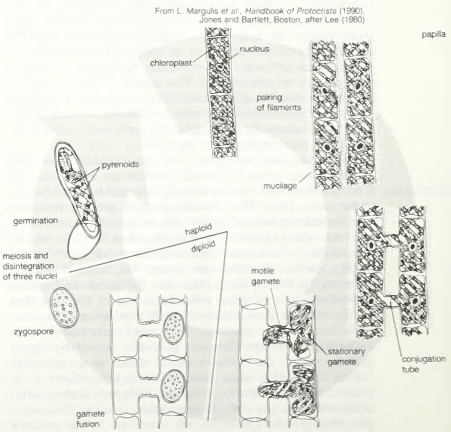


Figure 4: Life cycle of *Spirogyra*, a conjugating algal protist.

Protist life cycles range from relatively simple ones that may involve only periodic binary fissions to very complex schemes that may contain asexual and sexual phases, encystment and excystment, and—in the case of many symbiotic and parasitic forms—an alternation of hosts. In the more complicated life cycles in particular, the morphology of the organism may be strikingly different (polymorphism) from phase to phase in the entire life cycle. Among certain ciliate groups in which a larval or migratory form (known as a swarmer) is produced by the parent, the offspring may demonstrate such a differing morphology that it might well be assigned taxonomically to an entirely different family, order, or even class.

Dormancy

Dormant stages in a life cycle are probably more common in algal protists than in protozoan protists. Such stages, somewhat analogous to hibernation in mammals, serve to preserve the species during unfavourable conditions, as in times of inadequate food supply or extreme temperatures. The occurrence of resistant cysts in the vegetative stage depends, therefore, on such environmental factors as season, temperature, light, water, and nutrient supply. The fertilized egg, or zygote, in a number of algal groups may also pass into a dormant stage (a zygospore). Temporary or long-lasting cysts may occur among other protist species as well. Many sporozoa and members of other totally parasitic phyla form a highly resistant stage—for example, the oocyst of the coccidians, which may survive for a long time in the fecal material of the host or in the soil. This cyst is the infective stage for the next host in the parasite's life cycle.

Some life cycles involve not only multiple hosts but also a vector—that is, a particular metazoan organism that can act as either an active or a passive carrier of the parasite to the next host. In malaria, for example, a mosquito is required to transfer the *Plasmodium* species to the next vertebrate host (Figure 5).

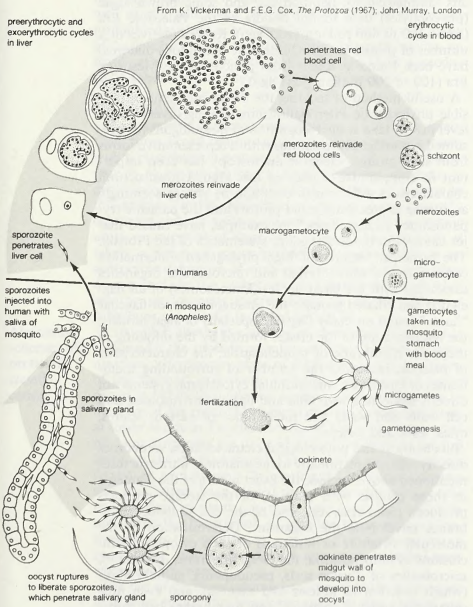


Figure 5: *Life cycle of Plasmodium vivax, a protozoan that is parasitic in the blood of humans.*
Meiosis occurs when the diploid ookinete undergoes its first nuclear division, so that all stages in the life cycle other than the ookinete (zygote) are haploid.

ECOLOGY

The distribution of protists is worldwide; as a group, these organisms are both cosmopolitan and ubiquitous. Every individual species, however, has preferred niches and microhabitats, and all protists are to some degree sensitive to changes in their surroundings. The availability of sufficient nutrients and water, as well as sunlight for photosynthetic forms, is, however, the only major factor restraining successful and heavy protist colonization of practically any habitat on Earth.

Free-living forms are particularly abundant in natural aquatic systems, such as ponds, streams, rivers, lakes, bays, seas, and oceans. Certain of these forms may occur at specific levels in the water column, or they may be bottom-dwellers (benthic). More specialized, sometimes human-made, habitats are also often well populated by both pigmented and nonpigmented members of various taxa. Such sites include thermal springs, briny pools, cave waters, snow and ice, beach sands and intertidal mud flats, bogs and marshes, swimming pools, and sewage treatment plants. Many are commonly found in various terrestrial habitats, such as soils, forest litter, desert sands, and the bark and leaves of trees. Cysts and spores may be recovered from considerable heights in the atmosphere, and some researchers claim that certain algal protists actually live, and perhaps reproduce, in air streams.

Fossilized forms are plentiful in the geologic record. They are found in strata of all ages, as far back, in the case of red alga fossils, as the Precambrian (1.9 billion years ago). Entire classes or even phyla of protists have left no record of their now extinct forms, making speculation about early phylogenetic and evolutionary relationships within the kingdom difficult to verify with the types of hard data available in the study of animal and plant evolution.

Symbiotic protists are as widespread as free-living forms, since they occur everywhere their hosts are to be found. Hundreds or even thousands of kinds of protists live as ectosymbionts or episymbionts, finding suitable niches with plants, fungi, vertebrate and invertebrate animals, or even other protists. Seldom are the hosts harmed; in fact, these often mobile substrates are actually used as a means of dispersal.

Endosymbionts include commensals, facultative parasites, and obligate parasites; the latter category embraces forms that have effects on their hosts ranging from mild discomfort to death. Protozoan and certainly nonphotosynthetic protists are implicated far more often in such associations than are algal forms. In a few protists, both cytoplasm and nuclei can be invaded by other protists, and intimate, mutually beneficial relationships between protistan hosts and protistan symbionts have been seen, such as foraminiferans or ciliates that nourish symbiotic algae in their cytoplasm. When higher eukaryotes are hosts to protists, all body cavities and organ systems are susceptible to invasion, although terrestrial plants bear relatively few such parasites. In animal hosts, the three principal areas serving as sites for endosymbiotic species are the coelom, the digestive tract and its associated organs, and the circulatory system.

Endosymbionts

The numbers of individuals in populations of many protists reach staggering figures. There are, on the average, tens of thousands of protists in a gram of arable soil, hundreds of thousands in the gut of a termite, millions in the rumen of a bovine mammal, billions in a tiny patch of floating plankton in the sea, and trillions in the bloodstream of a person infected with severe malaria. Fossil forms reach similar, if not greater, concentrations.

Some of the worst diseases of humans are caused by protists, primarily blood parasites. Malaria (caused by a protozoan protist of the phylum Sporozoa [Apicomplexa]), the various trypanosomiases (one type is African sleeping sickness) and leishmaniasis (caused by tissue-invading flagellates), toxoplasmosis (caused by another sporozoan group), and amoebic dysentery (caused by sarcodine rhizopod species) are debilitating or fatal afflictions. Biomedical research still needs to be carried out to find ways of controlling and eradicating such diseases of humans.

Protist parasites infecting domesticated livestock, poultry, hatchery fishes, and other such food sources deplete

supplies or render them unpalatable. The economic losses can be considerable. Certain free-living marine dinoflagellates are the causative agents of the so-called red tide outbreaks that occur periodically along coasts throughout the world; a toxin released by the blooming protists kills fishes in the area by the hundreds of tons. Other dinoflagellates produce a toxin that may be taken up by certain shellfish (bivalve mollusks) and which causes paralysis, even death, when the mollusk is eaten by humans. Some of the "lower" fungal protists have had significant effects on human history. One species was responsible for the great Irish potato blight of the mid-19th century, and later, another nearly ruined the entire French wine industry before a fungicide was developed to destroy it.

Many protists provide humans with benefits, some more obvious than others. Because protists are located near the bottom of the food chain in nature (just above the bacteria), they serve a crucial role in sustaining the higher eukaryotes in fresh and marine waters. In addition to directly and indirectly supplying organic molecules (such as sugars) for other organisms, the pigmented (chlorophyll-containing) algal protists produce oxygen as a by-product of photosynthesis. Algae may supply up to half of the net global oxygen. Deposits of natural gas and crude oil are derived from fossilized populations of algal protists. Much of the nutrient turnover and mineral recycling in the oceans and seas comes from the activities of the heterotrophic (nonpigmented) flagellates and the ciliates living there, species that feed on the bacteria and other primary producers present in the same milieu. Seaweeds (*e.g.,* brown algae) have long been used as fertilizers.

Benefits of protistan activities

Several hundred species of algae are consumed as food, either directly or indirectly in prepared items. For example, alginates (extracted from brown algae) and agars (from red algae) occur in such foods as ice cream, candy bars, puddings, and pie fillings.

The calcareous test, or shell, of the foraminiferans is preservable and constitutes a major component of limestone rocks. Assemblages of certain of these protists, which are abundant and usually easily recognized, are known to have been deposited during various specific periods in the Earth's geologic history. Geologists in the petroleum industry study foraminiferan species present in samples of drilled cores in order to determine the age of different strata in the Earth's crust, thus making possible the identification of rich oil deposits. Before synthetic substitutes, blackboard chalk consisted mostly of calcium carbonate derived from the scales (coccoliths) of certain algal protists and from the tests of foraminiferans. Diatoms and some ciliate species are useful as indicators of water quality and therefore of the amount of pollution in natural aquatic systems and in sewage purification plants. Selected species of parasitic protozoans from among microsporidians (Microspora) may play a significant role as biological control organisms against certain insect predators of food plants.

Uses as laboratory models

Protists have been used as model cells in laboratory research, some of which is directed against major human diseases. The combination of characteristics that has made them superior to both prokaryotic cells and other eukaryotic cells includes their easy availability and maintenance, convenient size for handling in large numbers, short generation time, broad physiological adaptability, basic structural and functional similarity to the eukaryotic cells of animal organisms, and, most importantly for sophisticated work requiring purity of material and rigidity of controls, culturability (*i.e.,* their successful growth axenically—free of other living organisms—and on chemically definable media). The culturability of some unicellular free-living protists has made them invaluable as assay organisms and pharmacological tools. Among those that have proven to be useful this way, the most important is the ciliate *Tetrahymena,* which serves as a superb model cell in investigations in cell and molecular biology. The value of such research in such biomedically important fields as cancer chemotherapy is potentially great.

EVOLUTION AND PALEOPROTISTOLOGY

Students of the evolution of most lines of plants and animals have relied heavily on the fossil records of their forms to indicate ancestor-descendant relationships over time. In the case of most protist groups, extinct forms are rare or too scattered to be of much help in evolutionary studies. For certain major taxa, fossil forms are abundant, and such material is useful in an investigation of their probable interrelationships, but only at lower taxonomic levels within those groups themselves. Speculation about the possible degrees of phylogenetic closeness among the various phyla within the entire kingdom Protista is frustrated by the lack of appropriate fossil material. There are other ways and means of determining relationships, but these are also only partially helpful. The application of modern techniques of sequencing proteins and genes to problems of evolutionary protistology is offering invaluable assistance in these investigations.

Paleoprotistology, the study of extinct protists (*i.e.,* of parts that were capable of becoming fossilized: cell, cyst, or spore walls; internal or external skeletons of appropriate preservable materials; and scales, loricae, tests, or shells) has thrown light on the probable interrelationships of both fossil and contemporary forms within classes, orders, and genera and on the paleoecology of the geologic eras and periods in which the fossil forms once lived. In addition, it has provided valuable information on the antiquity of the groups being examined. Caution is necessary, however, since species with no hard parts left no fossil record, and the extinct forms that are studied may have been preceded by species that have left specimens not yet discovered.

Paleo-protistology

The antiquity of several major groups of protists, however, has been quite well established. The rhodophytes (red algae) may have arisen as early as 1.9 billion years ago, in the Precambrian Era, although most of their fossils are from more recent geologic periods. The polycystine actinopods (classically known as the radiolarians) and various green algal protist lines also have origins in the late Precambrian (1.2 to 1.3 billion years ago). Foraminiferans, dinoflagellates, haptophytes, and some brown algae (phaeophytes) date to the middle of the Paleozoic Era (some 300 to 400 million years ago). Representatives of a number of protist taxa (including the ubiquitous diatoms) have been found as fossils from periods of the Mesozoic Era (100 to 200 million years ago).

A useful method of tackling the broad problems of possible phylogenetic interrelationships among diverse high-level protist taxa is the recognition of homologous (or presumed homologous) structures within representative forms from such groups. Electron microscopy has been important in comparative studies of this kind. Ultrastructural characteristics exhibited in common by groups seemingly as diverse as green euglenoid protists and the parasitic trypanosome "zooflagellates," for example, have caused major changes in the subkingdom systematics of the Protista. The principal features of high phylogenetic-information content are the microfibrillar and microtubular organelles associated with the basal bodies (kinetosomes) of all flagellated and ciliated protists; the mastigonemes, or flagellar "hairs," found on many flagella, especially of algal protists; the configuration of the cristae formed by the infolding of the inner membrane of mitochondria; the characteristics of plastids, including the number of surrounding membranes or envelopes; microtubular cytoskeletal systems not directly associated with cilia and flagella; extrusomes; and cell walls and walls and membranes of various spores, cysts, tests, and loricae.

Biochemical and physiological characteristics, sometimes directly related functionally to the anatomic ultrastructures mentioned above, include the exact nature of the pigments in those protists with plastids, of the storage products produced (food reserves), and of the cell walls or membranes enveloping the organism. Determination of the molecular structure or functions of such cytoplasmic inclusions as mitochondria, the Golgi apparatus, lysosomes, microbodies of diverse sorts, pseudopodia, spindle fibres (which function in mitosis and meiosis), and even miscellaneous vesicles, vacuoles, and membranes can throw light on group affinities. Comparing metabolic pathways can be valuable as well; for example, the choice of lysine biosynthesis differs among various protist taxa. Modes of nutrition are also investigated.

General ecological factors or characteristics have not played an important role in these studies. Specifically implicated in hypotheses of the origin of eukaryotic cells from prokaryotic ancestries (eukaryogenesis), however, is the phenomenon of endosymbiosis, which in a broad sense might be considered an ecological factor in the very early evolution of organisms destined to comprise the eukaryotic kingdoms. The serial endosymbiosis theory (or SET) offers one explanation of the origin of such cytoplasmic organelles as the mitochondria and plastids found in so many protists. According to SET, certain primitive prokaryotes were engulfed by other, different prokaryotes. The structures and functions of the first were ultimately incorporated into the second. The second form—now more highly evolved and presumably favoured by selection—could subsequently engulf, or be invaded by, still other types of primitive prokaryotes, acquiring from them additional, and different, structures and functions. Through its own internal evolution as well, this more complex organism eventually came to possess the characteristics recognizable as eukaryotic. This exogenous theory is to be contrasted with the endogenous hypothesis, which has held that all cellular organelles have been derived, in a long evolutionary process, from materials (especially membranes) already present in the (potential) eukaryotic cell.

The serial endosymbiosis theory

Ribosomal RNA sequencing is a molecular technique that has had a major impact on conventional schemes of classification of the protists. It has, however, also strengthened or confirmed older systems that were based either on intuitive deductions or on the determination of ultrastructural homologies.

The protists are thought to have arisen from eubacterial (not archaebacterial) prokaryotes, with symbiotic associations being involved in some way. The first, or "eoprotist," was probably a nonpigmented heterotrophic form. From within the vast array of protists there must have arisen the early members of the other eukaryotic kingdoms, as well as still additional protist groups. Numerous groups undoubtedly arose as evolutionary experiments, and many of these subsequently became extinct, generally leaving no fossil record. The protists are themselves likely someday to be subdivided into several separate kingdoms.

CLASSIFICATION

Macrosystems of protist classification. There are essentially three broad options with respect to treating protists within classification systems that embrace all living things. One is to recognize that a single kingdom, Protista, is evolutionarily and taxonomically justifiable, as is done in this article. Protists, by virtue of sharing many common characteristics, do seem to manifest an overall taxonomic unity or integrity of their own. Yet, if this approach is taken, a series of major problems remains: what is an acceptable definition of such an assemblage; exactly what does it include (*i.e.*, what are its boundaries); and what are the phylogenetic interrelationships of the high-level subgroups specifically included within it?

Alternative schemes of protistan classification

A popular alternative among evolutionary biologists is to consider the protists as only a structural grade of organization, a temporary state of transition in the evolution of the "higher" eukaryotic kingdoms from a prokaryote ancestry. While this view has appeal, it leaves confusion in its wake: if the protists belong to distinct taxonomic units at lower levels in the classificational hierarchy, then what phyla or kingdoms are to be identified for them at the top levels in the macrosystem? The fact that certain protists served as evolutionary "gap-bridgers" in eukaryogenesis and that others have played an ancestor-descendant role in the origin of plants, animals, and fungi by itself does not forbid the recognition of separate taxonomic distinctiveness for the protists as a group. Furthermore, many present-day protist taxa do not appear to have led anywhere evolutionarily.

The last of the three options proposes that there are more than four eukaryotic kingdoms and that the protists are scattered throughout them, sometimes sharing a particular kingdom with some plant, fungal, or animal groups. In this option, there is generally no specific kingdom bearing the name (or concept) Protista. For example, in the late 1980s the biochemical cytologist Tom Cavalier-Smith argued, based on his interpretation of a number of facts mostly ultrastructural in nature, that within the Eukaryota there are six kingdoms: Archezoa, Protozoa, Chromista, Plantae, Fungi, and Animalia. The organisms treated as protists in this article appear in all his kingdoms except the Animalia, although only a few are in his Fungi. The huge and diverse group of heterokonts (mostly algal protists in this article) comprise the bulk of his Chromista; all the red and green algae are placed in his kingdom Plantae. Admittedly, the green algae, especially, are closely related to plants and are likely their direct progenitor group. Cavalier-Smith's kingdom Protozoa includes the typical nonpigmented, motile, heterotrophic protists long claimed by protozoologists, but not all such protozoa are included in his kingdom bearing that name. (For example, some are distributed among several other eukaryotic kingdoms, including what he has called Chromista and, especially, Archezoa, the latter containing groups considered in this article as the phyla Metamonadea, Karyoblastea, and Microsporidia.)

A scheme of classification is an effort to set up discrete units containing a great diversity of living organisms that have been evolving gradually over hundreds of millions of years, an evolution that does not necessarily show taxonomically convenient breaks in the succession of forms. The challenge is to recognize major lines of evolution within the diverse assemblage and to organize them into named groups and ranks with minimal violation of their probable phylogenetic interrelationships. The single greatest handicap to the successful production of an ideal macrosystem for the protists is the scarcity of unambiguous data about the comparative morphology, biochemistry, and molecular biology of practically any taxon of these lower eukaryotes above the level of genus or species. Problems arise when the same group or part of a particular taxon of organisms has been treated quite differently systematically at the higher levels by workers of different scientific backgrounds or training.

Purpose of classification

Application of a protist perspective, taxonomically mixing algal, protozoan, and fungal groups to the degree required by their phylogenetic interrelationships, would mean the dropping of such groups and their formal nomenclatural designations as "Protozoa," "Algae," "Phytomastigophora," "Zoomastigophora," "Sarcomastigophora," and the like.

The phyla and the classes listed in the following working high-level classification of the kingdom Protista are themselves grouped into sections, supraphyletic assemblages given only vernacular names because they do not have an official nomenclatural rank. This is done in order to indicate, in a general way, the supposed phylogenetic closeness of some protist taxa to others. Section I, for example, contains a dozen phyla sharing basic characteristics while also showing major differences that allow them to remain separate at the high level of phylum. It may be noted that one of these phyla has been claimed taxonomically as fungi in the past; three as protozoa only; four as algae only; and four, wholly or partially, as both—simultaneously—protozoa and algae. Only one section is composed solely of algae (the one containing only the unique rhodophytes); seven, all with nonpigmented members, are purely protozoan in nature; four contain mixtures of algal and protozoan phyla; and one contains protozoan and fungal groups (as indicated by their former classifications). It is this commingling of phyla formerly assigned to widely separated assemblages of organisms that makes impossible any recognition—at a formal taxonomic level—of distinct and discrete protozoan protists, algal protists, or fungal protists.

The order or the arrangement of the 16 sections below has no particular phylogenetic significance; in fact, a number of biologists today consider the most primitive protists to be members of Sections IX and X. Neighbouring sections may sometimes be closer phylogenetically than more distant ones, but not always (particularly in view of the vast ignorance of most intersectional affinities). In some publications, dinoflagellates and ciliates are postulated as being rather closely related; but, partly in an attempt to

keep (former) algal groups close together, the dinoflagellates, in the scheme below, are in Section VI, while the ciliates form Section XVI.

Diagnostic characterization. Eukaryotic organisms possessing, at most, one tissue—tissue being an aggregation of similar cells and their products forming a definite, specialized kind of structural material—protistan species are predominantly unicellular in organization and microscopic in size. The relatively few syncytial (coenocytic), coenobial, or multicellular forms, which generally appear as filaments, colonies, coenobia, or thalli, still do not exhibit a true multitissue organization in the active (vegetative) stage. Macroscopic sizes are attained by species of a few groups (notably the brown algae). There are no truly vascular protists. All eukaryotic modes of nutrition are shown by the kingdom, with both phototrophic and heterotrophic types being common. Cysts or spores occur widely. Motility is frequently exhibited, principally via flagella, cilia, or pseudopodia; in general, motility in at least one stage of the life cycle is more common among the protists than are completely nonmotile forms. Both intracellular and extracellular elaborations (such as the organelles and the skeleton) show considerable complexity in protists. The diversity that exists among the numerous characteristics of the group supports the hypothesis that protists were ancestral to the other three eukaryotic kingdoms. For example, the distribution of the protists is ubiquitous and cosmopolitan; they show all modes of nutrition, and some species may exhibit only aerobic respiration and others only anaerobic respiration; in aerobic groups, the mitochondrial cristae are tubular, vesicular, lamellar (flattened), or discoidal; and mitotic and meiotic mechanisms and types are diverse. The total number of acceptably described species, extinct and extant, may be estimated to reach at least 120,000, with another 80,000 (mostly fossil forms) on record but of questionable validity.

Summary of characteristics

Annotated classification. In the following abbreviated classification, phyla are generally the only formal taxonomic categories presented. In selected sections, classes are also included, especially if they are an aid in relating the present classification to the older and more conventional schemes. Thus, a number of classes and many important orders, suborders, families, and so on are not mentioned at all. Some of the names used and several that are not shown here may occur at the same or lower taxonomic levels in the articles ALGAE and PROTOZOA. This does not necessarily mean that the classifications presented in these articles are contradictory. The protists are considered as a single integrated assemblage in this article, while the algal and protozoan protist types are treated in more detail in their respective articles. Differences, relatively minor though they are, between the classification presented here and those appearing in the articles ALGAE and PROTOZOA also reflect variations that arise from individual interpretations. Finally, it should be noted that "phylum" and "division" represent the same level of organization; the former is the zoological term, and the latter the botanical term.

Section I. Chromobionts (heterokonts or Chromophyta sensu lato). Predominantly golden-brown, yellow-green, and brown algae plus some lower fungal groups and 3 nonpigmented zooflagellate taxa; tubular mitochondrial cristae; pigmented moiety with chlorophylls *a, c,* and *d* and chloroplasts located within rough endoplasmic reticulum, tubular mastigonemes on anterior flagellum, and food reserves stored outside plastids; ubiquitous; more than 30,000 confirmed species described, about half of which are fossils, with a possible additional 50,000 to 70,000 recorded species.

PHYLUM CHRYSOPHYTA

PHYLUM SYNUROPHYTA

PHYLUM HAPTOPHYTA (PRYMNESIOPHYTA)

PHYLUM XANTHOPHYTA

PHYLUM PEDINELLOPHYTA

PHYLUM CHLORARACHNIOPHYTA

PHYLUM EUSTIGMATOPHYTA

PHYLUM BACILLARIOPHYTA (diatoms)

PHYLUM PHAEOPHYTA (brown algae)

PHYLUM OOMYCOTA

PHYLUM HYPHOCHYTRIDIOMYCOTA

PHYLUM PROTEROMONADEA

PHYLUM OPALINATA

Section II. Chlorobionts. Essentially the green algae; flattened mitochondrial cristae; chlorophylls *a* and *b* (except for glaucophytes); flagellates and nonflagellates; unicellular and multicellular cellulosic cell walls; starch stored within chloroplasts; flagella bear no tubular hairs; sometimes classified as plants because the ancestry of the kingdom Plantae is found in this group; 10,000 described species, only relatively few as fossils; additional desmid species may be considered questionable.

PHYLUM CHLOROPHYTA

PHYLUM CHAROPHYTA

PHYLUM MICROMONADOPHYTA

PHYLUM PLEURASTROPHYTA

PHYLUM ULVOPHYTA

PHYLUM GLAUCOPHYTA (controversial)

Section III. Euglenozoa. Discoidal mitochondrial cristae; large nuclear endosome; sheets of cortical microtubules under the pellicle; paraflagellar rods; cytochrome *c* and 5S rRNA homologies known for euglenoids and kinetoplastideans; euglenoid plastids enclosed in 3 membranes, no stored starch, and no cellulosic wall; kinetoplastideans with large DNA body in mitochondrion; approximately 1,600 acceptable species.

PHYLUM EUGLENOPHYTA

PHYLUM KINETOPLASTIDEA

Class Bodoninea

Class Trypanosomatea

PHYLUM PSEUDOCILIATEA

PHYLUM HEMIMASTIGOPHOREA

Section IV. Rhodophytes (red algae). Flattened mitochondrial cristae; no centrioles or basal bodies; no flagella; photosynthetic species with chlorophyll and accessory phycobilipigments that mask green colour; predominantly marine, filamentous forms; a few may reach lengths of 1 metre or more; 5,000 species described, 750 as fossils.

PHYLUM RHODOPHYTA

Section V. Cryptomonads. Algal protists; flattened mitochondrial cristae; chloroplasts contain chlorophylls *a* and *c* and some phycobilipigments; typically biflagellate and phagotrophic; a few species are nonpigmented; nucleomorph and ejectisomes (extrusomes) are unique to this group; approximately 200 species.

PHYLUM CRYPTOPHYTA

Section VI. Dinozoa. Predominantly biflagellates with flagella uniquely located, one essentially longitudinal and the other transverse; tubular mitochondrial cristae; photosynthetic species possess chlorophylls *a* and *c* as well as xanthophylls and carotenes; cortical alveoli present; nucleus contains condensed chromosomes; many also feed phagotrophically; of approximately 4,200 known species, half are fossil forms.

PHYLUM DINOFLAGELLATA (PYRRHOPHYTA)

Class Peridinea

Class Syndinea

Nonphotosynthetic; endosymbiotic; unique life cycles; low chromosome numbers; marine.

Section VII. Chytrids.

PHYLUM CHYTRIDIOMYCETES

Section VIII. Choanoflagellates.

PHYLUM CHOANOMONADEA

Section IX. Polymastigotes. Essentially the "higher zooflagellates"; nonpigmented; mostly endosymbiotic;

multiflagellated; mitochondria absent; hydrogenosomes, always present in cytoplasm, perform mitochondrial functions; anaerobes; unique organelles associated with the base of the flagellar apparatus; of 750–800 reported species, only 500–600 acceptable.

PHYLUM METAMONADEA
Class Retortamonadea
Class Diplomonadea
Class Oxymonadea

PHYLUM PARABASALIA
Class Trichomonadea
Class Hypermastiginea

Section X. Rhizopod sarcodines. Many possess nonaxopod pseudopodia in at least some stage of the life cycle, or a shuttle-type flow of cytoplasm is exhibited; tubular mitochondrial cristae; biflagellate stage is common in many species; pseudopodia are often employed in locomotion and holozoic feeding; some 44,000 described species, of which 85 percent are foraminiferans, with about 75 percent of the total represented by fossil forms.

PHYLUM KARYOBLASTEA

PHYLUM LOBOSEA
Rhizopod amoebas, including parasitic forms, plus amoeboflagellates and many testaceous amoebas.

PHYLUM FILOSEA

PHYLUM ACARPOMYXEA

PHYLUM GRANULORETICULOSA
Class Foraminiferidea

PHYLUM MYCETOZOA (MYXOMYCETES)
Class Protosteliidea
Class Myxogastrea

PHYLUM DICTYOSTELIIDEA

PHYLUM ACRASIDEA

PHYLUM PLASMODIOPHOREA

PHYLUM XENOPHYOPHOREA

PHYLUM LABYRINTHOMORPHA

Section XI. Actinopod sarcodines. All with axopodia; pseudopodia with microtubular cores; elaborate endoskeletal systems generally present; tubular mitochondrial cristae; complex central capsule characteristic of many; primarily marine; 11,000 to 12,000 reported species, more than half of which are extinct forms.

PHYLUM ACTINOPHRYIDEA

PHYLUM CENTROHELIDEA

PHYLUM GYMNOSPHAERIDEA

PHYLUM DESMOTHORACIDEA

PHYLUM TAXOPODA

PHYLUM ACANTHARIA

PHYLUM POLYCYSTINA

PHYLUM PHAEODARIA

Section XII. Apicomplexans. Endosymbionts, mostly true parasites; unique apical complex of specialized organelles clearly visible only under the electron microscope; spores common in most life cycles; tubular mitochondrial cristae; host organisms are terrestrial, marine, and freshwater animals; often pathogenic; approximately 5,000 species.

PHYLUM SPOROZOA (APICOMPLEXA)
Class Gregarinidea
Class Coccidea
Class Hematozoea
Class Perkinsidea

Section XIII. Microsporidia. Minute intracellular parasites primarily of insects and fishes; resistant unicellular spores characterized by a single polar filament or tube; uninucleate or binucleate amoeboid sporoplasm emerges through the eversible tube on hatching of the spore in a new host, often developing into a syncytial plasmodial stage; no plastids, mitochondria, or flagella; chitin in one of the spore walls; may be one of the most ancient of all protist assemblages; 800 described species.

PHYLUM MICROSPORIDIA (MICROSPORA)

Section XIV. Haplosporidia. Small endoparasites of cells and tissues of mostly certain marine invertebrates; spores structurally complex but without polar filaments or tubes; flagella not present; flattened mitochondrial cristae; infective sporoplasms contain unique and enigmatic haplosporosomes; about 25 described species.

PHYLUM HAPLOSPORIDIA (ASCETOSPORA)

Section XV. Myxozoa. Coelozoic or histozoic parasites of mainly cold-blooded vertebrates; one or more polar capsules within valved spores and exhibiting multinuclear plasmodial and multicellular developmental stages; polar capsules contain coiled, nonhollow polar filaments, which are not used for inoculation of sporoplasms into new hosts, but to anchor the organism in tissues to be infected; no flagella; flattened mitochondrial cristae; shell valves may be extensively drawn out and elaborately sculptured; at least 1,200 species.

PHYLUM MYXOSPORIDIA (MYXOSPORA)

PHYLUM PARAMYXIDIA

Section XVI. Ciliates. Dual nuclear apparatus; infraciliary or cortical system containing distinct microtubular and microfibrillar structures; exhibit conjugation; tubular mitochondrial cristae; pellicle contains many cilia and usually alveoli (organelles rarely found in other protists except dinoflagellates); homothetogenic binary fission; heterotrophic, mostly phagotrophic; functional, complex oral apparatuses; mostly free-living, some symbiotic; marine, soil, and freshwater habitats; classes distinguished on the basis of kinetid structure, 8,000 described species, some fossil forms.

PHYLUM CILIOPHORA
Class Karyorelictea
Class Spirotrichea (Polyhymenophorea)
Class Litostomatea
Class Prostomatea
Class Phyllopharyngea
Class Nassophorea
Class Oligohymenophorea
Class Colpodea

BIBLIOGRAPHY. Most modern works on protists are to be found in specialized biological journals, and very few books have appeared that are concerned with protists overall. To begin to understand "the protist perspective," the following publications are especially recommended for perusal: HERBERT F. COPELAND, *The Classification of Lower Organisms* (1956); JOHN O. CORLISS, "Progress in Protistology During the First Decade Following Reemergence of the Field as a Respectable Interdisciplinary Area in Modern Biological Research," *Progress in Protistology*, 1:11–64 (1986); LYNN MARGULIS *et al.* (eds.), *Handbook of Protoctista: The Structure, Cultivation, Habitats, and Life Histories of the Eukaryotic Microorganisms and Their Descendants Exclusive of Animals, Plants, and Fungi* (1990); MARK A. RAGAN and DAVID J. CHAPMAN, *A Biochemical Phylogeny of the Protists* (1978); MICHAEL A. SLEIGH, *Protozoa and Other Protists* (1989); F.J.R. TAYLOR, "Problems in the Development of an Explicit Hypothetical Phylogeny of the Lower Eukaryotes," *BioSystems*, 10(1/2):67–89 (1978); and R.H. WHITTAKER and LYNN MARGULIS, "Protist Classification and the Kingdoms of Organisms," *BioSystems*, 10(1/2):3–18 (1978).

Works on more specific topics include G.C. AINSWORTH, F.K. SPARROW, and A.S. SUSSMAN (eds.), *The Fungi: An Advanced Treatise*, 4 vol. (1965–73), especially vol. 4; CONSTANTINE J. ALEXOPOULOS and CHARLES W. MIMS, *Introductory Mycology*, 3rd ed. (1979); O. ROGER ANDERSON, *Comparative Protozoology: Ecology, Physiology, Life History* (1988); R.S.K. BARNES (ed.), *A Synoptic Classification of Living Organisms* (1984); HAROLD C. BOLD and MICHAEL J. WYNNE, *Introduction to the Algae: Structure and Reproduction*, 2nd ed. (1985); T. CAVALIER-SMITH, "The Kingdom Chromista: Origin and Systematics," *Progress in Phycological Research*, 4:309–348 (1987); JOHN O. CORLISS, "A Puddle of Protists: There's More to Life than Animals and Plants," *The Sciences*, 23(3):34–39 (May/June 1983), "The Kingdom Protista and Its 45 Phyla," *BioSystems*, 17(2):87–126 (1984), "Protista Phylogeny and Eukaryogene-

sis," *International Review of Cytology,* 100:319–370 (1987), and "The Protozoon and the Cell: A Brief Twentieth-Century Overview," *Journal of the History of Biology,* 22(2):307–324 (1989); TOM FENCHEL, *Ecology of Protozoa: The Biology of Free-living Phagotrophic Protists* (1987); JOSEPH G. GALL (ed.), *The Molecular Biology of Ciliated Protozoa* (1986); J. GRAIN, "The Cytoskeleton in Protists: Nature, Structure, and Functions," *International Review of Cytology,* 104:153–185 (1986); J.C. GREEN, B.S.C. LEADBEATER, and W.L. DIVER (eds.), *The Chromophyte Algae: Problems and Perspectives* (1989); A.M. JOHNSON, "Phylogeny and Evolution of Protozoa," *Zoological Science,* 7(Suppl.):179–188 (1990); BRYCE KENDRICK, *The Fifth Kingdom* (1985); J.P. KREIER and J.R. BAKER, *Parasitic Protozoa* (1987); JØRGEN KRISTIANSEN and ROBERT A. ANDERSEN (eds.), *Chrysophytes: Aspects and Problems* (1986); JOHANNA LAYBOURN-PARRY, *A Functional Biology of Free-living Protozoa* (1984); JOHN J. LEE, SEYMOUR H. HUTNER, and EUGENE C. BOVEE (eds.), *An Illustrated Guide to the Protozoa* (1985); N.D. LEVINE *et al.,* "A Newly Revised Classification of the Protozoa," *The Journal of Protozoology,* 27(1):37–58 (1980); D.H. LYNN and E.B. SMALL, "An Update on the Systematics of the Phylum Ciliophora Doflein, 1901: The Implications of Kinetic Diversity," *BioSystems,* 21(3/4):317–322 (1988); LYNN MARGULIS, "The Classification and Evolution of Prokaryotes and Eukaryotes," in ROBERT C. KING (ed.), *Handbook of Genetics,* vol. 1, *Bacteria, Bacteriophages, and Fungi* (1974), pp. 1–41; LYNN MARGULIS and KARLENE V. SCHWARTZ, *Five Kingdoms: An Illustrated Guide to the Phyla of Life on Earth,* 2nd ed. (1988); LINDSAY S. OLIVE, *The Mycetozoans* (1975); SYBIL P. PARKER (ed.), *Synopsis and Classification of Living Organisms,* 2 vol. (1982); D.J. PATTERSON, J. LARSEN, and JOHN O. CORLISS, "The Ecology of Heterotrophic Flagellates and Ciliates Living in Marine Sediments," *Progress in Protistology,* 3:185–277 (1989); HAYDEN N. PRITCHARD and PATRICIA T. BRADT, *Biology of Nonvascular Plants* (1984); L.J. ROTHSCHILD and P. HEYWOOD, "Protistan Phylogeny and Chloroplast Evolution: Conflicts and Congruence," *Progress in Protistology,* 2:1–68 (1987); JOHN MCNEILL SIEBURTH, *Sea Microbes* (1979); MITCHELL L. SOGIN *et al.,* "Phylogenetic Meaning of the Kingdom Concept: An Unusual Ribosomal RNA from *Giardia lamblia,*" *Science,* 243(4887):75–77 (Jan. 6, 1989); HELEN TAPPAN, *The Paleobiology of Plant Protists* (1980); and C.R. WOESE, "Bacterial Evolution, *Microbiological Reviews,* 51(2): 221–271 (1987).

(J.O.C.)

Protozoa

The protozoa are a collection of single-celled eukaryotic (*i.e.,* possessing a well-defined nucleus) organisms. As such, they are among the simplest of all living organisms. Although they comprise a subkingdom in the kingdom Protista, protozoans are not necessarily related to one another. In biological terms, they are not a natural group but simply a collection of organisms. There are more than 65,000 described species, of which over half are fossil.

Protozoa are ubiquitous in most soils and in aquatic habitats from the South to the North poles. Most are invisible to the naked eye. Many are symbionts of other organisms, and about one-third of the living species are parasites. The classification of protozoans requires regular revision because modern electron microscopy and new biochemical and genetic techniques provide an ever-increasing pool of knowledge about the relationships of various protistan species and groups, often showing previous assignments to be incorrect.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* section 313, and the *Index.*

This article is divided into the following sections:

General features 279
Natural history 280
  Size range and diversity of structure
  Distribution and abundance
  Importance
Form and function 282
  The protozoan cell
  Respiration
  Metabolism and nutrition
  Reproduction and life cycles
  Adaptations
Evolution and paleontology 286
Classification 287
  General principles
  Diagnostic features
  Annotated classification
Bibliography 288

GENERAL FEATURES

Free-living protozoan groups that inhabit soils and natural waters are extremely diverse, not only in their structure (Figure 1) but also in the manner in which they feed, reproduce, and move. Among the mainly free-living groups are the flagellates (Mastigophora). Their name derives from the whiplike structures, or flagella, that are used for movement and feeding. Each flagellate cell bears one or more of these organelles.

The flagellates exhibit the greatest diversity of nutrition among the protozoa. Many contain pigments also shared by plants, such as chlorophyll, which capture light energy during photosynthesis for the manufacture of carbohydrates and other complex nutrient substances. Thus they possess a plantlike nutrition and are called autotrophs (self-feeders). Other flagellates are colourless (*i.e.,* contain no photosynthetic pigments); they obtain their nutrients by feeding on algae, bacteria, and other protozoa. Such flagellates have an animal-like nutrition and are called heterotrophs. Some colourless flagellates have photosynthetic ecto- and endosymbionts—for example, *Oikomonas syncyanotica,* which carries cyanobacterial (blue-green algal) symbionts on its surface, and the dinoflagellate *Amphisolenia,* which contains endosymbiotic cyanobacteria. Some dinoflagellates, such as *Noctiluca* and *Gyrodinium,* may have other flagellates living within them as symbionts. Many autotrophic flagellates must also consume bacteria because photosynthesis alone is not sufficient. These flagellates and those with symbiotic algae exhibit a metabolism known as mixotrophy, in which heterotrophy and autotrophy are combined in a variety of ways and to different degrees. Thus, flagellates exhibit the complete nutritional spectrum, from totally plantlike nutrition to completely animal-like nutrition, with varying degrees of both.

In fact, nutrition is not taxonomically significant because many of the phytoflagellates (*i.e.,* the plantlike groups, or Phytomastigophorea) do not contain photosynthetic pigments but feed in a heterotrophic manner. The dinoflagellates are a good example: about one-half do not contain plant pigments, but they are classified as dinoflagellates because in every other respect they are like their coloured relatives. Moreover, even among the coloured dinoflagellates, many are mixotrophs.

While the majority of flagellates are free-living, some have evolved a parasitic way of life. These include the hemoflagellates (class Zoomastigophorea, order Kinetoplastida), so called because at some stage in their life cycle they live in the blood of a vertebrate host. Several hemoflagellates cause disease, such as sleeping sickness and Chagas' disease.

The amoebas (phylum Sarcomastigophora, subphylum Sarcodina) are a diverse group of free-living protozoa that probably evolved from a number of different primitive protozoan ancestors. While they are often regarded as the simplest of protozoans, because many resemble a blob of protoplasm with no apparent organized shape, some members of the Sarcodina are actually extremely complex. The most sophisticated are the shell-bearing foraminiferans (superclass Rhizopoda, class Granuloreticulosea). There are two categories of amoebas: the naked amoebas, which lack skeletal structures, and those that possess a skeleton or shell of some type. Members of the Sarcodina move and feed by means of protoplasmic extrusions called pseudopodia (false feet). Pseudopodia vary in both structure and number among the different species. Some possess only one leading pseudopod, while others have a complex multibranched network of pseudopodia. Like the flagellates, the sarcodines also include some parasitic species among their number. A well-known example is *Entamoeba histolytica,* which causes amoebic dysentery in humans. Amoebas

The most evolved and complex protozoans are the ciliates (phylum Ciliophora). The cell surface is covered with hundreds of hairlike structures, or cilia, arranged in ordered rows called kineties. The cilia beat in synchronized waves and thereby propel the organism through the water. Most ciliates possess a cell mouth (cytostome) through which food enters the cell. (Some flagellates also have cytostomes.) In some ciliates, the cilia around the cytostome have become specially modified into sheets called membranelles, which create a feeding current and act as a sieve to trap food particles. Other important ciliate characteristics include the possession of two types of nuclei (a large nucleus, or macronucleus, and one or more small nuclei, or micronuclei, occurring in each cell); sexual reproduction by conjugation; and asexual reproduction by binary fission in an equatorial, or transverse, plane.

A number of the protozoan phyla are exclusively parasitic, either in higher animals or, as in the phylum Labyrinthomorpha, on algae (although some members of this phylum feed saprotrophically on the surface of marine grasses and algae by secreting extracellular enzymes). The entirely parasitic phylum Apicomplexa is particularly important to humans because among its members are those

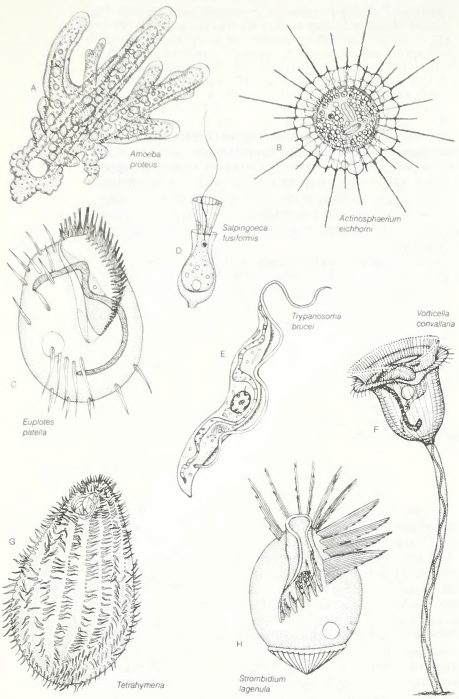


Figure 1: *Representative protozoans.*
(A) *Amoeba proteus*, a naked amoeba. (B) *Actinosphaerium eichhorni*, a heliozoan. (C) *Euplotes patella*, a hypotrich ciliate. (D) *Salpingoeca fusiformis*, a collar flagellate. (E) *Trypanosoma brucei*, a parasitic flagellate; note the undulating membrane. (F) *Vorticella convallaria*, a peritrich ciliate. (G) *Tetrahymena*, a ciliate. (H) *Strombidium lagenula*, a marine oligotrich.

From (A,B) J. Lee, S.H. Hunter, and E.C. Bovee, *An Illustrated Guide to the Protozoa* (1985), Society of Protozoologists, Lawrence, Kan., after (A) J. Leidy and (B) R. Kudo, (C,D,E,F,H) M.A. Sleigh, *Protozoa and Other Protists* (1989), Hodder and Stoughton, Ltd., London; (G) R.Y. Stanier, J.L. Ingraham, M.L. Wheelis, P.R. Painter, *THE MICROBIAL WORLD*, 5th ed., © 1986, p 535, reprinted by permission of Prentice Hall, Inc., Englewood Cliffs, N.J.

species responsible for causing such diseases as malaria and toxoplasmosis.

The major parasites causing pathological conditions in humans and other vertebrates are found in the apicomplexans and the two mainly free-living groups, the sarcomastigophorans and the ciliophorans. This fact, coupled with the importance of free-living protozoa in ecological processes, means that more is known about these three groups, and this article, therefore, concentrates on the functioning and biology of these protozoans.

NATURAL HISTORY

Size range and diversity of structure. Protozoa range in diameter from a few thousandths of a millimetre to several millimetres. Because the subkingdom contains many unrelated or loosely related groups, there is enormous diversity in structure and form. Even within a single phylum, the variation in form can be considerable.

The flagellates range from a simple oval cell with one or more flagella to the structural sophistication of the collared flagellates (order Choanoflagellida). The collared flagellates lack photosynthetic pigments and are therefore colourless. They have a single flagellum surrounded by a delicate circular collar of fine pseudopodia on which they trap food particles. In some marine species, the whole cell is enclosed in an elaborate, open latticelike basket formed from strands of silica. The dinoflagellates, half of which contain plant pigments and rely to a greater or lesser degree on photosynthesis, may be surrounded by a cell wall armour with a complicated pattern. In some species (*e.g., Ceratium*), long spines arise from the cell surface and aid in flotation. Dinoflagellates possess two flagella; one beats in a transverse plane around the equator of the cell while the other beats in a longitudinal plane. Many of the flagellates live in colonies. In *Volvox*, for example, hundreds of individual organisms are embedded in a gelatinous sphere.

The sarcodines also are extremely diverse. They have four types of pseudopodia—lobopodia, filopodia, axopodia, and reticulopodia. The simplest (lobopodia) are blunt extentions of the protoplasm, and the most complex (reticulopodia) form a complicated branching network. The simplest of the sarcodines, the naked amoebas (Gymnamoebia), have no defined shape and extend one or many pseudopodia. At the opposite extreme are the complex foraminifera, which live inside multichambered calcareous shells up to several millimetres in diameter. The pseudopodia (reticulopodia) of foraminiferans extend from the aperture of the largest chamber of the shell and form a complicated, sticky branching network. Other sarcodines, known commonly as radiolarians (class Polycystinea), form shells from silica; in some, the shell has so many holes that the structure resembles a sponge. Some of the most exquisite sarcodines are the sun protozoans, or heliozoans. Their radiating pseudopodia (axopodia), extend like spokes from the central body; microtubules support an outer layer of cytoplasm. Naked amoebas

The ciliates are the most structurally homogeneous group, although even they have evolved considerable variation on the cilia-covered cell. In some species (*e.g.*, the hypotrich *Euplotes*) the cilia are combined to form thick conical structures, called cirri, which the ciliate uses to crawl along surfaces, rather like little legs. In others the cilia virtually disappear from the main body of the cell, but the circle of cilia around the mouth becomes well developed (as in the oligotrich *Strombidium* and the tintinnid ciliates). The peritrich ciliates have developed stalks and attach to plants and animals as a means of dispersal. Many peritrichs (*e.g., Epistylis*) form branching colonies.

A group of ciliates, the suctorians, have completely lost their cilia in the adult phase. They have instead developed a stalk and many tentacles, which they use to capture passing prey, usually other ciliates. Because they cannot swim, they produce motile ciliated offspring, which settle elsewhere and then transform into the feeding stage, thus avoiding overcrowding.

Although the parasitic protozoa tend to be less structurally complex than free-living forms, considerable variation may occur during the course of their life cycles. *Plasmodium*, the malarial parasite that lives inside the liver and red blood cells of humans and the gut of its insect vector (the *Anopheles* mosquito), undergoes various changes in form through its asexual and sexual phases of development. Among the parasitic flagellates, the trypanosomes and their relatives (hemoflagellates), morphological variation occurs during the various stages of the life cycle in both the mammalian and insect host. Among species of *Leishmania*, which cause visceral leishmaniasis (kala-azar), cutaneous leishmaniasis (Oriental sore), and mucocutaneous leishmaniasis (espundia), two distinctly different forms occur. In humans, rounded, nonflagellated forms called amastigotes feed and divide inside macrophage cells in different regions of the body, while in the gut of the insect vector there occurs a flagellated form called a promastigote. Members of the genus *Trypanosoma*, which cause sleeping sickness and other diseases, have flagellated forms with different morphologies. At some stage in the life cycle, all assume the trypomastigote form—*i.e.*, slender with part of the flagellum running over the body and attached to it by a finlike extension to form an undulating membrane. They may also occur as amastigote (stumpy flagella) or promastigote forms.

Distribution and abundance. Protozoa have colonized a wide array of aquatic and terrestrial habitats from the Habitats

Arctic and Antarctic to equatorial zones. In soils and bogs, ciliates, flagellates, and amoebas form part of a complex microbial community. They live in the moisture films surrounding soil particles, so that they are actually aquatic organisms, even though living in a terrestrial environment. Between 10,000 and 100,000 organisms per gram of soil may inhabit fertile land; the relative proportions of each group vary depending on soil type and latitude. In Antarctic soils, flagellates and testate (shell-dwelling) amoebas predominate, while in temperate woodland soils, ciliates are more numerous.

In the open waters of lakes, estuaries, and the sea, protozoa form an important component of the floating plant and animal community (plankton). They are often present in densities of tens of thousands per litre of water. During photosynthesis, flagellates carrying plant pigments transform the energy from the Sun into organic matter, which, along with many algal species, forms the base of the aquatic food chain. Most planktonic protozoa, however, feed on bacteria, algae, other protozoa, and small animals. The most common planktonic protozoa are the zooflagellates, ciliates—especially members of the oligotrichs and the tintinnids (which live inside small tubes, or loricae)—and the exclusively marine foraminiferans and radiolarians.

Although few data exist for oceanic deeps, foraminiferans have been found at depths of 4,000 metres, and some protozoans have been observed around hydrothermal vents on the ocean floor.

Importance. Protozoa play important roles in the fertility of soils. By grazing on soil bacteria, they regulate bacterial populations and maintain them in a state of physiological youth—*i.e.,* in the active growing phase. This enhances the rates at which bacteria decompose dead organic matter. Protozoa also excrete nitrogen and phosphorus, in the form of ammonium and orthophosphate, as products of their metabolism, and studies have shown that the presence of protozoa in soils enhances plant growth.

Role in sewage treatment

Protozoa play important roles in wastewater treatment processes, in both activated sludge and slow percolating filter plants. In both processes, after solid wastes are removed from the sewage, the remaining liquid is mixed with the final sludge product, aerated, and oxidized by aerobic microorganisms to consume the organic wastes suspended in the fluid. In the former process, aerobic ciliates consume aerobic bacteria, which have flocculated; in the latter process, substrates are steeped in microorganisms, such as fungi, algae, and bacteria, which provide food for oxidizing protozoa. In the final stages of both processes, solids settle out of the cleaned effluent in the settlement tank. Treatment plants with no ciliates and only small numbers of amoebas and flagellates produce turbid effluents containing high levels of bacteria and suspended solids. Good-quality, clean effluents are produced in the presence of large ciliated protozoan communities because they graze voraciously on dispersed bacteria and because they have the ability to flocculate suspended particulate matter and bacteria.

Protozoa probably play a similar role in polluted natural ecosystems. Indeed, there is evidence that, by feeding on oil-degrading bacteria, they increase bacterial growth in much the same way they enhance rates of decomposition in soils, thereby speeding up the breakdown of oil spillages.

Some radiolarians and foraminiferans harbour symbiotic algae that provide their protozoan hosts with a portion of the products of photosynthesis. The protozoans reciprocate by providing shelter and carbon and essential plant nutrients. Many ciliates contain endosymbiotic algae, and one species, *Mesodinium rubrum,* has formed such a successful relationship with its red-pigmented algal symbiont that it has lost the ability to feed and relies entirely on symbiosis for its livelihood. *Mesodinium* often forms dense, nontoxic red blooms (or red tides) when it reaches high densities in plankton. Among the ciliates with endosymbionts, *Mesodinium* is the only completely photosynthetic species. Other ciliates achieve photosynthesis in another way. Although they do not have symbiotic algae, they consume plantlike flagellates, sequester the organelles that contain the plant pigments, and use them for photosynthesis. Because the isolated plastids eventually age and die, they must be replaced continuously.

The impact of protozoan grazing on phytoplankton can be considerable. It has been estimated that at least half of the phytoplankton production in marine waters is consumed by protozoa. Like the soil protozoa, these planktonic protozoans excrete nitrogen and phosphorus at high rates. The protozoans are a fundamental component in recycling essential nutrients (nitrogen and phosphorus) to the phytoplankton.

Parasitic protozoa

Parasitic protozoa have invaded and successfully established themselves in hosts from practically every animal phylum, although it is about parasitic species of medical and agricultural importance that most is known. The trypanosomes, for example, cause a number of important diseases in humans. African sleeping sickness is produced by two subspecies of *Trypanosoma brucei,* namely, *T. brucei gambiense* and *T. brucei rhodesiense.* The life cycle of *T. brucei* has two hosts, humans and other mammals and the blood-sucking tsetse fly, which transmits the parasite between humans.

Trypanosomes live in the blood plasma and the central nervous system of humans and have evolved an ingenious way of fooling the immune system of the host. Upon contact with a parasite, the immune system generates antibodies that recognize the specific chemical and physical nature of the parasite and actively neutralize it. Just as the host's immune system is beginning to win the battle against the parasite and the bulk of the population is being recognized and destroyed by host antibodies, the parasite is able to shed its glycoprotein coat, which is attached to the cell surface, and replace it with a coat containing different amino acid sequences. Thus, the parasite essentially changes its makeup. These alternate forms are known as antigenic variants, and it has been estimated that each species may have as many as 100 to 1,000 such variants. The host must produce a new set of antibodies against each new variant; in the meantime, the parasite has time in which to replenish its numbers. Ultimately, unless the disease is treated, the parasite wins the battle and the host dies. Such antigenic variation makes the development of an effective vaccine against certain parasitic protozoan diseases virtually impossible.

A close relative of *T. brucei, Trypanosoma cruzi,* causes Chagas' disease, or American trypanosomiasis. The vector hosts are bugs (*Rhodnius*) and other arthropods, such as lice and bedbugs. In humans the nonflagellated (amastigote) form of the parasite lives inside macrophage cells, the cells of the central nervous system, and muscle tissue, including the heart, where it grows and divides. Short trypomastigote flagellated forms periodically appear in the blood, where they are readily taken up by the bloodsucking vector hosts. These flagellated forms do not divide in the blood, reproduction occurring only in the amastigote intracellular forms.

Relatives of the trypanosomes, species of the genus *Leishmania,* cause a variety of diseases worldwide known as leishmaniasis. Like *T. cruzi,* these are intracellular parasites of the macrophage cells. The intermediate, or vector, hosts are a variety of sand fly species (*Phlebotominae*). In cutaneous leishmaniasis the infected macrophages remain localized at the site of the infection, causing an unsightly lesion, but in visceral leishmaniasis the infected macrophages are carried by the blood to the visceral organs. This latter disease is characterised by enlargement of the spleen and liver, leading to the distended abdomen that is typical of kala-azar. In mucocutaneous leishmaniasis the initial skin infection spreads to the mucous membranes of the face (the nose, mouth, and throat), producing a lesion that can cause the destruction of part of the face.

Plasmodium

Malaria, which is caused by the protozoan *Plasmodium,* remains a serious disease despite both measures that can be taken to control and eradicate the mosquito vector host and the availability of an array of antimalarial drugs. The life cycle is fundamentally identical among the four species of *Plasmodium,* but the pathology of the disease varies in the frequency and severity of attacks and in the occurrence of relapses. Problems in controlling the disease include the development of resistance to insecticides by

the mosquito and the evolution of drug resistance by the parasite. Prophylactic drugs taken before and during a visit to areas where malaria is endemic may prevent the disease from forming in persons who have no natural resistance. Since antigenic variation does not appear to occur in *Plasmodium*, modern genetic engineering techniques offer promise of producing a vaccine.

The apicomplexan *Cryptosporidium* (class Coccidea) is a protozoan parasite of humans and other mammals that has become particularly prominent since the 1970s. It has a one-host life cycle and lives inside the cells lining the intestines and sometimes the lungs. *Cryptosporidium* carries out all the asexual reproductive stages typical of an apicomplexan (see below) inside a single host and is passed from host to host in a resistant cyst stage called an oocyst. The disease caused by the parasite is typified by severe diarrhea and vomiting. Although there is no drug treatment, most healthy people recover quickly. In persons who have impaired immune systems, such as AIDS patients, however, *Cryptosporidium* can cause serious infections.

FORM AND FUNCTION

The protozoan cell. The protozoan cell carries out all of the processes—including feeding, growth, reproduction, excretion, and movement—necessary to sustain and propagate life, although it does so at a somewhat simpler level than do multicellular organisms. The cell is enclosed in a unit membrane called the plasma membrane. Like all membranous structures in the eukaryotic cell, the plasma membrane is composed of protein and lipid molecules. The membrane is a barrier between the cytoplasm and the outside liquid environment. Some substances, such as oxygen, readily pass through the membrane by diffusion (passive transport), while others must be transported across at the expense of energy (active transport). Cilia and flagella arising from the cell are also sheathed in the cell membrane.

Eukaryotic organelles

The cell also has internal membranes, which are not as thick as the plasma membrane. Among these are the endoplasmic reticulum, whose membranes separate out compartments of the cell, thereby allowing different conditions to be maintained in various parts—*e.g.*, separation of different substances. Enzymes are arranged on the surface of the endoplasmic reticulum; one such enzyme system catalyzes the activity of the ribosomes during protein synthesis. The Golgi apparatus is a cluster of flattened vesicles, or cisternae, associated with the endoplasmic reticulum. The vesicles are concerned with membrane maturation and the formation and storage of the products of cell synthesis, as, for example, in the formation of scales on the surface coat of some flagellates. The scales are formed within the Golgi and are transported by the vesicles to the plasma membrane, where they are incorporated onto the surface of the cell. The Golgi apparatus is well developed in flagellates, poorly seen in most ciliates, and absent from some amoebas.

All protozoa possess at least one nucleus, and many species are multinucleate. The genetic material DNA (deoxyribonucleic acid) is contained within the chromosomes of the nucleus. Each nucleus is bounded by two unit membranes through which pores provide a channel permitting the passage of molecules between the cytoplasm and the nucleoplasm. Most ciliates have two types of nuclei, micronuclei and macronuclei. Almost all protozoan cells contain at least one micronucleus, and many contain more than one. The macronucleus can be quite variable in shape, resembling in some species a string of beads or a horseshoe. It directs the normal functioning of the cell and usually disintegrates during sexual reproduction, to be reformed from the products of micronuclear division after the sexual phase is completed.

Almost all protozoa contain double-membrane mitochondria; the inner membrane forms fingerlike extensions (or cristae) into the mitochondrial interior, and the outer membrane forms the boundary of the organelle. Mitochondria are the sites of cellular respiration. Species that do not require oxygen (anaerobes), such as those that live in the intestinal tract of their hosts or those that occupy special anaerobic ecological niches, lack mitochondria. They have instead respiratory organelles called microbodies. These oblong or spherical membrane-bound organelles, about one to two micrometres in length, are believed to be the site of respiratory processes. They contain enzymes that oxidize pyruvate to acetate and carbon dioxide, resulting in the release of hydrogen sulphide under anaerobic conditions.

Photosynthetic pigments, when present, are housed in organelles called plastids. These also are bounded by at least two unit membranes. All plastids contain ribosomes and DNA of a type structurally similar to those of bacteria and other members of the kingdom Monera (prokaryotes). Ribosomes and DNA synthesize some of the protein and plastid RNAs of these protozoans independently from the synthesis carried out under the direction of the DNA of the macronucleus. Plastids are believed to have evolved from endosymbionts, and their structure and independent ability to produce proteins and RNA support this hypothesis.

Organisms that live in a liquid environment with a lower concentration of ions than is found in the interior of their cells—that is, a lower osmotic concentration—gradually gain water. If this remains unchecked, the cell swells and bursts. In protozoa the maintenance of the osmotic balance of the cell is achieved by the contractile vacuole. These membrane-bound organelles are situated close to the plasma membrane. They swell periodically and then suddenly contract and disappear, forcing their contents from the cell in repeated cycles. In the amoebas and the flagellates the contractile vacuole is formed when smaller vesicles combine with the main vacuole. In the ciliates the contractile vacuole is fed by a complex system of feeder canals, which are, in turn, fed by a complex of vesicles and fine tubules within the cytoplasm.

Contractile vacuoles

Heterotrophic protists have transitory food or digestive vacuoles. The number of these membrane-bound cell organelles depends on the feeding habits of the organism. Some species may have many, while others may contain only one or two at any one time. In ciliates the food vacuoles form at the base of the cytopharynx, while in species without a cell mouth the vacuoles form near the cell membrane at the site where food is ingested.

Within the cell, structural proteins of various types form the cytoskeleton (cell skeleton) and the locomotory appendages. They include microfilaments formed of a contractile protein also found in the muscles of animals (actin) and cylindrical microtubules formed from filaments of the protein tubulin. Microtubules are particularly important in the structural formation and functioning of cilia and flagella. Axopodia of certain flagellate species are supported by microtubules.

Locomotion

The protozoa exhibit diverse modes of locomotion across the various groups, but the modes of locomotion can be broadly divided into flagellar, ciliary, and amoeboid movement. Flagellar propulsion is employed by the flagellates and during some stages in the life cycles of certain sarcodines. The flagellum is a whiplike structure found not only in protozoans but in higher organisms as well (such as in sperm, the male reproductive cells of higher animals). The structure of all flagella is basically the same, consisting of a cylinder (axoneme) made up of a pair of central microtubules surrounded and joined by crossbridges to a circle of nine pairs of microtubules. This "nine-plus-two" arrangement of the microtubules in the axoneme is surrounded by cytoplasm and ensheathed in cell membrane. The flagellum arises from the basal body, or kinetosome, within the cell.

The undulating motion of the flagellum is normally generated at its base. The waves move along the flagellum to produce a force on the water acting along the long axis of the organelle in the direction of the wave. The speed of movement is determined by the length of the flagellum and by the size of, and distance between, the waves it generates. Some species have hairs (mastigonemes) arising at right angles to the flagellum along its length, while other species have slender hairs called flimmer filaments. Either structure has the effect of altering the movement of water produced by undulations of the flagellum by reversing its flow toward the flagellar base.

Swimming speeds achieved by flagellates are relatively

low. Ciliates have an increased number of beating flagella on the cell surface, thereby enabling greater power to be developed against viscous forces, for greater speeds. The structure of a cilium is identical to that of a flagellum, but the former is considerably shorter. Cilia are a type of flagellum arranged in closely aligned longitudinal rows called kineties. A complex system of fibres and microtubules arising from the basal bodies, or kinetosomes, of each cilium connects it to its neighbouring cilia in the kinety and to adjacent ciliary rows. In some species the body cilia may be reduced to specialized cirri, where the kinetosomes are not arranged in the usual rows but instead have a hexagonal pattern interlinked at several levels by fibres and microtubules.

The effective stroke of the cilium is usually planar, but in the recovery stroke the cilium sweeps out to the side, creating an overall beat with a three-dimensional pattern. The cilium performs work against the viscous force of the water during both the effective and the recovery strokes. To be effective, each cilium must beat in a coordinated manner with its neighbouring cilia. A synchronized beat is passed along a ciliary row by means of a hydrodynamic linkage between the cilia. During a beat, each cilium displaces a layer of surrounding water. Displaced water layers overlap between cilia and, as a consequence, interference occurs between the movements of adjacent cilia, creating a hydrodynamic linkage.

Amoeboid movement

Amoeboid movement is characteristic of the sarcodines and some of the apicomplexans. It is achieved by pseudopodia and involves the flow of cytoplasm as extensions of the organism. The process is visible under the light microscope as a movement of granules within the organism. The basic locomotory organelle is the pseudopodium; the way in which movement is effected varies.

A variety of pseudopodial types are found among the naked and testate amoebas (Gymnamoebia and Testacealobosia, respectively). In some species, a single pseudopodium is extended at any one time; in others, numerous tubular pseudopodia are extended simultaneously. Some amoebas appear saclike throughout locomotion, and no pseudopodia are obvious. The numerous long, stiff protoplasmic extensions (axopodia) of heliozoans shorten and lengthen—the forward axopodia lengthen and become attached, while the posterior axopodia detach and retract—and the amoeba rolls slowly along. The foraminiferans move by extending slender pseudopodia (filapodia), which may be several millimetres long in some species. The extending filopodia branch and fuse with each other so that there is a complex, continuously changing network of pseudopodia pulling the organism along.

Various theories have been proposed to explain how pseudopodia effect movement. A widely accepted model suggests that the ectoplasm at the front of the pseudopod contracts; isometric tension is maintained on the endoplasm, while isotonic contraction in the rear ectoplasm increases pressure on the tail endoplasm to push it forward. In addition, the contractile protein actin and the force-generating enzyme myosin—which can release the energy carried by ATP (adenosine triphosphate) and is found in the muscle contraction system of higher animals—have been isolated from the cytoplasm of amoebas. During movement, the endoplasm moves toward the cell surface at the pseudopodial tip and then gels, while actin filaments polymerize to form a longitudinal network of endoplasm that interacts with myosin- and actin-binding proteins—some of which are attached to the plasma membrane—to create a contractile cytoskeleton. This contraction increases internal hydrostatic pressure, resulting in a flow of cytoplasm toward any area where the endoplasm can extend.

Respiration. Aerobic microorganisms are so small that they are able to obtain the oxygen they require for metabolism from the surrounding liquid medium by simple diffusion. The special pigments or structures required for the acquisition and transport of oxygen found in higher organisms are not required in the protozoa. The respiratory pigment hemoglobin has been found in some ciliates (*e.g., Tetrahymena*), but it does not function as an oxygen-carrying pigment as in humans.

Most species of free-living protozoa appear to be obligate aerobes; that is, they cannot survive without oxygen. As in the cells of higher organisms, their respiration is based on the oxidation of the six-carbon glucose molecule to single-carbon carbon dioxide molecules and water via the Embden-Meyerhof pathway, tricarboxylic acid cycle (Krebs cycle), and cytochrome systems, the last two metabolic processes taking place in the mitochondria. Within a single species, the rate of oxygen consumption varies in relation to such factors as temperature, the stage in the life cycle, and the cell's nutritional status (*i.e.,* whether or not it is well fed).

Aerobic respiration

Obligate anaerobes, in which metabolism must take place in the absence of oxygen, are rarely found among eukaryotic organisms. Some parasitic anaerobic species, however, live in the gastrointestinal tract of humans and other vertebrates or, in one ecological group of ciliates (*e.g., Metopus, Plagiopyla,* and *Caenomorpha*), are associated with sulfide-containing sediments. The latter have been found to lack cytochrome activity, and both anaerobes contain microbodies rather than typical protozoan mitochondria. Along with the microbodies, the sulfur protozoa also harbour endosymbiotic and ectosymbiotic bacteria, which may take the metabolic end products released by the ciliates and reutilize them for growth and energy-yielding processes. These ciliates are believed to have reverted from an aerobic metabolism to an anaerobic metabolism in order to exploit a specialized ecological niche rich in bacteria as a food source.

The type of microbodies of the anaerobic intestine-dwelling species, which are called hydrogenosomes, function as respiratory organelles. They possess enzymes that oxidize pyruvate to acetate and carbon dioxide. Under anaerobic conditions this also results in the release of hydrogen; when oxygen is present, the hydrogen combines with the oxygen to form water.

Certain parasitic protozoa that live in the blood, such as *Trypanosoma brucei,* have evolved a system of aerobic respiration that does not involve the mitochondria. The initial stages of glycolysis in the Embden-Meyerhof pathway are the same, but glucose, rather than being broken down completely to carbon dioxide and water, is broken down only to the three-carbon molecule pyruvic acid, which is then excreted. The subsequent stages (the tricarboxylic acid cycle and the cytochrome system), which usually take place in the mitochondria, do not occur; instead, the terminal respiration is mediated by an L-α-glycerophosphate oxidase–L-α-glycerophosphate dehydrogenase system located in small membrane-bound vesicles throughout the cytoplasm.

Metabolism and nutrition. The protozoa display a range of nutritional types, from the entirely plantlike photosynthetic (or autotrophic) nutrition to the totally animal-like (or heterotrophic) nutrition, in which bacteria, algae, other protozoa, and small animals like the crustacean copepods constitute the food source (Figure 2).

Nutritional types

The coloured flagellates, or phytoflagellates (Phytomastigophorea), contain a variety of pigments that trap the Sun's radiant energy and use it to synthesize complex carbohydrates from carbon dioxide and water in the process of photosynthesis. Many coloured flagellates combine autotrophy with heterotrophy and are, strictly speaking, mixotrophs. Some members of the Euglenida, Cryptomonadida, and Volvocida, for example, are commonly called the acetate flagellates because their preferred food sources are acetates, simple fatty acids, and alcohols. In the presence of the correct nutrients, these flagellates are able to switch from carbohydrate-producing photosynthesis when light is available to heterotrophy on acetate and other substrates when it is not. Many planktonic marine and freshwater phytoflagellates also feed voraciously on bacteria. Indeed, in some lakes they may be the main consumers of bacteria suspended in the plankton. It is believed that this ingestion of bacteria (phagotrophy) provides the flagellates not only with an additional source of carbon to supplement what is gained by photosynthesis but also with phosphorus and nitrogen, which are often scarce in planktonic waters, and possibly with vitamins, all of which are essential to photosynthesis. Bacteria are

more efficient at taking up these nutrients because they have a higher surface-to-volume ratio than do the flagellates. Thus, one way for the flagellates to acquire essential nutrients is to consume the bacteria.

Heterotrophic protozoans (Zoomastigophorea, or zooflagellates) may take food into the cell at a specific point, such as the cytostome, at a particular region of the cell surface, or at any random point of entry. In the collared flagellates, for example, the collar and flagellum operate in feeding. The collar, composed of fine pseudopodia, surrounds the flagellum. The beating flagellum creates a water current, causing water to move through the collar. Particles of food in the current are trapped on the collar and are ingested by pseudopodia at its base. The ingested food is then enclosed in a membrane-bound digestive or food vacuole.

Filter feeding

(Left) Roman Vishniac, (right) Philip Feinberg, Fellow, New York Microscopical Society

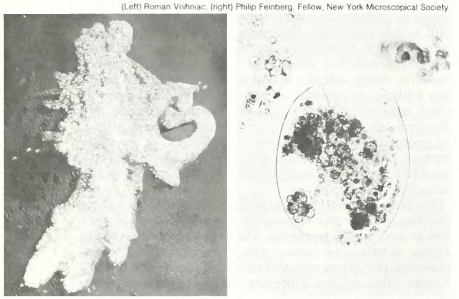


Figure 2: (Left) *Amoeba* engulfing a ciliate. (Right) *Euplotes patella* digesting algae; movement of cilia has started a whirlpool of algae at right.

Many ciliates are also filter feeders, creating water currents with special ciliary structures associated with the cytostome. The synchronized beating of these ciliary structures pushes a stream of water against a membranelle composed of cilia; the membranelle acts as a collecting sieve, where the food particles become trapped in the free spaces between the cilia. Using this mode of feeding, ciliates can shift considerable volumes of water in relation to their size. *Tetrahymena,* for example, can filter 3,000 to 30,000 times its own volume in one hour.

Other ciliates lack complex oral cilia and gather their food by other means. *Nassula* has a complex cytostome and cytopharynx supported by a basketlike cytopharyngeal structure composed of microtubules. This species ingests filamentous algae by grasping the filament, bending it like a hairpin, and drawing it into the cytopharynx, where it is broken up into fragments and enclosed in digestive vacuoles. Predatory ciliates such as *Didinium nasutum, Lacyrmaria olor,* and *Dileptus anser* apprehend their prey with special structures called extrusomes. Among the various types of extrusomes are the toxicysts, which are found in the oral region and release toxins that paralyze the prey. The suctorians are ciliate predators that usually possess tentacles of two functional types, feeding tentacles and piercing tentacles; the latter trap and immobilize the prey, usually other ciliates that make chance contact with the outstretched tentacles of the suctorian. The cell contents of the prey are transported up through the feeding tentacles into the suctorian, where digestive vacuoles are formed. The transporting mechanism is mediated by a complex array of microtubules within the tentacle. A single suctorian can often feed on several prey at the same time, and frequently the prey are larger than the predator.

The sarcodines, all of which lack a cell mouth, or cytostome, also exhibit a diverse array of feeding mechanisms and diet. Some feed on filaments of cyanobacteria (blue-green algae)—which are composed of long chains of individual cells—by taking in the entire filament at any point on the cell surface and rolling it up into a coil inside a digestive vacuole. Others, such as the testate amoeba *Pontigulasia,* pierce single cells in algal filaments and remove the contents. The radiolarians and foraminiferans trap a wide range of prey, including protozoans, algae, and small crustaceans, in their complex pseudopodial networks and then convey the food items to the main body of the cell for ingestion.

Parasitic protozoa feed in a variety of ways. Many live in the nutrient-rich medium of the body fluids—*e.g.,* the blood or cells of their host. There they take in energy-rich fluids by pinocytosis, in which small amounts of the medium are pinched off into digestive vacuoles either at a specific site, such as the cytostome in ciliates or the flagellar pocket in trypanosomes, or along the surface of the cell in amoebas. Other parasitic protozoa engulf portions of the host tissue (phagocytosis) in much the same way that free-living amoebas feed. *Plasmodium,* for example, engulfs portions of the red blood cells or liver cells in which they live. The hemoglobin in the cytoplasm of the red blood cell is only partially digested by the parasite; the protein portion of the hemoglobin molecule is degraded to its constituent amino acids, but the iron-containing portion is converted into insoluble iron-containing hemozoin, which remains within the parasite's endosomes until discarded at the next division. This process removes free hematin from the parasite cytoplasm, where it would otherwise prevent further metabolism within the parasite because it inhibits the actions of succinic dehydrogenase, an enzyme in the Krebs cycle.

Feeding in parasitic species

Whatever the mode of heterotrophic nutrition or diet, the food material is enclosed in food vacuoles, which are bounded by cell membrane. Digestive enzymes are poured into the newly formed vacuole from the surrounding cytoplasm. In the ciliate *Paramecium,* where the process has been researched in detail, it is known that the digestive vacuoles initially decrease in size and the enclosed particles aggregate. As digestion proceeds, the vacuole increases in size and the contents become progressively acidic, before gradually becoming alkaline near the end of the process. The products of digestion are then absorbed into the surrounding cytoplasm, and the waste material is ejected from the cell anus, or cytoproct. The length of the digestive cycle varies and depends on the species and the diet.

Paramecium contains a reservoir of membrane-forming material in discoid vesicles for the purpose of producing food vacuoles. The food vacuoles form at the cytopharynx when the cytopharyngeal membrane and the discoid vesicles fuse. At the cytoproct, where the vacuoles are broken down and the waste material of digestion is ejected, the membrane material is retrieved and returned to the cytopharynx. Thus, the pool of digestive vacuole membrane is continuously recycled within the cell.

While they seem to lack a sensory system, protozoans are capable of food selection. Many of the filter feeders apparently discriminate solely on the basis of size, dictated by the dimensions of the spaces in the membranelle acting as a sieve. Some filter-feeding ciliates, such as the tintinnids, however, are known to be selective and appear to be able to capture or reject items that arrive at the feeding membranelles in the feeding current. The large ciliate *Stentor,* for example, takes ciliates in preference to flagellates and algae, and discrimination increases as the animal becomes less hungry. Carnivorous species exercise distinct selectivity. Most suctorians feed exclusively on particular ciliate taxa. They are selective feeders and usually do not capture flagellates, sarcodines, or their own ciliated swarmers. Evidence suggests that a reaction between chemical compounds on the surface of the prey and the tentacle tip of the suctorian is responsible for feeding selectivity. Sarcodines also display feeding selectivity. *Amoeba proteus,* for example, selects the flagellate *Chilomonas paramecium* in preference to *Monas punctum,* even when the number of *Monas* in the medium is high. In this case, selection may be based on the digestibility of the prey; the latter is digested in 3½ hours, the former in 3 to 18 minutes.

Mixotrophy is a common phenomenon among free-living ciliates and sarcodines. Moreover, the degree of mixotrophy varies from complete reliance on the symbiotic alga or algae to transitory retention of the plastids of phytoflagellate prey with only a partial dependence on

photosynthesis to supplement the cell's energy balance. Some phytoflagellates (e.g. *Dinobryon* and *Ochromonas*), which are primarily autotrophic, also feed on bacteria and are consequently mixotrophic, but this represents a different kind of mixotrophy from that practiced by the fundamentally heterotrophic ciliates and sarcodines.

Symbiotic algae

Many of the foraminiferans and radiolarians possess symbiotic algae. In some foraminiferans and radiolarians several different symbiotic species of algae may live within the protozoan cytoplasm. During the day, the endosymbionts are distributed in the pseudopodial network, but at night they are withdrawn close to the main body of the cell or into the shell. Many thousands of these algae may exist within a single protozoan, and a significant amount of the products of photosynthesis (*e.g.*, glucose, alanine, maltose) are transferred from the algae to the protozoan. Indeed, in some circumstances, the protozoan can survive on this source of energy if deprived of food, although its growth may be impaired.

In another form of mixotrophy, the sarcodines and ciliates sequester the plastids of their phytoflagellate prey and use them for photosynthesis. The plastids do not replicate inside the protozoan as they do in the symbiotic algae and must be replaced continuously. The large marine ciliate *Tontonia appendiculariformis*, for example, may contain thousands of plastids that have been derived from a variety of flagellates; moreover, the ciliate appears to be selective in its choice of prey from which to derive plastids.

Reproduction and life cycles. Asexual reproduction is the most common means of replication by protozoans (Figure 3). The ability to undergo a sexual phase is confined to the ciliates, the apicomplexans, and restricted taxa among the flagellates and sarcodines. Moreover, sexual reproduction does not always result in an immediate increase in numbers but may simply be a means of exchanging genetic material between individuals of the same species (conjugation). Free-living protozoans normally only resort to sexual reproduction when environmental conditions become adverse, because this mode of reproduction enhances the fitness of the population and increases the chance of mutation. When food and other conditions are favourable, asexual reproduction is practiced.

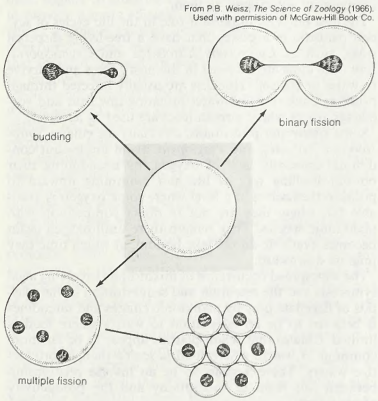


Figure 3: Asexual reproductive patterns in protozoans.

Asexual reproduction in free-living species usually involves nuclear division and the division of the cell into two identical daughter cells of equal size by binary fission. In parasitic protozoa and some free-living species, multiple fission, resulting in the production of many offspring that may not resemble the parent cell, is normal. During the cycle of growth and division, the protozoan undergoes a series of identifiable phases: a division phase, a growth phase during which the cell increases substantially in size, a phase of DNA synthesis, and a phase of preparation for division, which extends from the end of DNA synthesis until the initiation of division. The division of the cytoplasm is preceded by the division of the nucleus or nuclei.

Plane of division

The plane of division in protozoan cells varies among the different groups and is of taxonomic significance. The flagellates normally divide in a longitudinal plane. The usual process starts at the front end with the division of the flagella and the associated structures; simultaneously, the nucleus divides. The cytoplasm then splits from front to back into two identical daughter cells. The ciliates normally divide in an equatorial, or transverse, plane, thereby maintaining the correct number of ciliary rows, or kineties. The cell mouth and any specialized cilia around it are replicated in different ways among the various ciliate groups, depending on the complexity of the cytostome. The replication of the cytostome precedes the division of the cytoplasm. Some ciliates (*e.g., Colpoda*) divide within thin-walled reproductive cysts into two daughter ciliates, each of which then divides so that the cyst contains four progeny, which are released when the cyst wall ruptures.

The sedentary suctorians do not reproduce by binary fission because the production of an identical, nonswimming offspring would rapidly lead to overcrowding. They instead produce single ciliated offspring called swarmers by a process called budding. Budding can occur endogenously, in which the bud forms within the parent and is ejected when mature, or exogenously, in which the swarmer is formed outside the parent. The swarmers swim away from the parent, settle on a substrate, lose their cilia, and develop feeding tentacles and an attaching stalk.

Naked amoebas (rhizopods) have no fixed plane of division but simply round up and divide into two basically equal halves. The testate amoebas (also rhizopods), which live in single-chambered shells, or tests, exude the daughter from the aperture of the shell. In species that have a shell formed from silica plates, the daughter contains the plates used to produce the shell but remains attached to the mother cell until the shell is fully formed, when the final severing of the cytoplasm between the individuals occurs. Some of the testate amoebas live inside proteinaceous shells. There, too, the new shell is secreted before binary fission is completed.

The foraminiferan and radiolarian sarcodines have evolved multiple fission. Both produce many flagellated swarmers, or zoospores. The common planktonic foraminiferan *Globigerinoides sacculifer*, for example, can produce 30,000 swarmers at one time. Each swarmer is about 5 micrometres (0.005 millimetre) long. In planktonic species the parent usually loses buoyancy and sinks by shedding spines and withdrawing the complicated pseudopodial network into the shell. The swarmers are produced in deep water and migrate upward as they mature. Each secretes a shell around itself, which is added to as the organism grows.

The foraminiferans are unusual among free-living protozoans in that a sexual phase is a regular part of the life cycle, alternating with an asexual phase. During the life cycle two types of swarmer are produced. One type, zoospores, have half the number of chromosomes of the parent (*i.e.*, they are haploid); they grow until they become mature adults and can produce and release large numbers of gametic swarmers. These gametes are identical (isogamous) but are comparable to the eggs and sperm of higher organisms. The gametic swarmers fuse in pairs, thus restoring the full complement of chromosomes (*i.e.*, they are diploid), and each individual grows, matures, and ultimately produces haploid zoospores.

Sexual reproduction among the flagellates is not widespread and can involve identical gametes (isogamy) or distinct male and female gametes (anisogamy). The female gametes are larger and are stationary, whereas the male gametes are smaller, produced in larger numbers, and motile.

Conjugation

Sexual reproduction among the ciliated protozoans takes the form of conjugation. The process does not result in an increase in numbers, but is a simple exchange of genetic material between two individual cells. Conjugation occurs only between compatible mating strains within a species, and each species may contain many mating strains. Before

conjugation occurs, special chemical signals, called gamones, are released by some ciliates. The gamones cause compatible mating strains to undergo processes that facilitate conjugation. In other ciliates, such as *Paramecium,* gamones are bound to the cell surface and elicit their responses when the ciliates make physical contact.

During conjugation, two ciliates line up side by side. The macronucleus, which plays no part in the process, disintegrates. A series of nuclear divisions of the micronuclei in each ciliate then ensues, including a meiosis, during which a number of haploid micronuclei are produced in both cells. All but one of these haploid micronuclei disintegrate. The remaining haploid micronucleus in each cell then divides through mitosis into two haploid nuclei (gamete nuclei). A bridge of cytoplasm forms between the two ciliates, and one gametic nucleus from each cell passes into the other cell. The two gametic nuclei in each cell unite, thus restoring the diploid number of chromosomes. The micronucleus undergoes two mitotic divisions to produce four micronuclei; two of these will form the new micronuclei of the cell and two are destined to become the macronucleus. Following the process of conjugation, normal binary fission proceeds. The number of macronuclei and micronuclei formed is dependent on the species and remains the same as the original number.

Modified conjugation

When no suitable mating partner is available, ciliates may undergo a form of conjugation called autogamy, in which all of the nuclear processes described above occur. But, because only one individual is involved, there is no exchange of gametic nuclei; instead, the two gametic nuclei within the cell unite to form the restored micronucleus.

Specialized sedentary suctorian ciliates practice a modified form of conjugation. The conjugating individuals differ in appearance. The macroconjugants resemble the normal feeding individuals, and the microconjugants resemble the swarmers, although smaller. When a microconjugant locates a macroconjugant, it enters and fuses with it. This is quite different from the temporary association between two cells that occurs in most ciliates.

As is common with other parasitic organisms, parasitic protozoans face the problem of how to disperse from one host to another. In order to increase the probability of finding more hosts, most parasitic protozoa reproduce in high numbers. A representative life cycle of a parasitic protozoan can be found in members of the parasitic phylum Apicomplexa. These protozoans have a complex life cycle that involves a series of stages characterized by episodes of asexual multiple division called schizogony. In the parasite *Plasmodium,* for example, this phase of the life cycle occurs in the liver and red blood cells of humans. The parasite (sporozoite) enters the host's cells and grows while feeding on the cell contents. It then undergoes a multiple asexual division (schizogony) into many individuals (merozoites). The host's cell wall ruptures, permitting each individual to invade a new red blood cell and repeat the process.

In certain merozoites a sexual cycle is eventually initiated inside the red blood cell, and male and female gametes are produced. The male gametes (microgametocytes) are small, while the female gametes (macrogametocytes) are larger. The life cycle continues if the gametocytes are taken up by a feeding female mosquito of the genus *Anopheles.* Only the gametocytes can infect the mosquito. Inside the mosquito's gut the haploid gametes fuse to form a diploid zygote, which then undergoes sporogony, a process of multiple divisions in which many sporozoites are produced. The sporozoites migrate to the salivary glands of the insect and are injected into a new host when the mosquito next feeds. They are carried by the blood to the liver, where they undergo their first schizogony inside liver cells, thereafter invading the red blood cells for repeated cycles of schizogony.

The parasitic flagellates reproduce entirely by asexual means and do not appear to have a sexual phase in their life cycles. There is, however, evidence of genetic exchange between certain subspecies of *Trypanosoma brucei,* although the process by which this occurs is not known.

Adaptations. For the most part, parasitic protozoans live in a fairly constant environment. Temperature fluctuates very little, or not at all, inside the host, there is no risk of desiccation, and food is in constant supply. Free-living protists, on the other hand, face short- or long-term changes in temperature, acidity of the water, food supply, moisture, and light. Many protozoa respond to adverse environmental conditions by encysting. They secrete a thick, tough wall around themselves and effectively enter a quiescent state comparable to hibernation. The ability to form a resistant cyst is widespread among diverse protozoan groups and probably developed early in their evolutionary history. Resting cysts also are easily carried by the wind and form an important means of dispersal for species that live in the soil or are common in temporary ponds and pools. In climates with distinct cold seasons, the cyst may be an important phase in the annual life cycle.

The cyst wall is composed of a varying number of layers, the components of which are dependent on the species. During the encystment process, the protozoan cell undergoes a series of changes that considerably reduce the complexity of the organism. Flagellates and ciliates lose their flagella and cilia, the contractile vacuole and food vacuoles disappear, and the distribution of organelles within the cell may be reorganized. In some species, the cell volume reduces considerably. These changes are reversed during the process of excystment.

Encystment

Certain of the tintinnid ciliates that live in the plankton of seas are programmed to break out of their cysts en masse at times of the year when the food supply is abundant. *Helicostomella subulata,* for example, excysts in June in temperate waters and becomes numerous from July through October. It encysts again in October, sinking to the sediments, where it remains until the following year. The cyst is a normal part of the annual life cycle, and even laboratory populations of this ciliate encyst at the same time as the natural population. This type of life strategy pattern has been demonstrated in several other ciliates and in sarcodines.

For soil-dwelling protozoa, the cyst is an important refuge when soil moisture disappears or when soil water becomes frozen. In soils that are subject to freezing and periodic short-term thawing, the protozoa rapidly excyst, feed, and reproduce and then encyst again when soil water becomes temporarily unavailable to them.

The cyst plays an important role in the life cycles of several parasitic protozoans that have a free-living dispersal stage, such as *Entamoeba histolytica* and *Cryptosporidium.* The cysts are excreted in the host's feces and survive in water or the soil. Humans are usually infected through drinking contaminated water or eating raw fruit and vegetables grown where human feces are used as fertilizer.

Some freshwater protozoans, especially the ciliates *Spirostomum, Loxodes,* and *Plagiopyla,* avoid unpleasant conditions, especially lack of oxygen, by abandoning their bottom-dwelling way of life and swimming upward to position themselves at a level where some oxygen is available but where they are not in direct competition with planktonic species. They remain there until oxygen again becomes available on the lake bottom, at which time they migrate downward.

The widespread occurrence of mixotrophy involving algal symbiosis and the retention and sequestration of the plastids of flagellate prey by planktonic ciliates and sarcodines is believed to be an adaptation to waters where food is limited. Ciliates that retain plastids appear to be far more common in waters where food is scarce than in productive waters. There appears to be an inverse relationship between this form of mixotrophy and the productivity of the waters.

Algal symbionts

EVOLUTION AND PALEONTOLOGY

Protists dominated life on Earth 1.5 billion years ago, giving rise eventually to multicellular organisms. While protozoa evolved early and have survived to the present day as unicellular organisms, they have undoubtedly undergone considerable evolutionary change. That many species must have become extinct as others appeared can be deduced from the limited fossil record. Extinct fossil foraminiferan species, for example, number around 34,000, while there are only about 4,000 described living species.

Only a small number of protozoans have left fossil remains. The calcareous shells of the sarcodine foraminiferans and calcium-secreting flagellate coccolithophores, for example, have produced substantial geologic strata in the chalk formed during the Cretaceous Period (144 to 66.4 million years ago) and the well-developed foram-limestones of the Upper Paleozoic Era (570 to 245 million years ago), Early Cretaceous Period (144 to 97.5 million years ago), and Cenozoic Era (66.4 million years ago to the present). Another fossil-forming group includes the radiolarians, which date to late Precambrian times.

The most abundant and important fossil protozoans are the foraminiferans. This entirely marine group is extremely important as stratigraphic markers in oil exploration. Because species have appeared and then become extinct frequently during geologic history and because they have fairly wide geographic distribution, particularly planktonic species, their value is in showing distinct phases in geologic history and, with specific species, in typifying particular beds of rock or strata.

The poor fossil record of protozoans has hampered attempts at unraveling the complexities of their evolution. Modern biochemical and electron microscopy techniques, however, are providing evidence for new affinities between groups and are elucidating possible evolutionary pathways. Comparisons of flagellar structures, mitochondria, and nuclear and plastid characteristics in conjunction with ribosomal RNA (ribonucleic acid) sequences are revealing the relationships of various taxa.

The main protozoan groups seen today—the ciliates, flagellates, sarcodines, and apicomplexans—almost certainly separated early in geologic history from amoeboid and flagellate ancestors and represent independent ancestral branches of the protozoan evolutionary tree. The ancestral eukaryote organism is thought to have been an amoeboid creature that relied on anaerobic metabolism. The evolution of mitchondria (the centres of aerobic respiration in the cell) as organelles from endosymbiotic bacteria and the establishment of oxidative pathways allowed a more efficient cellular energy balance, which led the way to the evolution of an enormously diverse array of eukaryotic organisms. Some of the early amoeboid eukaryotes developed flagella to enhance their food-gathering abilities and to provide a more efficient mode of propulsion. The flagellates gradually evolved different ways of life, and their structures became modified accordingly. As phagotrophs that ingested bacteria for food, they came in some cases to establish symbiotic associations with photosynthetic species, and ultimately the endosymbionts became plastids within the cell. Some of the flagellates came to depend entirely on photosynthesis and to abandon heterotrophy completely, though many still retain both heterotrophic and autotrophic nutrition as mixotrophs. (Some present-day mixotrophs, however, may be only secondarily mixotrophic, having reestablished heterotrophy in conjunction with photosynthesis.)

Development of flagella

A considerable number of protozoans became parasitic, a mode of life that evolved independently among the protozoa many times. Ciliates and sarcodines became symbionts in the intestinal tracts of both vertebrates and invertebrates as a result of surviving the digestive enzymes of the predator. (Most present-day parasites among these protists are intestinal parasites.) Once inside the intestine of the host, they multiplied and gradually, through mutation and selection, came to rely on the resistant cyst as a means of survival and dispersal, losing the ability to survive in a free-living feeding form. Some species subsequently started to invade the tissues of their host (*e.g., Entamoeba histolytica*), resulting in disease.

The process of parasitism probably arose several times among the parasitic flagellates. The trypanosomes, for example, evolved from free-living zooflagellates, which adapted to living in the alimentary canal of primitive invertebrates in Precambrian times (3.8 billion to 570 million years ago). They evolved with their hosts, becoming symbionts in a wide variety of invertebrates, including annelids, nematodes, and mollusks. It was in the insects, however, that they underwent their most extensive evolutionary explosion into two stocks. At this stage they were transmitted from insect to insect by resistant cysts passed in the feces and ingested by subsequent hosts. When insects developed the habit of sucking vertebrate blood as an energy source, which is believed to have occurred about 40 million years ago, the protozoan symbionts that lived in the gut entered the blood of vertebrates, probably as feces left by the insect were rubbed into the wound. The blood provided a rich environment for the flagellates and thus evolved the two-host life cycles seen today in the *Leishmania* and *Trypanosoma* groups.

Two stocks of parasites

The apicomplexans, which also inhabit the blood of vertebrates at some stage in their life cycle, probably evolved from a basal primitive stock seen today as the gregarines, which are parasites of invertebrates. They gave rise to a group of parasitic organisms of which the Coccidea, with a one-host life cycle, are primitive survivors. At first these protozoans lived in the guts of their vertebrate host, but they gradually began invading host tissues and eventually became adapted to spending part of their life cycle in the bloodstream. There they were taken up by blood-feeding insects, and an insect vector host became incorporated into the life cycle. Associated modifications in the reproductive pattern, as seen in *Plasmodium*, which belongs to the Haemosporina, also occurred. This series of events appears to have happened at least twice in the evolution of apicomplexan life cycles.

CLASSIFICATION

General principles. Until the 1970s the general view was that the protozoans were animals, and as such they were placed in the animal kingdom as the phylum Protozoa. Under this system of classification, zoologists placed the coloured flagellates in the phylum Protozoa despite their obvious plant affinities. Botanists, on the other hand, classified the same organisms as algae, which were regarded as plants. Protozoans are now regarded as a phylum or subkingdom of the kingdom Protista (sometimes called the Protoctista), which also includes the algae. The Society of Protozoologists, which periodically reviews the systematics of the group, favours the subkingdom level for the Protozoa.

Protozoan systematics remains a subject of debate and change. Protozoans comprise a large, unwieldy assemblage, and assignments of species to particular taxa change as new biochemical techniques and electron microscopy studies provide more details on the affinities of various species. The subkingdom Protozoa essentially represents a level of organization—that is, single-celled eukaryote organisms—and does not necessarily indicate single major branches of an evolutionary tree comparable with the kingdoms Plantae or Animalia. Thus, while affinities between some of the groups exist, this is not the case with all the phyla within the subkingdom. Moreover, certain groups (*e.g.*, the Sarcodina or amoebas) are considered by some researchers to have had many different ancestors and lineages (polyphyletic). The Labyrinthomorpha, Myxozoa, and Microspora have no clear affinities either with one another or with other protozoan groups; they are placed in Protozoa until more information is available.

Diagnostic features. The subkingdom Protozoa has been reviewed and divided by a committee of the Society of Protozoologists into six phyla. A major review was undertaken in 1980 by a committee headed by N.D. Levine, and a series of changes have been made subsequently as outlined in a major work on protozoan systematics published by the society in 1985. The six phyla include the Sarcomastigophora (flagellates and amoebas), the Ciliophora (ciliates), and the entirely parasitic Apicomplexa, all of which are discussed in this article. At least some of these groups show affinities in having common amoeboid or flagellate ancestors early in their evolutionary history. The remaining three phyla are totally unrelated to one another and to the phyla discussed here.

In this assemblage of organisms, the only common feature is a single-celled level of organization. Such a situation invariably means that within the whole group there are considerable differences in structure physiology, life mode, and life cycles.

In addition, the classification of protists generally contin-

ues to be debated, and a standard outline of the kingdom has not been established. The differences between the classification of protozoa given below and that given in the article PROTISTS reflect taxonomic variations that arise from individual interpretations.

Annotated classification.

PHYLUM SARCOMASTIGOPHORA

All have one type of nucleus, though some may be multinucleate; Foraminiferida may show nuclear dimorphism similar to that seen in ciliates; sexual reproduction is not widespread; flagella or pseudopodia, and in some species both, occur at one or another stage in the life cycle.

Subphylum Mastigophora

At least 1, and often many, flagella as feeding and locomotory organelles; binary fission resulting in 2 identical daughter cells is the usual mode of asexual reproduction (symmetrogenic binary fission); sexual reproduction occurs in some groups.

Class Phytomastigophorea. Typically possess chloroplasts, with a range of photosynthetic pigments; some members lack chloroplasts but have an evident relationship to the pigmented forms; all free-living, solitary or colonial; typical species include the green species *Volvox* (which belongs to the order Volvocida) and *Euglena* (of the order Euglenida) and the colourless species *Peranema* (Euglenida) and *Noctiluca* (a dinoflagellate from the order Dinoflagellida).

Class Zoomastigophorea. Colourless flagellates; heterotrophic nutrition; binary fission; some groups, such as the collared flagellates, produce flagellated swarmers; both free-living and parasitic species; includes many free-living groups, among which are the collared flagellates (Choanoflagellida) and the amoeboflagellates, having both amoeboid and flagellated forms in the life cycle (Cercomonadida), and various parasitic orders and families; includes *Trypanosoma brucei gambiense* and *Trypanosoma cruzi* (both of the order Kinetoplastida).

Subphylum Opalinata

Previously classified with the ciliates because of their outward appearance in being covered with rows of cilia; their kinetosomes differ from those of the ciliates; they have only 1 type of nucleus, whereas the ciliates have 2 nuclear types; they lack a cytostome; most divide by symmetrogenic binary fission; asexual and sexual phases of reproduction alternate in the life cycle; all members commensals in the alimentary tracts of amphibians.

Subphylum Sarcodina

Not necessarily closely related; pseudopodia of varying types for locomotion and feeding; most are free-living and divide by binary fission; sexual reproduction is not widespread but does occur in some groups, such as the foraminiferans; some groups are naked; others have shells, or tests, formed from calcium, silica, or protein; others have a structured internal cytoskeleton.

Superclass Rhizopoda

Includes all species that use different types of pseudopodia for movement (filopodia, lobopodia, and reticulopodia), such as naked amoebas (Gymnamoebia), testate forms (Testacealobosia), the foraminiferans (Granuloreticulosea), and some parasitic forms, but not forms that use axopodia.

Superclass Actinopoda

Radial stiffened axopods with a central rod composed of microtubules; many have mineral material incorporated into the cytoskeleton; mostly marine, many planktonic; includes Polycystinea (radiolarians, *e.g., Hexacontium*), Heliozoea (the heliozoans, or sun protozoans, *e.g., Actinosphaerium eichhorni*), Acantharea (*e.g., Acanthochiasma rubescens*), and Phaeodarea (*e.g., Planktonella atlantica*).

PHYLUM LABYRINTHOMORPHA

An ectoplasmic network of spindle-shaped or spherical non-amoeboid cells; in some, amoeboid cells move within the network by gliding, using a unique cell-surface organelle associated with the ectoplasmic network; produce zoospores; saprozoic or parasitic on algae; mostly in marine and brackish waters.

PHYLUM APICOMPLEXA

Previously called Sporozoa; parasitic in both vertebrates and invertebrates; vesicular nucleus; lack flagella or cilia, except in the flagellated microgamete stage; usually a sexual phase in the life cycle with male and female gametes; schizogony and sporogony are features of the life cycle, and cysts are often present at some stage in species with one-host cycles; typical species include *Plasmodium falciparum, Toxoplasma gondii, Cryptosporidium,* and *Eimeria* species.

PHYLUM MICROSPORA

Not analogous with any other protozoan group; include intracellular parasites of a wide variety of animals; produce spores at some stage in their life cycle, and all active stages develop in the host's tissues, specifically in the cytoplasm; invasive stage is the sporoplasm, a tiny uninucleate or binucleate structure only 1 micrometre in diameter; both schizogony and sporogony occur in the life cycle; a typical species is *Pleistophora typicalis.*

PHYLUM MYXOZOA

Parasites of the tissues and organ cavities of cold-blooded vertebrates, especially fish, and annelid worms; spore producers; typical species is *Myxidium giardi.*

PHYLUM CILIOPHORA

Each cell usually has at least 1 macronucleus and 1 or more micronuclei; most have a cytostome; all cilia arranged in an ordered fashion at some time in the life cycle; divide by homothetogenic binary fission, but some sedentary forms produce swarmers during asexual reproduction; sexual reproduction is universal in ciliates and normally involves conjugation between 2 cells; more than 7,000 described species, of which the majority are free-living.

Subphylum Postciliodesmatophora

Class Karyorelictea. Long, wormlike ciliates that may be extremely contractile; in some, 1 of the surfaces may be devoid of cilia; 2 to many macronulei; includes 4 orders; representive genera include *Protocruzia, Trachelonema, Loxodes,* and *Geleia.*

Class Spirotrichea. Conspicuous right and left oral or preoral ciliature; cytostome may be deep or shallow, and some species live inside loricae (tubes secreted by themselves or an agglomerate of material); lorica may be attached to a substrate; 3 subclasses, Heterotrichia, Choreotrichia (including the loricate tintinnid ciliates and the oligotrichs), and Stichotrichia.

Subphylum Rhabdophora

Class Prostomatea. Cytostome apical to subapical, with a shallow cytostomal cavity; oral kinetids may be tangential to the perimeter of the oral area; includes *Coleps* and *Tiarina.*

Class Litostomatea. Simple oral ciliature may have oral toxicysts; includes *Dileptus* and *Lacrymaria.*

Subphylum Cyrtophora

Class Phyllopharyngea. Where the feeding stage is sessile, the dispersal stage has simple ciliature; includes suctorians, which lack cilia in the trophic stage and feed via tentacles, and the chonotrichs, which are usually sessile ectosymbionts of crustaceans.

Class Nassophorea. Often well-developed oral/cytopharyngeal zones; typical genera are *Nassula, Paramecium, Frontonia,* and hypotrichs, such as *Euplotes.*

Class Oligohymenophorea. Complex oral cilia distinct from the body cilia; oral cavity may be deep, with ciliary structures extending into it; includes *Tetrahymena* and *Vorticella.*

Class Colpodea. Some inhabit gelatinous sheaths (*Mycterothrix tuamotuensis*); includes *Woodruffia* and *Colpoda.*

BIBLIOGRAPHY. JOHN J. LEE, SEYMOUR H. HUTNER, and EUGENE C. BOVEE (eds.), *An Illustrated Guide to the Protozoa* (1985), includes sections on the different groups, with many excellent illustrations. MICHAEL A. SLEIGH, *Protozoa and Other Protists* (1989), covers the features and physiology of the major groups, with a section on symbiosis and ecology, and is well illustrated. Coverage of the ecology and physiology of ciliates, sarcodines, and flagellates may be found in JOHANNA LAYBOURN-PARRY, *A Functional Biology of Free-Living Protozoa* (1984); and TOM FENCHEL, *The Ecology of Protozoa* (1987). RICHARD R. KUDO, *Protozoology,* 5th ed. (1966, reissued 1971), although slightly out-of-date, is nonetheless a useful book covering the major groups in detail. MICHAEL LEVANDOWSKY and SEYMOUR H. HUTNER (eds.), *Biochemistry and Physiology of Protozoa,* 2nd ed., 4 vol. (1979–81), contains articles on all aspects of protozoan physiology and biochemistry, including structure, locomotion, metabolism, mating behaviour, and immunology of parasitic protozoan diseases. J.P. KREIER and J.R. BAKER, *Parasitic Protozoa* (1987), is an easy-to-read overview covering physiology, biochemistry, and life cycles. R.S. PHILLIPS, *Malaria* (1983), a small, easy-to-read book, discusses pathology, immunology, and control methods. JOHN R. HAYNES, *Foraminifera* (1981), covers the fossil record and the importance of these sarcodines, with many illustrations. O. ROGER ANDERSON, *Radiolaria* (1983), covers the biology of this marine group. F.J.R. TAYLOR (ed.), *The Biology of Dinoflagellates* (1987), contains articles on the physiology and other aspects of these flagellates. LYNDA J. GOFF (ed.), *Algal Symbiosis* (1983), contains articles on endosymbionts of some protozoans and their role. JOHN O. CORLISS, *The Ciliated Protozoa: Characterization, Classification, and Guide to the Literature,* 2nd ed. (1979), is a specialized text providing further references. LYNN MARGULIS, *Symbiosis and Cell Evolution: Life and Its Environment on the Early Earth* (1981), is an interesting text on the evolution of cellular structures. (J.E.M.L.-P.)

Psychological Tests and Measurement

The word "test" refers to any means (often formally contrived) used to elicit responses to which human behaviour in other contexts can be related. When intended to predict relatively distant future behaviour (*e.g.*, success in school), such a device is called an aptitude test. When used to evaluate the individual's present academic or vocational skill, it may be called an achievement test. In such settings as guidance offices, mental-health clinics, and psychiatric hospitals, tests of ability and personality may be helpful in the diagnosis and detection of troublesome behaviour. Industry and government alike have been prodigious users of tests for selecting workers. Research workers often rely on tests to translate theoretical concepts (*e.g.*, intelligence) into experimentally useful measures.

For coverage of related topics in the *Micropædia* and *Macropædia*, see the *Propædia*, section 434, and the *Index*.

The article is divided into the following sections:

General problems of measurement in psychology 289
  Types of measurement scales
  Primary characteristics of methods or instruments
  Other characteristics
Types of instruments and methods 290
Development of standardized tests 292
  Test content
  Test norms
Assessing test structure 293
  Factor analysis
  Profile analysis
Bibliography 293

GENERAL PROBLEMS OF MEASUREMENT IN PSYCHOLOGY

Physical things are perceived through their properties or attributes. A mother may directly sense the property called temperature by feeling her infant's forehead. Yet she cannot directly observe colicky feelings or share the infant's personal experience of hunger. She must infer such unobservable private sensations from hearing her baby cry or gurgle; from seeing him flail his arms, or frown, or smile. In the same way, much of what is called measurement must be made by inference. Thus, a mother suspecting her child is feverish may use a thermometer, in which case she ascertains his temperature by looking at the thermometer, rather than by directly touching his head.

Indeed, measurement by inference is particularly characteristic of psychology. Such abstract properties or attributes as intelligence or introversion never are directly measured but must be inferred from observable behaviour. The inference may be fairly direct or quite indirect. If persons respond intelligently (*e.g.*, by reasoning correctly) on an ability test, it can be safely inferred that they possess intelligence to some degree. In contrast, people's capacity to make associations or connections, especially unusual ones, between things or ideas presented in a test can be used as the basis for inferring creativity, although producing a creative product requires other attributes, including motivation, opportunity, and technical skill.

Types of measurement scales. To measure any property or activity is to assign it a unique position along a numerical scale. When numbers are used merely to identify individuals or classes (as on the backs of athletes on a football team), they constitute a nominal scale. When a set of numbers reflects only the relative order of things (*e.g.*, pleasantness-unpleasantness of odours), it constitutes an ordinal scale. An interval scale has equal units and an arbitrarily assigned zero point; one such scale, for example, is the Fahrenheit temperature scale. Ratio scales not only provide equal units but also have absolute zero points; examples include measures of weight and distance.

Although there have been ingenious attempts to establish psychological scales with absolute zero points, psychologists usually are content with approximations to interval scales; ordinal scales often are used as well.

Primary characteristics of methods or instruments. The primary requirement of a test is validity—traditionally defined as the degree to which a test actually measures whatever it purports to measure. A test is reliable to the extent that it measures consistently, but reliability is of no consequence if a test lacks validity. Since the person who draws inferences from a test must determine how well it serves his purposes, the estimation of validity inescapably requires judgment. Depending on the criteria of judgment employed, tests exhibit a number of different kinds of validity.

Validity

Empirical validity (also called statistical or predictive validity) describes how closely scores on a test correspond (correlate) with behaviour as measured in other contexts. Students' scores on a test of academic aptitude, for example, may be compared with their school grades (a commonly used criterion). To the degree that the two measures statistically correspond, the test empirically predicts the criterion of performance in school. Predictive validity has its most important application in aptitude testing (*e.g.*, in screening applicants for work, in academic placement, in assigning military personnel to different duties).

Alternatively, a test may be inspected simply to see if its content seems appropriate to its intended purpose. Such content validation is widely employed in measuring academic achievement but with recognition of the inevitable role of judgment. Thus, a geometry test exhibits content (or curricular) validity when experts (*e.g.*, teachers) believe that it adequately samples the school curriculum for that topic. Interpreted broadly, content covers desired skills (such as computational ability) as well as points of information in the case of achievement tests. Face validity (a crude kind of content validity) reflects the acceptability of a test to such people as students, parents, employers, and government officials. A test that looks valid is desirable, but face validity without some more basic validity is nothing more than window dressing.

In personality testing, judgments of test content tend to be especially untrustworthy, and dependable external criteria are rare. One may, for example, assume that a man who perspires excessively feels anxious. Yet his feelings of anxiety, if any, are not directly observable. Any assumed trait (anxiety, for example) that is held to underlie observable behaviour is called a construct. Since the construct itself is not directly measurable, the adequacy of any test as a measure of anxiety can be gauged only indirectly; *e.g.*, through evidence for its construct validity.

A test exhibits construct validity when low scorers and high scorers are found to respond differently to everyday experiences or to experimental procedures. A test presumed to measure anxiety, for example, would give evidence of construct validity if those with high scores ("high anxiety") can be shown to learn less efficiently than do those with lower scores. The rationale is that there are several propositions associated with the concept of anxiety: anxious people are likely to learn less efficiently, especially if uncertain about their capacity to learn; they are likely to overlook things they should attend to in carrying out a task; they are apt to be under strain and hence feel fatigued. (But anxious people may be young or old, intelligent or unintelligent.) If people with high scores

on a test of anxiety show such proposed signs of anxiety, that is, if a test of anxiety has the expected relationships with other measurements as given in these propositions, the test is viewed as having construct validity.

Reliability

Test reliability is affected by scoring accuracy, adequacy of content sampling, and the stability of the trait being measured. Scorer reliability refers to the consistency with which different people who score the same test agree. For a test with a definite answer key, scorer reliability is of negligible concern. When the subject responds with his own words, handwriting, and organization of subject matter, however, the preconceptions of different raters produce different scores for the same test from one rater to another; that is, the test shows scorer (or rater) unreliability. In the absence of an objective scoring key, a scorer's evaluation may differ from one time to another and from those of equally respected evaluators. Other things being equal, tests that permit objective scoring are preferred.

Reliability also depends on the representativeness with which tests sample the content to be tested. If scores on items of a test that sample a particular universe of content designed to be reasonably homogeneous (*e.g.,* vocabulary) correlate highly with those on another set of items selected from the same universe of content, the test has high content reliability. But if the universe of content is highly diverse in that it samples different factors (say, verbal reasoning and facility with numbers), the test may have high content reliability but low internal consistency.

For most purposes, the performance of a subject on the same test from day to day should be consistent. When such scores do tend to remain stable over time, the test exhibits temporal reliability. Fluctuations of scores may arise from instability of a trait; for example, the test taker may be happier one day than the next. Or temporal unreliability may reflect injudicious test construction.

Reliability estimates

Included among the major methods through which test reliability estimates are made is the comparable-forms technique, in which the scores of a group of people on one form of a test are compared with the scores they earn on another form. Theoretically, the comparable-forms approach may reflect scorer, content, and temporal reliability. This ideally demands that each form of the test be constructed by different but equally competent persons and that the forms be given at different times and evaluated by a second rater (unless an objective key is fixed).

In the test-retest method, scores of the same group of people from two administrations of the same test are correlated. If the time interval between administrations is too short, memory may unduly enhance the correlation. Or some people, for example, may look up words they missed on the first administration of a vocabulary test and thus be able to raise their scores the second time around. Too long an interval can result in different effects for each person due to different rates of forgetting or learning. Except for very easy speed tests (*e.g.,* in which a person's score depends on how quickly he is able to do simple addition), this method may give misleading estimates of reliability.

Internal-consistency methods of estimating reliability require only one administration of a single form of a test. One method entails obtaining scores on separate halves of the test, usually the odd-numbered and the even-numbered items. The degree of correspondence (which is expressed numerically as a correlation coefficient) between scores on these half-tests permits estimation of the reliability of the test (at full length) by means of a statistical correction.

This is computed by the use of the Spearman-Brown prophecy formula (for estimating the increased reliability expected to result from increase in test length). More commonly used is a generalization of this stepped-up, split-half reliability estimate, one of the Kuder-Richardson formulas. This formula provides an average of estimates that would result from all possible ways of dividing a test into halves.

Other characteristics. A test that takes too long to administer is useless for most routine applications. What constitutes a reasonable period of testing time, however, depends in part on the decisions to be made from the test. Each test should be accompanied by a practicable and economically feasible scoring scheme, one scorable by machine or by quickly trained personnel being preferred.

A large, controversial literature has developed around response sets; *i.e.,* tendencies of subjects to respond systematically to items regardless of content. Thus, a given test taker may tend to answer questions on a personality test only in socially desirable ways or to select the first alternative of each set of multiple-choice answers or to malinger (*i.e.,* to purposely give wrong answers).

Response sets

Response sets stem from the ways subjects perceive and cope with the testing situation. If they are tested unwillingly, they may respond carelessly and hastily to get through the test quickly. If they have trouble deciding how to answer an item, they may guess or, in a self-descriptive inventory, choose the "yes" alternative or the socially desirable one. They may even mentally reword the question to make it easier to answer. The quality of test scores is impaired when the purposes of the test administrator and the reactions of the subjects to being tested are not in harmony. Modern test construction seeks to reduce the undesired effects of subjects' reactions.

TYPES OF INSTRUMENTS AND METHODS

Psychophysical scales and psychometric, or psychological, scales. The concept of an absolute threshold (the lowest intensity at which a sensory stimulus, such as sound waves, is perceived) is traceable to the German philosopher Johann Friedrich Herbart. The German physiologist Ernst Heinrich Weber later observed that the smallest discernible difference of intensity is proportional to the initial stimulus intensity. Weber found, for example, that, while people could just notice the difference after a slight change in the weight of a 10-gram object, they needed a larger change before they could just detect a difference from a 100-gram weight. This finding, known as Weber's law, is expressed more technically in the statement that the perceived (subjective) intensity varies mathematically as the logarithm of the physical (objective) intensity of the stimulus.

Weber's law

In traditional psychophysical scaling methods, a set of standard stimuli (such as weights) that can be ordered according to some physical property is related to sensory judgments made by experimental subjects. By the method of average error, for example, subjects are given a standard stimulus and then made to adjust a variable stimulus until they believe it is equal to the standard. The mean (average) of a number of judgments is obtained. This method and many variations have been used to study such experiences as visual illusions, tactual intensities, and auditory pitch.

Psychological (psychometric) scaling methods are an outgrowth of the psychophysical tradition just described. Although their purpose is to locate stimuli on a linear (straight-line) scale, no quantitative physical values (*e.g.,* loudness or weight) for stimuli are involved. The linear scale may represent an individual's attitude toward a social institution, his judgment of the quality of an artistic product, the degree to which he exhibits a personality characteristic, or his preference for different foods. Psychological scales thus are used for having a person rate his own characteristics as well as those of other individuals in terms of such attributes, for example, as leadership potential or initiative. In addition to locating individuals on a scale, psychological scaling can also be used to scale objects and various kinds of characteristics: finding where different foods fall on a group's preference scale; or determining the relative positions of various job characteristics in the view of those holding that job. Reported degrees of similarities between pairs of objects are used to identify scales or dimensions on which people perceive the objects.

The American psychologist L.L. Thurstone offered a number of theoretical-statistical contributions that are widely used as rationales for constructing psychometric scales. One scaling technique (comparative judgment) is based empirically on choices made by people between members of any series of paired stimuli. Statistical treatment to provide numerical estimates of the subjective (perceived) distances between members of every pair of stimuli yields a psychometric scale. Whether or not these computed scale values are consistent with the observed comparative judgments can be tested empirically.

Equal-appearing intervals

Another of Thurstone's psychometric scaling techniques (equal-appearing intervals) has been widely used in attitude measurement. In this method judges sort statements reflecting such things as varying degrees of emotional intensity, for example, into what they perceive to be equally spaced categories; the average (median) category assignments are used to define scale values numerically. Subsequent users of such a scale are scored according to the average scale values of the statements to which they subscribe. Another psychologist, Louis Guttman, developed a method that requires no prior group of judges, depends on intensive analysis of scale items, and yields comparable results. Quite commonly used is the type of scale developed by Rensis Likert in which perhaps five choices ranging from strongly in favour to strongly opposed are provided for each statement, the alternatives being scored from one to five. A more general technique (successive intervals) does not depend on the assumption that judges perceive interval size accurately. The widely used graphic rating scale presents an arbitrary continuum with preassigned guides for the rater (*e.g.,* adjectives such as superior, average, and inferior).

Tests versus inventories. The term "test" most frequently refers to devices for measuring abilities or qualities for which there are authoritative right and wrong answers. Such a test may be contrasted with a personality inventory, for which it is often claimed that there are no right or wrong answers. At any rate, in taking what often is called a test, the subjects are instructed to do their best; in completing an inventory, they are instructed to represent their typical reactions. A distinction also has been made that in responding to an inventory the subjects control the appraisal, whereas in a test they do not. If a test is more broadly regarded as a set of stimulus situations that elicit responses from which inferences can be drawn, however, then an inventory is, according to this definition, a variety of test.

Free-response versus limited-response tests. Free-response tests entail few restraints on the form or content of response, whereas limited-response tests restrict responses to one of a smaller number presented (*e.g.,* true-false). An essay test tends toward one extreme (free response), while a so-called fully objective test is at the other extreme (limited response).

Essay and objective tests

Response to an essay question is not completely unlimited, however, since the answer should bear on the question. The free-response test does give practice in writing, and, when an evaluator is proficient in judging written expression, his comments on the test may aid the individual to improve his writing style. All too often, however, writing ability unfortunately affects the evaluator's judgment of how well the test taker understands content, and this tends to reduce test reliability. Another source of unreliability for essay tests is found in their limited sampling of content, as contrasted with the broader coverage that is possible with objective tests. Often both the scorer and the content reliability of essay tests can be improved, but such attempts are costly.

The objective test, which minimizes scorer unreliability, is best typified by the multiple-choice form, in which the subject is required to select one from two or (preferably) more responses to a test item. Matching items that have a common set of alternatives are of this form. The true-false test question is a special multiple-choice form that may tend to arouse antagonism because of variable standards of truth or falsity.

The more general multiple-choice item is more acceptable when it is specified only that the best answer be selected; it is flexible, has high scorer reliability, and is not limited to simple factual knowledge. The ingenious test constructor can use multiple-choice items to test such functions as generalization, application of principles, and the ability to educe unfamiliar relationships.

Forced-choice items

Some personality tests are presented in a forced-choice format. They may, for example, force the person to choose one of two favourable words or phrases (*e.g.,* intelligent-handsome) as more descriptive of himself or one of two unfavourable terms as less descriptive (*e.g.,* stupid-ugly). Marking one choice yields a gain in score on some trait but may also preclude credit on another trait. This technique is intended to eliminate any effects from subjects' attempts to present themselves in a socially desirable light; it is not fully successful, however, because what is highly desirable for one person may be less desirable for another.

The forced-choice technique for self-appraisals is exemplified in a widely used interest inventory. Forced-choice ratings were introduced for evaluation of one military officer by another during World War II. They were an effort to avoid the preponderance of high ratings typically obtained with ordinary rating scales. Raters tend to give those being rated the benefit of any doubt, especially when they are fellow workers. Also, supervisors or teachers may give unduly favourable ratings because they believe good performance of subordinates or students reflects well on themselves.

Falling between free- and limited-response tests is a type that requires a short answer, perhaps a single word or a number, for each item. When the required response is to fit into a blank in a sentence, the test is called a completion test. This type of test is susceptible to scorer unreliability.

A personality test to which a subject responds by interpreting a picture or by telling a story it suggests resembles an essay test except that responses ordinarily are oral. A personality inventory that requires the subject to indicate whether or not a descriptive phrase applies to him is of the limited-response type. A sentence-completion personality test that asks the subject to complete statements such as "I worry because . . . " is akin to the short-answer and completion types.

Verbal versus performance tests. A verbal (or symbol) test poses questions to which the subject supplies symbolic answers (in words or in other symbols, such as numbers). In performance tests, the subject actually executes some motor activity; for example, he assembles mechanical objects. Either the quality of performance as it takes place or its results may be rated.

The verbal test, permitting group administration, requiring no special equipment, and often being scorable by relatively unskilled evaluators, tends to be more practical than the performance test. Both types of devices also have counterparts in personality measurement, in which verbal tests as well as behaviour ratings are used.

Written (group) versus oral (individual) tests. The oral test is administered to one person at a time, but written tests can be given simultaneously to a number of subjects. Oral tests of achievement, being uneconomical and prone to content and scorer unreliability, have been supplanted by written tests; notable exceptions include the testing of illiterates and the anachronistic oral examinations to which candidates for graduate degrees are liable.

Proponents of individually administered intelligence tests (*e.g.,* the Stanford-Binet) state that such face-to-face testing optimizes rapport and motivation, even among literate adult subjects. Oral tests of general aptitude remain popular, though numerous written group tests have been designed for the same purpose.

Interviews as tests

The interview may provide a personality measurement and, especially when it is standardized as to wording and order of questions and with a key for coding answers, may amount to an individual oral test. Used in public opinion surveys, such interviews are carefully designed to avoid the effects of interviewer bias and to be comprehensible to a highly heterogeneous sample of respondents.

Appraisal by others versus self-appraisal. In responding to personality inventories and rating scales, a person presumably reveals what he thinks he is like; that is, he appraises himself. Other instruments may reflect what one person thinks of another. Because self-appraisal often lacks objectivity, appraisal by another individual is common in such things as ratings for promotions. Ordinary tests of ability clearly involve evaluation of one person by another, although the subject's self-evaluation may intrude; for example, he may lack confidence to the point where he does not try to do his best.

Projective tests. The stimuli (*e.g.,* inkblots) in a projective test are intentionally made ambiguous and open to different interpretations in the expectation that each sub-

ject will project his own unique (idiosyncratic) reactions in his answers. Techniques for evaluating such responses range from the intuitive impressions of the rater to complex, coded schemes for scoring and interpretation that require extensive manuals; some projective tests are objectively scorable.

Speed tests versus power tests. A pure speed test is homogeneous in content (*e.g.,* a simple clerical checking test), the tasks being so easy that with unlimited time all but the most incompetent of subjects could deal with them successfully. The time allowed for testing is so short, however, that even the ablest subject is not expected to finish. A useful score is the number of correct answers made in a fixed time. In contrast, a power test (*e.g.,* a general vocabulary test) contains items that vary in difficulty to the point that no subject is expected to get all items right even with unlimited time. In practice, a definite but ample time is set for power tests.

Speed tests are suitable for testing visual perception, numerical facility, and other abilities related to vocational success. Tests of psychomotor abilities (*e.g.,* eye–hand coordination) often involve speed. Power tests tend to be more relevant to such purposes as the evaluation of academic achievement, for which the highest level of difficulty at which a person can succeed is of greater interest than his speed on easy tasks.

In general, tests reflect unknown combinations of the effects of speed and power; many consist of items that vary considerably in difficulty, and the time allowed is too limited to allow a large proportion of subjects to attempt all items.

Teacher-made versus standardized tests. A distinction between teacher-made tests and standardized tests is often made in relation to tests used to assess academic achievement. Ordinarily, teachers do not attempt to construct tests of general or special aptitude or of personality traits. Teacher-made tests tend instead to be geared to narrow segments of curricular content (*e.g.,* a sixth-grade geography test). Standardized tests with carefully defined procedures for administration and scoring to ensure uniformity can achieve broader goals. General principles of test construction and such considerations as reliability and validity apply to both types of test.

Sociodrama and psychodrama

Special measurement techniques. Sociodrama and psychodrama were originally developed as psychotherapeutic techniques. In sociodrama, group members participate in unrehearsed drama to illuminate a general problem. Psychodrama centres on one individual in the group whose unique personal problem provides the theme. Related research techniques (*e.g.,* the sociometric test) can offer insight into interpersonal relationships. Individuals may be asked to specify members of a group whom they prefer as leader, playmate, or coworker. The choices made can then be charted in a sociogram, from which cliques or socially isolated individuals may be identified at a glance.

Research psychologists have grasped the sociometric approach as a means of measuring group cohesiveness and studying individual reactions to groups. The degree to which any group member chooses or is chosen beyond chance expectation may be calculated, and mathematical techniques may be used to determine the complex links among group members. Sociogram-choice scores have been useful in predicting such criteria as individual productivity in factory work and combat effectiveness.

DEVELOPMENT OF STANDARDIZED TESTS

Test content. *Item development.* Once the need for a test has been established, a plan to define its content may be prepared. For achievement tests, the test plan may also indicate thinking skills to be evaluated. Detailed content headings can be immediately suggestive of test items. It is helpful if the plan specifies weights to be allotted to different topics, as well as the desired average score and the spread of item difficulties. Whether or not such an outline is made, the test constructor clearly must understand the purpose of the test, the universe of content to be sampled, and the forms of the items to be used.

Tryouts and item analysis. A set of test questions is first administered to a small group of people deemed to be representative of the population for which the final test is intended. The trial run is planned to provide a check on instructions for administering and taking the test and for intended time allowances, and it can also reveal ambiguities in the test content. After adjustments, surviving items are administered to a larger, ostensibly representative group. The resulting data permit computation of a difficulty index for each item (often taken as the percentage of the subjects who respond correctly) and of an item-test or item-subtest discrimination index (*e.g.,* a coefficient of correlation specifying the relationship of each item with total test score or subtest score).

Item validation

If it is feasible to do so, measures of the relation of each item to independent criteria (*e.g.,* grades earned in school) are obtained to provide item validation. Items that are too easy or too difficult are discarded; those within a desired range of difficulty are identified. If internal consistency is sought, items that are found to be unrelated to either a total score or an appropriate subtest score are ruled out, and items that are related to available external criterion measures are identified. Those items that show the most efficiency in predicting an external criterion (highest validity) usually are preferred over those that contribute only to internal consistency (reliability).

Estimates of reliability for the entire set of items, as well as for those to be retained, commonly are calculated. If the reliability estimate is deemed to be too low, items may be added. Each alternative in multiple-choice items also may be examined statistically. Weak incorrect alternatives can be replaced, and those that are unduly attractive to higher scoring subjects may be modified.

Cross validation. Item-selection procedures are subject to chance errors in sampling test subjects, and statistical values obtained in pretesting are usually checked (cross validated) with one or more additional samples of subjects. Typically, it is found that cross-validation values tend to shrink for many of the items that emerged as best in the original data, and further items may be found to warrant discard. Measures of correlation between total test score and scores from other, better known tests are often sought by test users.

Differential weighting. Some test items may appear to deserve extra, positive weight; some answers in multiple-choice items, though keyed as wrong, seem better than others in that they attract people who earn high scores generally. The bulk of theoretical logic and empirical evidence, nonetheless, suggests that unit weights for selected items and zero weights for discarded items and dichotomous (right versus wrong) scoring for multiple-choice items serve almost as effectively as more complicated scoring. Painstaking efforts to weight items generally are not worth the trouble.

Negative weight for wrong answers is usually avoided as presenting undue complication. In multiple-choice items, the number of answers a subject knows, in contrast to the number he gets right (which will include some lucky guesses), can be estimated by formula. But such an average correction overpenalizes the unlucky and underpenalizes the lucky. If the instruction is not to guess, it is variously interpreted by persons of different temperament; those who decide to guess despite the ban are often helped by partial knowledge and tend to do better.

A responsible tactic is to try to reduce these differences by directing subjects to respond to every question, even if they must guess. Such instructions, however, are inappropriate for some competitive speed tests, since candidates who mark items very rapidly and with no attention to accuracy excel if speed is the only basis for scoring; that is, if wrong answers are not penalized.

Test norms. Test norms consist of data that make it possible to determine the relative standing of an individual who has taken a test. By itself, a subject's raw score (*e.g.,* the number of answers that agree with the scoring key) has little meaning. Almost always, a test score must be interpreted as indicating the subject's position relative to others in some group. Norms provide a basis for comparing the individual with a group.

Centiles

Numerical values called centiles (or percentiles) serve as the basis for one widely applicable system of norms. From

a distribution of a group's raw scores the percentage of subjects falling below any given raw score can be found. Any raw score can then be interpreted relative to the performance of the reference (or normative) group—eighth-graders, five-year-olds, institutional inmates, job applicants. The centile rank corresponding to each raw score, therefore, shows the percentage of subjects who scored below that point. Thus, 25 percent of the normative group earn scores lower than the 25th centile; and an average called the median corresponds to the 50th centile.

Another class of norm system (standard scores) is based on how far each raw score falls above or below an average score, the arithmetic mean. One resulting type of standard score, symbolized as z, is positive (*e.g.,* +1.69 or +2.43) for a raw score above the mean and negative for a raw score below the mean. Negative and fractional values can, however, be avoided in practice by using other types of standard scores obtained by multiplying z scores by an arbitrarily selected constant (say, 10) and by adding another constant (say, 50, which changes the z score mean of zero to a new mean of 50). Such changes of constants do not alter the essential characteristics of the underlying set of z scores.

The French psychologist Alfred Binet, in pioneering the development of tests of intelligence, listed test items along a normative scale on the basis of the chronological age (actual age in years and months) of groups of children that passed them. A mental-age score (*e.g.,* seven) was assigned to each subject, indicating the chronological age (*e.g.,* seven years old) in the reference sample for which his raw score was the mean. But mental age is not a direct index of brightness; a mental age of seven in a 10-year-old is different from the same mental age in a four-year-old.

To correct for this, a later development was a form of IQ (intelligence quotient), computed as the ratio of the subject's mental age to his chronological age, multiplied by 100. (Thus, the IQ made it easy to tell if a child was bright or dull for his age.)

Ratio IQs for younger age groups exhibit means close to 100 and spreads of roughly 45 points above and below 100. The classical ratio IQ has been largely supplanted by the deviation IQ, mainly because the spread around the average has not been uniform due to different ranges of item difficulty at different age levels. The deviation IQ, a type of standard score, has a mean of 100 and a standard deviation of 16 for each age level. Practice with the Stanford-Binet test reflects the finding that average performance on the test does not increase beyond age 18. Therefore, the chronological age of any individual older than 18 is taken as 18 for the purpose of determining IQ.

Deviation IQ

The Stanford-Binet has been largely supplanted by several tests developed by the American psychologist David Wechsler between the late 1930s and the early 1960s. These tests have subtests for several capacities, some verbal and some operational, each subtest having its own norms. After constructing tests for adults, Wechsler developed tests for older and for younger children.

ASSESSING TEST STRUCTURE

Factor analysis. Factor analysis is a method of assessment frequently used for the systematic analysis of intellectual ability and other test domains, such as personality measures. Just after the turn of the 20th century the British psychologist Charles E. Spearman systematically explored positive intercorrelations between measures of apparently different abilities to provide evidence that much of the variability in scores that children earn on tests of intelligence depends on one general underlying factor, which he called *g*. In addition he believed that each test contained an *s* factor specific to it alone. In the United States, Thurstone developed a statistical technique called multiple-factor analysis, with which he was able to demonstrate, in a set of tests of intelligence, that there were primary mental abilities, such as verbal comprehension, numerical computation, spatial orientation, and general reasoning. Although later work has supported the differentiation between these abilities, no definitive taxonomy of abilities has become established. One element in the problem is the finding that each such ability can be shown to be composed of narrower factors.

The first computational methods in factor analysis have been supplanted by mathematically more elegant, computer-generated solutions. While earlier techniques were primarily exploratory, the Swedish statistician Karl Gustav Jöreskog and others have developed procedures that permit the researcher to test hypotheses about the structure in a set of data.

Rooted in extensive applications of factor analysis, a structure-of-intellect model developed by the American psychologist Joy Paul Guilford posited a very large number of factors of intelligence. Guilford envisaged three intersecting dimensions corresponding respectively to four kinds of test content, five kinds of intellectual operation, and six kinds of product. Each of the 120 cells in the cube thus generated was hypothesized to represent a separate ability, each constituting a distinct factor of intellect. Educational and vocational counselors usually prefer a substantially smaller number of scores than the 120 implied by this model.

Factors of intelligence

Factor analysis has also been widely used outside the realm of intelligence, especially to seek the structure of personality as reflected in ratings by oneself and by others. Although there is even less consensus here than for intelligence, a number of studies suggest that four prevalent factors can be approximately labeled, namely, conformity, extroversion, anxiety, and dependability.

Profile analysis. With the fractionation of tests (*e.g.,* to yield scores measuring separate factors or clusters), new concern has arisen for interpreting differences among scores measuring the underlying variables, however conceived. Scores of an individual on several such measures can be plotted graphically as a profile; for direct comparability, all raw scores may be expressed in terms of standard scores that have equal means and variabilities. The difference between any pair of scores that have less than perfect reliability tends to be less reliable than either, and fluctuations in the graph should be interpreted cautiously. Nevertheless, various features of an individual's profile may be examined, such as scatter (fluctuation from one measure to another) and relative level of performance on different measures. (The particular shape of the graph, it should be noted, partly depends upon the arbitrary order in which measures are listed.) One may also statistically express the degree of similarity between any two profiles. Such statistical measures of pattern similarity permit quantitative comparison of profiles for different persons, of profiles of the same individual's performance at different times, of individual with group profiles, or of one group profile with another. Comparison of an individual's profile with similar graphs representing the means for various occupational groups, for example, is useful for vocational guidance or personnel selection.

BIBLIOGRAPHY. DOROTHY C. ADKINS, *Test Construction,* 2nd ed. (1974), a simplified treatment of measurement principles, rules for test construction, and statistical techniques; ANNE ANASTASI, *Psychological Testing,* 5th ed. (1982), an authoritative text and reference book, with emphasis on current psychological tests; LEE J. CRONBACH, *Essentials of Psychological Testing,* 4th ed. (1984), a modern and insightful text and general reference; J.P. GUILFORD, *Psychometric Methods,* 2nd ed. (1954), a widely used book that attempts to integrate psychophysical scaling and psychological measurement methods; HAROLD GULLIKSEN, *Theory of Mental Tests* (1950), a basic theoretical reference; HARRY H. HARMAN, *Modern Factor Analysis,* 3rd rev. ed. (1976), an eclectic treatment of factor-analytic theory and methods; PAUL HORST, *Psychological Measurement and Prediction* (1966), a discussion of practical requirements of psychological measurement as well as of technical problems in prediction; FREDERIC M. LORD and MELVIN R. NOVICK, *Statistical Theories of Mental Test Scores* (1968), a highly technical presentation; GEORG RASCH, *Probabilistic Models for Some Intelligence and Attainment Tests* (1980), with a new model for tests; and ROBERT L. THORNDIKE (ed.), *Educational Measurement,* 2nd ed. (1971), with specially prepared chapters by authorities in particular fields of measurement.

(D.C.A./D.W.Fi.)

Public Administration

Public administration, traditionally defined, denotes the implementation of government policies. Today public administration is often regarded as including also some responsibility for determining the policies and programs of governments. Specifically, it is the planning, organizing, directing, coordinating, and controlling of government operations.

Public administration is a feature of all nations, whatever their system of government. Within nations public administration is practiced at the central, intermediate, and local levels. Indeed, the relationships between different levels of government within a single nation constitute a growing problem of public administration.

In most of the world the establishment of highly trained administrative, executive, or directive classes has made public administration a distinct profession. The body of public administrators is usually called the civil service. In the United States and a few other countries, the elitist class connotation traditionally attached to the civil service has been either consciously abandoned or avoided, with the result that professional recognition has come slowly and only partially.

Traditionally the civil service is contrasted with other bodies serving the state full time, such as the military, the judiciary, and the police. Specialized services, sometimes referred to as scientific or professional civil services, provide technical rather than general administrative support. Traditionally, in most countries, a distinction is also made between the home civil service and those persons engaged abroad on diplomatic duties. A civil servant, therefore, is one of a body of persons who are directly employed in the administration of the internal affairs of the state and whose role and status are not political, ministerial, military, or constabulary.

In most countries the civil service does not include local government or public corporations, such as, in the United Kingdom, the National Coal Board. In some countries, however—particularly those unitary states in which provincial administration is part of the central government—some provincial staffs are civil servants. In the United States, all levels of government have their own civil services, federal, state, and local, and a civil service is specifically that part of governmental service entered by examination and offering permanent tenure.

Certain characteristics are common to all civil services. Senior civil servants are regarded as the professional advisers to those who formulate state policy. In some countries entry requirements for a career in the higher civil service stress qualifications in technical fields such as accounting, economics, medicine, and engineering. In other countries legal training is deemed appropriate, and in others no specific technical or academic discipline is required among candidates for senior posts. Whatever their precise qualifications, senior civil servants are professional in the sense that their experience of public affairs is thought to provide them with the knowledge of the limits within which state policy can be made effective and of the probable administrative results of different courses of action. Civil servants in every country are expected to advise, warn, and assist those responsible for state policy and, when this has been decided, to provide the organization for implementing it. The responsibility for policy decisions lies with the political members of the executive (those members who have been elected or appointed to give political direction to government and, customarily, career civil servants). By custom, civil servants are protected from public blame or censure for their advice. The acts of their administration may, however, be subject to special judicial controls from which no member of the executive can defend them.

Civil services are organized upon standard hierarchical lines, in which a command structure rises pyramid-fashion from the lowest offices to the highest. This command implies obedience to the lawful orders of a superior, and in order to maintain this system the hierarchy of offices is marked by fixed positions, with well-defined duties, specific powers, and salaries and privileges objectively assessed. In some countries there may be direct appointment to higher office of persons not previously employed by the service, but even then a recognized system of internal promotion emphasizes the nature of the hierarchical pyramid.

This article discusses the growth of public administration through history as well as its development under different political systems. Special attention is paid to the problems of administrative law and bureaucratic structure. For discussion of a subject integral to public administration, see GOVERNMENT FINANCE. For further discussion of the various regimes under which public administration operates, see POLITICAL SYSTEMS.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 536, 542, and 552.

(F.C.M./B.Ch./E.C.P.)

The article is divided into the following sections:

History 294
  Early systems 294
  Modern developments 295
    Prussia
    France
    The British Empire
    The United States
    The Soviet Union
    China
    Japan
    Developing nations
Principles of public administration 298
  The classical definition 298
  Recent interpretations 298
  Education and training 299
Organization of civil service 300
  Appointment 300
  Conditions of service 300
  Patterns of control 301
  International civil service 302
Administrative law 302
  Defining principles 303
    Distinctions between public administration and private action
    The need for legal safeguards over public administration
    Bureaucracy and the role of administrative law
  Judicial review of administration 304
    The common-law system
    The council of state system
    The procurator system
  The ombudsman 307
  Administrative procedure 307
Bibliography 308

History

EARLY SYSTEMS

Public administration has ancient origins. In antiquity the Egyptians and Greeks organized public affairs by office, and the principal officeholders were regarded as being principally responsible for administering justice, maintaining law and order, and providing plenty. The Romans developed a more sophisticated system under their empire, creating distinct administrative hierarchies for justice,

military affairs, finance and taxation, foreign affairs, and internal affairs, each with its own principal officers of state. An elaborate administrative structure, later imitated by the Roman Catholic Church, covered the entire empire, with a hierarchy of officers reporting back through their superiors to the emperor. This sophisticated structure disappeared after the fall of the Roman Empire in western Europe in the 5th century, but many of its practices continued in the Byzantine Empire in the east, where civil service rule was reflected in the pejorative use of the word Byzantinism.

Medieval Europe

Early European administrative structures developed from the royal households of the medieval period. Until the end of the 12th century official duties within the royal households were ill-defined, frequently with multiple holders of the same post. Exceptions were the better-defined positions of butler (responsible for the provision of wine), steward (responsible for feasting arrangements), chamberlain (often charged with receiving and paying out money kept in the royal sleeping chamber), and chancellor (usually a priest with responsibilities for writing and applying the seal in the monarch's name). With the 13th century a separation began between the purely domestic functions of the royal household and the functions connected with governing the state. The older household posts tended to disappear, become sinecures, or decline in importance. The office of chancellor, which had always been concerned with matters of state, survived to become the most important link between the old court offices and modern ministries, and the development of the modern treasury or finance ministry can be traced back to the chamberlain's office in the royal household.

From the middle of the 13th century three institutions began to emerge as the major bodies for handling affairs of state: the high court (evolving primarily from the chancellery), the exchequer, and the collegial royal council. In England and France, however, it was not until the early 14th century that such bodies emerged. In Brandenburg, which was governed by an elector (a prince with a right to elect the Holy Roman emperor) and which later formed the basis of the Prussian state, they became distinct entities only at the beginning of the 17th century.

Apart from justice and treasury departments, which originated in old court offices, modern ministerial structures in Europe developed out of the royal councils, which were powerful bodies of nobles appointed by the monarch. From the division of labour within these bodies the monarchs' secretaries, initially given low status within a council, emerged as perhaps the first professional civil servants in Europe in the modern sense. The proximity of the secretaries to the monarch gave them more knowledge of royal intentions, and their relative permanence gave them greater expertise in particular matters of state than could be found among the more transient nobles on the council. They were also assisted by staffs. The secretaries grew in importance in the 15th and 16th centuries as they became more or less full members of the council.

The distribution of functions among secretaries was initially based upon geography. In England this geographical allocation—with, for example, a secretary of the North and a secretary of the South—persisted until 1782, when the offices of home and foreign secretary were created. In France a more complex allocation of territorial responsibilities among secretaries of state had begun to give way to functional responsibilities by the end of the ancien régime in 1789.

Chinese civil service

The civil service in China was undoubtedly the longest lasting in history; it was first organized, along with a centralized administration, during the Han dynasty (206 BC–AD 220) and improved under the T'ang (618–907) and Sung (960–1279). The administration was organized so well that the pattern stood until 1912. During the Sung dynasty there developed the full use of civil service examinations. Candidates were subjected to successive elimination through written tests on three levels, more than a hundred persons beginning the ordeal for each one who emerged successful. Although there was strong emphasis on the Chinese Classics (because knowledge of the Classics was thought to form the virtues of a good citizen), there was also an effort to devise objective and meaningful tests for practical qualities, and there were always long contentions over subject matter and testing methods. To preserve the anonymity of the candidate and to ensure fairness in grading, examination papers were copied by clerks, examinees were identified by number only, and three examiners read each paper. Higher officials were privileged to nominate junior relatives for admission to the bureaucracy, but the great stress on examination grades in promotion, the use of annual merit ratings, and the practice of recruiting many lower officials from the ranks of the clerical service ensured a considerable freedom of opportunity.

MODERN DEVELOPMENTS

Prussia. The foundations of modern public administration in Europe were laid in Prussia in the late 17th and 18th centuries. The electors of Brandenburg (who from 1701 were the kings of Prussia) considered a rigidly centralized government a means of ensuring stability and furthering dynastic objectives. Their principal effort was devoted in the first instance to the suppression of the autonomy of the cities and to the elimination of the feudal privileges of the aristocracy. Civil servants were therefore appointed by the central government to administer the provinces, where the management of crown lands and the organization of the military system were combined in a Kriegs-und-Domänen-kammer ("Office of War and Crown Lands"). Subordinate to these offices were the *Steuerräte* ("tax councillors"), who controlled the administration of the municipalities and communes. These officials were all appointed by the central government and were responsible to it. At the apex of the new machinery of government was the sovereign.

Military functions of the Prussian bureaucracy

This centralized system was strengthened by creating a special corps of civil servants. In the beginning these civil servants—in a real sense servants of the crown—were sent out from Berlin to deal with such purely military matters as recruiting, billeting, and victualing the troops, but in the course of time they extended their supervision to civil matters as well. By 1713 there were clearly recognizable administrative units dealing in civil affairs and staffed by crown civil servants.

Special ordinances in 1722 and 1748 regulated recruitment to the civil service. Senior officials were required to propose to the king the names of candidates suitable for appointment to the higher posts, while the adjutant general proposed noncommissioned officers suitable for subordinate administrative posts. Further steps were taken throughout the 18th century to regularize the system of recruitment, promotion, and internal organization. All of these matters were brought together in a single General Code promulgated in 1794. The merit system of appointment covered all types of posts, and the general principle laid down was that "special laws and instructions determine the appointing authority to different civil service rank, their qualifications, and the preliminary examinations required from different branches and different ranks." Entry to the higher civil service required a university degree in cameralistics, which, though strictly speaking the science of public finance, included also the study of administrative law, police administration, estate management, and agricultural economics. After the degree course, candidates for the higher civil service spent a further period of supervised practical training in various branches of the administration, at the end of which they underwent a further oral and written examination. The basic principles of modern civil services are to be found in this General Code.

France. A fundamental change in the status of the civil servant came about as a result of the French Revolution of 1789. The fall of the ancien régime and the creation of a republic meant that the civil servant was seen as the servant no longer of the king but rather of the state—even though rule by a king or emperor was soon brought back and continued in France for nearly another century. The civil servant became an instrument of public power, not the agent of a person. This depersonalization of the state encouraged a rapid growth in the field of public law concerned with the organization, duties, and rights of "the

public power," of which civil servants were the principal component. To the ordered structure of the Prussian bureaucracy there began to be added the logical development of administrative law.

Innovations under Napoleon I

This bureaucratization was greatly fostered by Napoleon I, who built up a new civil service marked not only by some of the features of military organization but also by the principles of rationality, logic, and universality that were the inheritance of the Enlightenment. There was a clear chain of command and a firmly established hierarchy of officials, with duties clearly apportioned between authorities. Authority was depersonalized and went to the office and not the official—although Napoleon insisted that each official should be responsible for action taken in the name of his office. France was divided into new territorial units: *départements, arrondissements,* and *communes.* In each of these, state civil servants had a general responsibility for maintaining public order, health, and morality. They were all linked in a chain to the national Ministry of the Interior. A special school, the École Polytechnique, was set up to provide the state with technical specialists in both the military and the civil fields—particularly in general administration. In the field of general administration, the Conseil d'État ("Council of State"), descended from the old Conseil du Roi ("Council of the King"), imposed an intellectual as well as a judicial authority over the rest of the civil service; as the first major European administrative court, it became the creator of a new type of administrative jurisprudence. The prestige of the new French administrative organization and the logical arrangement of its internal structure prompted many other European countries to copy its principal features. And the expansion of the French Empire spread many of its features across the world.

In France under the Third Republic (1870–1940) there developed, however, considerable political interference in some branches of the civil service; and much of its vitality was diminished as its bureaucratic practices tended to become unwieldy and its personnel lethargic. Not until 1946 was the system reformed—which involved overhauling the administrative structure of the central government, centralizing personnel selection, creating a special ministry for civil service affairs, and setting up a special school, the École National d'Administration, for the training of senior civil servants. This school in particular has attracted worldwide attention for its ability to instill in its graduates both specialist and generalist skills.

British civil service in India

The British Empire. The first attempts by Great Britain to create efficient administrative machinery arose from its commitment to govern India and to avoid in that country the periodic scandals that marked some of the rule of the East India Company. Robert Clive, appointed governor of Bengal for the second time in 1764, introduced a code of practice that prohibited servants of the company from trading on their own account or accepting gifts from native traders. Subsequent governors strengthened the ban, compensating for the loss of benefits by substantially increasing salaries, introducing promotion by seniority, and reorganizing the higher echelons of administration. Recruitment was carried on by the company in London, and after 1813 entrants to the civil service had to study the history, language, and laws of India for a period of four terms at Haileybury College, England, and to obtain a certificate of good conduct before taking up their posts. As a result of advocacy by Thomas Macaulay, secretary to the board of control, examination rather than patronage was adopted as a recruitment method. New rules from 1833 stipulated that four candidates had to be nominated for each vacancy and that they were to compete with one another in "an examination in such branches of knowledge and by such examinations as the Board of the Company shall direct."

Foundation of the modern civil service in Britain

There was further criticism of the way India was run, however, and in 1853 another legislative reform of the administration was proposed. The experience of the Indian Civil Service influenced the foundation of the modern civil service in the United Kingdom. A report was published in 1854 on the organization of the Permanent Civil Service in Britain. Its principal author, Sir Charles Trevelyan, had acquired a reputation for searching out corruption in the Indian Civil Service during 14 years of service there. The report of 1854 recommended the abolition of patronage and recruitment by open competitive examination. It further recommended (1) the establishment of an autonomous semijudicial body of civil service commissioners to ensure the proper administration of recruitment to official posts, (2) the division of the work of the civil service into intellectual and routine work, the two sets of offices to have separate forms of recruitment, and (3) the selection of higher civil servants more decidedly on the basis of general intellectual attainment than specialized knowledge. The Civil Service Commission was established in 1855, and during the next 30 years patronage was gradually eliminated. The two original classes were increased to four, and some specialized branches were amalgamated to become the Scientific Civil Service. The new civil service managed to attract to its senior levels highly capable, discreet, and self-effacing university graduates. Graduates of Oxford and Cambridge became—and remain to the present—especially prominent in the ranks of senior civil servants in Britain.

The United States. In the United States patronage remained the norm for considerably longer than in Britain. From the early days of the federation two principles were firmly held. First, there was antipathy to the notion of a cadre of permanent civil servants; President Jackson clearly dismissed this notion of a highly professional caste when he said, in 1829, that "the duties of all public officers are . . . so plain and simple that men of intelligence may readily qualify themselves for their performance." As a consequence, he said, "I can not but believe that more is lost by the long continuance of men in office than is generally to be gained by their experience. No one man has any more intrinsic right to official station than another." The second principle—that as far as possible public office should be elective—followed more or less automatically. But because this principle could not be practically applied to the subordinate levels of administration, there developed the "spoils system," in which public office became a perquisite of political victory, being widely used to reward political support. This system was susceptible to persistent, blatant, and ultimately unacceptable degrees of inefficiency, corruption, and partisanship. These particular faults were strongly felt after the Civil War (1861–65), during the period of rapid economic and social development. Under considerable pressure, the federal government accepted a restricted principle of entry by competitive open examination, and in 1883 the U.S. Civil Service Commission was established to control entry to office in the federal service. The work of the commission was mainly restricted to the lower grades of employment, and it was not until the first 20 years of the 20th century that the merit system of recruitment was expanded to cover half the posts in the federal service. After that period the commission's control gradually increased, mainly over the lower, middle, and managerial offices in the federal service. After 1978 the functions of the commission were divided between the Office of Personnel Management and the Merit Systems Protection Board. Principal policy-making posts, numbering some 2,000, remain outside the jurisdiction of these two bodies, being filled instead by presidential nomination.

The U.S. Civil Service Commission

The development of civil service in U.S. local government varied among states, counties, and cities. The adoption of a merit system can usually be dated from the early 20th century, during the reform period of the muckrakers. In some states the merit system became well established, with a central personnel office that included a civil service commission or board similar to the federal model. At the other extreme there was simply a central personnel office headed by a single personnel director with no advisory board. At the municipal level, by the mid-20th century, most large cities in the United States had developed some sort of merit system; in smaller cities, however, merit systems were correspondingly less common. In the counties, the majority of which were rural and had relatively few public employees, formally established merit systems were rare.

The Soviet Union. In Russia the Revolution of 1917 swept away the tsarist civil service. The Communist Party at first held that a strong administrative organization was bound to damage the revolution by dampening spontaneity and other revolutionary virtues. But it soon became clear that a regime dedicated to social engineering, economic planning, and world revolution needed trained administrators. The party fell back, albeit reluctantly, upon the expertise of the more reliable tsarist civil servants. It did, however, surround the new civil service with elaborate controls in an attempt to ensure that its members remained loyal to party directives.

As the Communist Party itself became bureaucratized and as the more enthusiastic revolutionary leaders were eliminated, special industrial academies were set up for party members who had shown administrative talent. With the First Five-Year Plan (1928–32) the status of civil servants was improved, and their conditions of service were made less rigid, even though the party never relaxed its tight system of control over all branches of the state apparatus. In 1935 the State Commission on the Civil Service was created and attached to the Commissariat of Finance with responsibility for ensuring general control of personnel practice. This commission laid down formal patterns of administrative structure, reformed existing bureaucratic practices, fixed levels of staffing, standardized systems of job classification, and eliminated unnecessary functions and staff. The inspectorate of the Ministry of Finance ensured that the commission's general policies were carried out in the ministries. The commission itself remained under the close supervision of the Council of People's Commissars to ensure that it complied with party directives, and the commission's members were appointed directly by the council.

Creation of Soviet commission

The Soviet commission, unlike those in such countries as Great Britain and the United States, was given no jurisdiction over the recruitment of civil servants, which remained the function of the ministries and agencies. The highest administrative and technical staff members were recruited by each ministry. Each branch of industry and administration had its own training schools, from which it selected qualified students with satisfactory records. On appointment, the student was bonded for a minimum of three years and liable to criminal proceedings if he refused or subsequently relinquished his assignment. At the lower levels of administration, recruitment and job placement were the responsibility of the Commissariat of Labour Reserves.

Merging of civil service and party elite

The Communist Party made determined attempts to recruit higher civil servants as party members. These drives, which followed periodically after the 1930s, went a long way toward transforming the party itself into an administrative and managerial elite and uniting the party and the state administration. The highest levels of the civil service came to constitute an influential apparatus and power centre in their own right. The internal structure of the civil service, moreover, had been fashioned along classic French and German lines; and titles, ranks, insignia, and uniforms officially appeared in various parts of the public services.

China. The People's Republic of China also illustrates the conflict between revolutionary suspicion of bureaucracy and the need to construct strong administrative machinery in order to attain revolutionary goals. China's long tradition of bureaucracy remained important even after the Communist Party came to power in 1949. Within a decade the weight of the administration had already led, according to party dogma, to a gap between the elite and the masses and also to excessive stratification among the ruling bureaucrats, or cadres, themselves. There was not only a distinction between "old cadres" and "new cadres," depending on nothing more substantial than the date of an official's entry into the revolutionary movement, but also a complex system of job evaluation that divided the civil service into 24 grades, each with its own rank, salary scales, and distinctions. The number of ratings represented very considerable differences of power, prestige, and prerogatives and produced psychological barriers between the highest and lowest grades at least as great and as conspicuous as between the cadres and the masses. These distinctions and discrepancies were widely attacked during the Cultural Revolution of the 1960s and '70s, but they remained deeply ingrained in the administrative structure.

Japan. Until the 17th century, Japan under the shogunate was administered by a military establishment made up of vassals and enfeoffed nobles. After the 1630s a civil bureaucracy developed and began to assume a more important role than the military. Appointment within the bureaucracy was based upon family rank, and officials were loyal primarily to the feudal lord. It was not until after Matthew C. Perry sailed four U.S. warships into Uraga Harbour in 1853, thus forcibly ending more than two centuries of Japan's isolation from the rest of the world, that the Japanese bureaucracy moved away from feudal rank as the basis of appointments, establishing in its place loyalty to the emperor rather than to feudal lords. Merit appointments were made on a modest scale immediately after Japan was opened to the West, yet it was not until the 1880s, during the Meiji Restoration, that a modern civil service was created on the basis of job security, career paths, and entry by open competition. Tokyo University law graduates tended to dominate this new civil service. Personal allegiance to the emperor was reflected in the status of Japanese civil servants as "Emperor's Officials."

Introduction of Western-style bureaucracy into Japan

After World War II the Allied occupation authorities directed the passage of a Japanese law guaranteeing that all public officials should be servants of the people rather than of the emperor. The National Public Service Law of 1947 set up an independent National Personnel Authority to administer recruitment, promotion, conditions of employment, standards of performance, and job classification for the new civil service. Technically the emperor himself became a civil servant, and detailed regulations brought within the scope of the new law all civil servants from labourers to the prime minister. Civil servants were classified into two groups, the regular service and a special service. Civil servants in the former category entered the service by competitive examination on a standard contract with tenure. The special service included elected officials and political appointees and covered such officials as members of the Diet (legislature), judges, members of the audit boards, and ambassadors.

Although in theory the sovereign people have an inalienable right to choose and dismiss all public officials—who are constitutionally described as "servants of the whole community"—both tradition and political practice have allowed the civil service in Japan to retain and consolidate its old position in government. The idealization of the scholar-bureaucrat (a Confucian tradition borrowed from China) makes the civil service an independent power centre. Political struggles in the Diet have led to constantly changing ministries, and individual ministers rarely stay at a post long enough to establish firm control of their administration. As in many democratic countries with volatile political systems, administrative control has tended to pass to senior civil servants.

Tradition of the scholar-bureaucrat

Developing nations. Less-developed countries have had to face the opposite problem with their civil services. After World War II many such countries became independent before they had developed effective administrative structures or bodies of trained civil servants. Few of the colonial powers had trained indigenous administrators sufficiently. The British left a viable administrative structure in India and a partly Indianized civil service, but the newly independent Pakistan had few experienced civil servants. The Belgians left the Congo without any trained administrative or technical staff, and for some years there was near anarchy.

Even when they inherited reasonably efficient administrative organizations, the newly independent countries' politicians frequently proved incapable of fulfilling their supporters' expectations. Civil servants from the old colonial powers who remained behind often found radical policies and new masters uncongenial. The resulting exodus of many such civil servants worsened matters, for indigenous civil servants were seldom an adequate substitute.

The lack of qualified personnel sometimes led to not only a reduction in efficiency but also a decline in admin-

istrative morality. Nepotism, tribalism, and corruption as well as inefficiency in the civil service were difficulties often added to the other trials of independence. In many countries the incapacity of the civil service was a factor leading to military rule, as were the political failings of the elected leaders. Military regimes have frequently been the last resort of a country where the civil power has failed to cope with the problems of independence. Consequently, the United Nations (UN), in conjunction with the governments of advanced countries, began to develop training programs for civil servants from underdeveloped countries. The first request came from Latin America, which led to the founding of a school of public administration in Brazil, followed in 1953 by an Advanced School of Public Administration for Central America. Various other international organizations, including the Organisation for Economic Co-operation and Development and the World Bank, supported institutions for the training of administrators in the less-developed countries. Such institutions included the Arab Planning Institute in Kuwait, the Arab Organization of Administrative Sciences in Jordan, and the Inter-American School of Public Administration in Brazil. Civil servants from the less-developed nations also studied administration at such places as the Institute of Social Studies in The Hague, Neth., the Institute of Local Government Studies in Birmingham, Eng., and the International Institute of Public Administration in Paris.

International training programs for civil servants

After the 1970s the international agencies gave less help toward training, on the assumption—often unrealized—that the less-developed nations would take on greater responsibility themselves. Training also tended to be generalist and academic, leading to acute shortages of trained administrators in specialized fields such as finance and planning. However, organizations such as the British Council began in the early 1980s to remedy some of these deficiencies. (B.Ch./E.C.P.)

Principles of public administration

THE CLASSICAL DEFINITION

Throughout the 20th century the study and practice of public administration has been essentially pragmatic and normative rather than theoretical and value free. This may explain why public administration, unlike some social sciences, developed without much concern about an encompassing theory. Not until the mid-20th century and the dissemination of the German sociologist Max Weber's theory of bureaucracy was there much interest in a theory of public administration. Most recent bureaucratic theory, however, has been addressed to the private sector, and there has been little effort to relate organizational to political theory.

A prominent principle of public administration has been economy and efficiency, that is, the provision of public services at the minimum cost. This has usually been the stated objective of administrative reform. Despite growing concern about other kinds of values, such as responsiveness to public needs, justice and equal treatment, and citizen involvement in government decisions, efficiency continues to be a major goal.

In its concern with efficiency and improvement, public administration has focused frequently on questions of formal organization. It is generally held that administrative ills can be at least partly corrected by reorganization. Many organizational principles originated with the military, a few from private business. They include, for example: (1) organizing departments, ministries, and agencies on the basis of common or closely related purposes, (2) grouping like activities in single units, (3) equating responsibility with authority, (4) ensuring unity of command (only one supervisor for each group of employees), (5) limiting the number of subordinates reporting to a single supervisor, (6) differentiating line (operating or end-purpose) activities from staff (advisory, consultative, or support) activities, (7) employing the principle of management by exception (only the unusual problem or case is brought to the top), and (8) having a clear-cut chain of command downward and of responsibility upward.

Classical principles of organization

Some critics have maintained that these and other principles of public administration are useful only as rough criteria for given organizational situations. They believe that organizational problems differ and that the applicability of rules to various situations also differs. Nonetheless, and despite much more sophisticated analyses of organizational behaviour in recent decades, such principles as those enumerated above continue to carry force.

Public administration has also laid stress upon personnel. In most countries administrative reform has involved civil service reform. Historically, the direction has been toward "meritocracy"—the best individual for each job, competitive examinations for entry, and selection and promotion on the basis of merit. Attention has increasingly been given to factors other than intellectual merit, including personal attitudes, incentives, personality, personal relationships, and collective bargaining.

In addition, the budget has developed as a principal tool in planning future programs, deciding priorities, managing current programs, linking executive with legislature, and developing control and accountability. The contest for control over budgets, particularly in the Western world, began centuries ago and at times was the main relationship between monarchs and their subjects. The modern executive budget system in which the executive recommends, the legislature appropriates, and the executive oversees expenditures originated in 19th-century Britain. In the United States during the 20th century, the budget became the principal vehicle for legislative surveillance of administration, executive control of departments, and departmental control of subordinate programs. It has been assuming a similar role in many of the developing countries of the world.

RECENT INTERPRETATIONS

The classical approach to public administration described above probably reached its fullest development in the United States during the 1930s, although since that time, through educational and training programs, technical assistance, and the work of international organizations, it has also become standard doctrine in many countries. However, some of its elements have been resisted by governments with British or continental-legal perspectives, and even during the 1930s it was being challenged from several quarters. Since that time study of the subject has greatly developed. It has also become somewhat confused as a result of certain inconsistencies in approach.

The orthodox doctrine rested on the premise that administration was simply the implementation of public policies determined by others. According to this view, administrators should seek maximum efficiency but should be otherwise neutral about values and goals. During the Great Depression of the 1930s, and even more so during World War II, however, it became increasingly evident that many new policies originated within the administration, that policy and value judgments were implicit in most significant administrative decisions, that many administrative officials worked on nothing except policy, and that, insofar as public policies were controversial, such work inevitably involved administrators in politics. The supposed independence of administration from policy and politics was seen to be illusory. Since the 1930s there has thus been increasing concern with policy formation and the development of techniques to improve policy decisions. Although the concept of a value-free, neutral administration is regarded by many as no longer tenable, no fully satisfactory substitute has been offered. How to ensure that responsible and responsive policy decisions are made by career administrators, and how to coordinate their work with the policies of politically elected or appointive officials, remain key preoccupations, especially in democratic states.

Changes in the basic premises of public administration

It was with governmental efforts to combat the Depression that new informational devices were introduced, including national income accounting and the scrutiny of gross national product as a major index of economic health. The applied techniques of fiscal and monetary policy have become established specializations of public administration. Economists occupy key posts in the administrations of most nations, and many other adminis-

trators must have at least elementary knowledge of the economic implications of government operations. France, Sweden and other Scandinavian nations, Great Britain, and the United States were among the leaders in developing economic planning techniques. Such planning has become a dominating concern of public administration in many of the developing countries.

As economic and social intervention by governments has increased, the limitations of "incrementalism" as a public administration practice have become increasingly apparent. Incrementalism is the tendency of government to tinker with policies rather than to question the value of continuing them. A number of techniques have been introduced to make decisions more rational. One such technique, widely applied, is cost–benefit analysis. This involves identifying, quantifying, and comparing the costs and benefits of alternative proposals. Another, less successful, technique was the Planning, Programming, and Budgeting System (PPBS), introduced into the U.S. Department of Defense in 1961 and extended to the federal budget in 1965. According to PPBS, the objectives of government programs were to be identified, and then alternative means of achieving these objectives were to be compared according to their costs and benefits. In practice, PPBS made little difference in federal budgeting, partly because the objectives of governmental programs were difficult to specify and partly because comprehensive evaluation took too long. PPBS was abandoned in 1971, and similar attempts, such as Management by Objectives and Zero-Base Budgeting, both introduced in the 1970s, were equally short-lived and ineffective. Comparable schemes in western Europe, such as the method called "rationalization of budgetary choice" introduced into France in the late 1960s and the so-called Programme Analysis and Review in Great Britain in the 1970s, were likewise unsuccessful.

Difficulty of measuring social costs and benefits

Quantitative economic measurement is useful up to a certain point, but the value of human life, of freedom from sickness and pain, of safety on the streets, of clean air, and of opportunity for achievement are hardly measurable in monetary terms. Public administration has thus increasingly concerned itself with developing better social indicators, quantitative and qualitative—that is, better indexes of the effects of public programs and new techniques of social analysis.

Another development has been an increasing emphasis on human relations. This originated in the 1930s when what became known as the Hawthorne research, involving the workers and management of an industrial plant near Chicago, brought out the importance to productivity of social or informal organization, good communications, individual and group behaviour, and attitudes (as distinct from aptitudes).

Awareness of the importance of human relations influenced the conduct of public administration. Many shibboleths of administration (hierarchy, directive leadership, set duties, treatment of employees as impersonal "units" of production, and monetary incentives) were challenged.

By the late 1930s the human relations approach had developed into a concept known as "organization development." Its primary goal was to change the attitudes, values, and structures of organizations so that they could meet new demands. Trained consultants, usually from outside the organization, undertook intensive interviewing of senior and junior staff, and sensitivity training and confrontation meetings were also held. Unlike the rationalistic PPBS approach, organization development stressed the identification of personal with organizational goals, the "self-actualization" of workers and managers, effective interpersonal communication, and broad participation in decision making. Its direct use within governmental agencies has been limited and has not always been successful, but it has had considerable indirect influence upon administrators.

Another modern movement in public administration has been the greater participation of citizens in government. It was stimulated during the 1950s and '60s by a growing feeling that governments were not responding to the needs of their citizens, particularly minority groups and the poor. A variety of experiments to involve citizens or their representatives in making governmental decisions were begun in the 1960s. These involved the delegation of decision making from central to local offices and, at the local level, the sharing of authority with citizen groups.

From the early 1970s increasing analysis of the way government policies affected the public resulted in a concept called the "public policy approach" to administration. This examines to what extent each stage in devising and executing a policy affects the overall shape and impact of the policy. According to the concept, the way a problem is conceived in the first place influences the range of remedies considered. The nature of the decision-making process may determine whether a course of action is merely incremental or truly radical. Indeed, it has been argued that the nature of the decision-making process shapes the outcome of the decision itself, particularly when the process is dominated by a powerful interest group. Moreover, the willingness of the government to evaluate programs, and modify them if necessary, affects the outcome. Many supporters of the public policy approach regard the concept as an important tool for constructing a body of knowledge on which recommendations can be based.

Until World War II there was relatively little exchange among nations of ideas about public administration. As early as 1910, however, a professional organization, which eventually became the International Institute of Administrative Sciences (IIAS), had been established. At first its membership consisted principally of scholars and practitioners of administrative law in the countries of continental Europe. By the late 1980s the IIAS had a membership drawn from some 70 countries. Its triennial congresses have covered all aspects of the field.

International interest in public administration

Since World War II international interest in administrative systems has grown, precipitated by the necessity of cooperation during the war, by the formation of international organizations, by the occupation of conquered nations and the administration of economic recovery programs for Europe and the Far East, and by aid programs for developing countries. One by-product of aid programs was a renewed appreciation of how crucial effective administration is to national development. It has also become apparent how parochial and culture-bound styles of public administration have often remained within individual countries.

Another effect of this international communication and sharing of experiences has been the realization that development is not exclusive to the so-called underdeveloped countries. All countries have continued to develop, and public administration has increasingly been perceived as the administration of planned change in societies that themselves have undergone rapid change, not all of it planned. Government has no longer been merely the keeper of the peace and the provider of basic services: in the postindustrial era government has become a principal innovator, a determinant of social and economic priorities, and an entrepreneur on a major scale. On virtually every significant problem or challenge—from unemployment to clean air—people have looked to the government for solutions or assistance. The tasks of planning, organizing, coordinating, managing, and evaluating modern government have likewise become awesome in both dimension and importance.

EDUCATION AND TRAINING

Shortage of trained public administrators

European universities have traditionally produced administrative lawyers for their governments, but legal skills alone are hardly adequate for handling contemporary problems. U.S. universities began graduate programs in the early years of the 20th century, and by the late 1980s there were more than 300 university programs in public administration. Nevertheless, very few of the scientists and other specialists who become administrators in their fields attend such programs.

Training programs have particularly flourished since World War II, many of them with government help. Some are attached to universities. In establishing the École Nationale d'Administration as one of its civil service reforms of 1946–47, France provided an extensive course for re-

cruits to the higher civil service. It was not until 1969 that Britain established a Civil Service College under the new Civil Service Department. In the United States the government established a variety of educational and training programs during the 1960s, including the Federal Executive Institute and the Executive Seminar Centers. Many less-advanced countries have since established centres for the training of public administrators. (F.C.M./E.C.P.)

Organization of civil service

APPOINTMENT

In earlier times, when civil servants were part of the king's household, they were literally the monarch's personal servants. As the powers of monarchs and princes declined and as, in some countries, their sovereignty was denied them, appointment became a matter of personal choice by ministers and heads of departments. The influence senior civil servants may wield over policy and the need for them to work in close harmony with ministers induce all governments to insist on complete freedom of choice in appointments, even when, as in Great Britain, the freedom is rarely invoked. In some countries, notably the United States, senior advisers usually are replaced whenever a new administration takes office.

In Europe in the 19th century, appointment and promotion frequently depended on personal or political favour, but tenure was common in the lower and middle ranks once an appointment had been made.

Reaction against patronage

Dependency on a superior's favour led civil servants to ally themselves with liberal public opinion, which was critical of the waste and corruption involved in political patronage. Pressure for reform led to official formulations of basic qualifications for different posts; appointments and promotions boards were established within each department to prevent or obstruct overt political favouritism and nepotism; and salary scales were introduced for different grades to provide a civil servant with increments for good service while still holding the same post. In many countries civil service commissions were set up to ensure impartiality in selection procedures and to lay down broad principles for personnel management in the civil service. Recruitment in many European countries corresponded to the national educational systems: the highest class of civil servants entered service after graduation from a university, the executive class after full completion of secondary school, the clerical class after the intermediate school examination. The manual workers in the service were mainly recruited from persons of mature age who had left school after primary education or, in such countries as France and Germany, from military veterans. As public administration became more complex in the 20th century, specialized categories of civil servants were created to bring into the service doctors, scientists, architects, naval constructors, statisticians, lawyers, and so on. In several countries the establishment of these special classes caused some difficulties because their salary scales had to be linked with those of competing professional groups outside the service. The distinction between foreign service and home service personnel has sometimes caused difficulty because of inadequate liaison between overseas representatives and the makers of foreign policy at home. In the United States, the Rogers Act of 1924 unified the overseas service itself, but the civil servants of the State Department in Washington, D.C., continued to be regarded as part of the federal civil service.

The posts that fall under the rules of the U.S. merit system are not grouped into a small number of general classes but have individual job specifications and entry qualifications. Although designed to select entrants with special knowledge or skills for individual posts, this system has been criticized for failing to make the best use of the talent available to the government. In 1978 the Senior Executive Service was created to achieve more effective promotion and deployment.

All countries base appointments on some kind of competition. In some countries great emphasis is placed on formal written examinations supplemented by interviews. Such is the situation in France, where entry into the higher civil service is channeled through specialist schools, or *grandes écoles,* of which the École Nationale d'Administration and the École Polytechnique are the most important. In Great Britain, traditionally one of the great advocates of entry by formal examination, the Civil Service Commission relies more on informal tests and a series of interviews and observations and tends to measure the candidate's intellectual competence by the quality of his university degree.

Determining qualifications

The conventional written examination is dispensed with also in such European countries as Finland, Switzerland, The Netherlands, and Portugal, as well as the German *Länder,* or states. In the *Länder* the qualifications and references of all candidates are compared, whereupon the most eligible are interviewed by a departmental board. Candidates are expected to have completed a lengthy program of academic work for professional qualification and a period of subsequent training in a variety of public institutions under official supervision. If successful in their interviews, candidates are recommended to the minister, who makes appointments to higher grade posts, or to the heads of department, who handle the middle and lower categories. On the face of it, this method offers fewer guarantees of impartiality than does the formal written examination, but a civil service career is less attractive now than formerly and the civil service has to compete, usually at lower salaries, with business and the professions for the best available talent. In Sweden a constitutional provision requires that nearly all public documents (including the proceedings of authorities that make appointments) be open for public inspection, thus providing a check upon corruption or favouritism.

Most federal and culturally diverse countries try to ensure an equitable distribution of posts among their constituent elements. In Switzerland the federal authorities try to maintain a balance of posts not only between the cantons but also between the political parties, religions, and languages. The federal civil service in Germany draws on the public service officers in the *Länder,* and some degree of proportional representation is attempted. There was considerable pressure in Canada in the 1970s to ensure a more equitable distribution of federal civil service posts between the English- and French-speaking populations. It is also clear that many African states are compelled to recognize regional and tribal origins in their appointments to the civil service.

CONDITIONS OF SERVICE

The forerunners of civil servants, being members of the royal household, had duties but no rights. The first attempts to formalize methods of appointment and conditions of service were among the administrative innovations introduced in Prussia in the 18th century. Elsewhere attempts were frustrated by political and public objections. Increased formal regulation of conditions of service came about when civil servants organized themselves into professional groups, sometimes barely distinguishable from trade unions. The fact that civil servants are agents of the public power, providing services on which law, order, and public health depend, has raised the question whether they should be permitted to strike; if they cannot lawfully strike, they are deprived of the main weapon in pressing for improvements in their conditions of service. Thus, there have developed special arrangements for reviewing conditions of service periodically and for settling contentious issues. In particular, it has been necessary to have a properly recognized system for regulating conduct and discipline. In the United Kingdom, traditional standards are supplemented or revised to accord with recommendations from periodic commissions of enquiry, which pay special attention to official conduct in relation to political activities and business dealings. In France and Germany these codes of conduct have been based mainly upon the rules of administrative law and the jurisprudence of administrative courts, although certain civil service rights and duties are specified in constitutional law. In other countries, particularly in the United States and India, conduct and discipline are regulated by administrative rules and codes promulgated by executive order after discussion and enquiry.

The need for codes of conduct

The standards placed upon a civil servant's conduct are partly those to be expected of any loyal, competent, and obedient employee and partly those enjoined upon a public employee. Ideally, the civil servant should be above any suspicion of partiality and should not let personal sympathies, loyalties, or interests affect the performance of duties; for example, a civil servant is obliged to be circumspect in private financial dealings. As a general rule, a civil servant is not allowed to engage directly or indirectly in any trade or business and may engage in social or charitable organizations only if these have no connection with official duties. There are always strict limits on a civil servant's right to lend or borrow money, and they are prohibited from accepting gifts.

Political involvements of civil servants

There are different attitudes about the extent to which civil servants may engage in political activities. One view is that a civil servant has the same constitutional rights as other citizens and that it is therefore unconstitutional to attempt to limit those rights other than by the common law. The opposing view is that, since civil servants are engaged in the unique function of national government, their integrity and loyalty to their political masters might be affected by active participation in political affairs, and public confidence in their impartiality could be shaken. Broadly speaking, those countries that traditionally expect civil servants to behave with complete impartiality and to conform to ministerial policy with energy and good will, whether they agree with the policy or not, expect all civil servants to behave with circumspection in political affairs. The United Kingdom has a total ban on its senior civil servants engaging in any form of political activity. The prohibition becomes progressively less strict, however, for the medium and lower grades of the service.

Another group of countries, including France and Germany, have deemed policy and administration to be so intimately connected that all top posts are filled at the discretion of the government of the day; thus, civil servants are allowed greater scope in political activities. They are nevertheless expected to act with greater discretion and public decorum than private citizens, and an excess of power or an abuse of office for political purposes renders a civil servant instantly liable both to statutory regulations and to severe internal disciplinary proceedings.

Unions of civil servants

Traditionally, governments have been hostile toward civil service unions, and in the past repressive laws made strike action unlawful. Strikes nevertheless occurred, and governments eventually adopted an attitude of open encouragement toward trade unionism. Most governments accept, in theory at least, that the state should be a model employer. It follows that, if the state genuinely pursues a policy of discussion and negotiation with civil servants and attempts properly to fulfill agreements with them, it should in return be freed from the threat of strike action. Mindful that the withdrawal of civil servants from some public services would lead to chaos, many governments have found it prudent to establish permanent channels for negotiating such matters as salaries and discipline. Organizations representing the staff and a management side of senior officials representing the state mirror the employer–employee relationship of private industry, although a higher percentage of public- than private-sector employees are members of unions. The United Kingdom was the first country to establish negotiating machinery for civil servants. Following a report in 1917, organizations known as Whitley Councils were set up, consisting of equal numbers of medium and lower staff on the one hand and directing and supervisory staffs on the other. These councils operate within the ministries, and a National Whitley Council performs central advisory functions for the government. They have no powers of decision, however, only of recommendation, because governments are never prepared to surrender their ultimate responsibility for determining the public interest. The councils have done a good deal to provide a sense of common purpose and joint responsibility within the civil service as a whole, although pay restraints from the early 1970s generated great friction between civil service unions and government.

In France each department has a comparable consultative body, but its work is broader in scope in that it can scrutinize recruitment, personnel records, promotions, and disciplinary procedures. There is also a national council, presided over by the prime minister or a specially nominated minister for civil service affairs, which is concerned with general personnel policy, conditions of service, and coordination of departmental committees.

Until after World War II, the commonly accepted view in the United States was that expressed by President Calvin Coolidge: "There is no right to strike against the public safety by anybody, anywhere, at any time." Although federal employees are still forbidden to strike, a rule illustrated by the dismissal of striking air traffic controllers in 1981, consultation has increased, and in many federal departments appeals committees comprising departmental heads and one or more members of the Merit Systems Protection Board may now hear appeals from civil servants against decisions adversely affecting their careers. These committees are also consulted on general matters of departmental interest, such as job classifications, pension schemes, promotion policies, and office procedures.

PATTERNS OF CONTROL

The expansion of public services, as well as the development of permanent civil service career structures, raised fears that civil services were becoming autonomous powers in their own right, no longer subject to the traditional forms of control. This view is associated with the sociologists Max Weber, who criticized the bureaucracy of imperial Germany, and Robert Michels, who formulated the "iron law of oligarchy." Michels' law suggested that every organization with a permanent staff produces an oligarchy running the organization according to the interests and values of the bureaucratic group. In addition, the growing complexity of modern government has greatly augmented the informal power of senior civil servants who act as advisers to ministers. This is particularly the case in countries (usually the more democratic ones) where ministries frequently change hands.

In the 19th century civil services were normally restricted to maintaining law and order and minor economic regulations such as those concerning weights and measures and factory laws. The subordination of civil servants to their political masters and their political masters' responsibility to the courts and legislatures seemed to provide an adequate safeguard against arbitrary administrative actions. But in some countries, notably Germany, France, and Austria, civil services became endowed with much greater authority, operating as part of the police power. This caused concern because civil servants were exempt from normal legal processes when performing their official functions. In response, special administrative courts were set up to which private citizens or corporations could appeal against administrative acts. Jurisdiction was limited, however, and redress was frequently slow. The courts themselves remained specialized institutions of the executive rather than normal parts of the judiciary.

Sweden provided a marked contrast. Before the constitution of 1809 the executive power had been absolute. Afterward, not only did it become subject to control by the legislature, but this control also was reinforced by the creation of a special post of ombudsman (see below *Ombudsman*).

World War I brought increased governmental activity almost everywhere. The area in which administrative discretion could be exercised grew; civil servants became as much adjudicators as administrators, and their influence upon economic life increased. By World War II the state had become, even in many conservative countries, an economic regulator, an industrial producer of overwhelming importance, and a conciliator between competing interests. In all of these matters civil servants were the effective agents of the state.

U.S. regulatory agencies

In the United States, Congress created an institution to counter the threatened increase in civil service power. As far back as the late 19th century Congress, when legislating for new areas of government, assigned powers to agencies or commissions, specifying their powers, competence, and composition and freeing them from direct presidential control. In this way large areas of government escaped the

control of the executive branch of government, including the federal civil service. These independent regulatory agencies have covered major economic fields and have included the Interstate Commerce Commission, the Federal Communications Commission, the Tennessee Valley Authority, and the Nuclear Regulatory Commission. This policy has laid Congress open to the charge that it has created a headless fourth branch of government, but it has successfully prevented the emergence of a monolithic federal civil service.

To counter charges that the U.S. civil service was encroaching on the powers of the judiciary, the Administrative Procedure Act of 1946 laid down detailed provisions to safeguard citizens' rights where the administration had powers of adjudication. These rights included the right to ample previous notice of proceedings, the right to submit evidence, the right to have independent hearing officers (to the exclusion of investigating or prosecuting officers), and the right to a decision based solely on testimony and papers actually entered in the proceedings.

Other democratic countries have been concerned about the growing powers of the civil service and about whether traditional forms of judicial and ministerial control are adequate. Many European countries have modeled their instruments of administrative jurisdiction and jurisprudence on the French Conseil d'État. In the United Kingdom the creation of a special administrative jurisdiction of this kind has been opposed by both parliamentary and judicial opinion, but it was because of mounting criticism of civil service immunity from detailed control that Parliament created the special office of parliamentary commissioner, or ombudsman. Public access to the office is by way of a member of Parliament, and the commissioner is excluded from inquiring into matters of policy, local government authorities, or lower judicial bodies.

Communist controls over the civil service

Special problems of control arose in Communist countries, where the main preoccupation of the regime, which was under the direct control of the Communist Party, had been to ensure the civil service's continual loyalty. Impartiality and objectivity in the administrative machine's dealings with the public were not of such high priority as in pluralist societies. A body of administrative procedure was built up, but this was always subordinated to the directives of the party leadership. Communist countries also had to establish new ways of judging performance, since the state monopoly of political power and means of production ensured that traditional incentives and yardsticks could not be applied.

Yet in their own way, Communist countries had elaborate controls. In the Soviet Union all ministries had a special section staffed by, and responsible to, operatives from the Ministry of Internal Affairs. This section provided security control over the ordinary civil servants, and its personnel were not part of that ministry's official structure. The Communist Party maintained further control through the party apparatus, and it closely supervised senior appointments.

Communist planning, financial, and personnel controls of a technical kind resembled those in democratic countries, but in the Soviet Union there were two additional special supervisory agencies. The Commission of State Control was responsible for vigilance over state property and administration. Its departments paralleled the different branches of state administration and maintained audits of their work. Its officers had the right of access to all administrative records and could issue directives to other institutions. They had powers to prosecute civil servants for criminal offenses, and they could apply a formidable range of disciplinary measures to civil servants, either by direct action or through the responsible minister.

The second agency of control arose because of the difficulty of reconciling disputes between production units and their controlling ministries in an economy that lacked the traditional forms of market discipline and could not rely upon an enforceable law of contract. A special system of compulsory arbitration operated through the State Arbitration Tribunal (known as Gosarbitrazh) under the Council of Ministers and through arbitration tribunals responsible to the councils of ministers in each of the republics. It settled all disputes concerning contracts, quality of goods, and other property disputes between various state enterprises. The system was staffed by civil servants charged with enforcing "contractual and plan discipline," but it was supported by technical experts qualified in economic and industrial matters. (B.Ch./E.C.P.)

INTERNATIONAL CIVIL SERVICE

The elements of an international civil service first appeared in the Universal Postal Union (established 1874–75). Some four and a half decades later the League of Nations and the International Labour Organization (ILO) were founded. They required a staff of almost 600 experts and subordinate personnel, which took the form of a true international civil service. It drew mainly on British, French, and Swiss personnel, but more than 40 states contributed members in order to spread recruitment as widely as possible. There were no formal methods of selection for the higher personnel; the secretary general of the League used personal contacts. The staff fell into three divisions: administrative, supervisory and clerical, and custodial. The history of the League shows that, at any rate as far as a secretariat was concerned, a broad measure of international loyalty could be achieved. The staff of the ILO was maintained after the League's disbandment in 1946.

Until World War II the League of Nations was the only major international organization of its kind. Since then, international organizations have multiplied, and compared with most of them the League had a small staff. The postwar organizations include the United Nations, the Organization of American States, the European Community (EC), and the Organisation for Economic Co-operation and Development. By the mid-1980s there were nearly 20,000 smaller bodies, ranging from the Nordic Council to small research institutes.

Two types of international civil service

International organizations have evolved into two general types. In the first—an example of which is the EC—a genuinely supranational civil service exists; its members have a career structure within the organization and can identify with it. In the second (the UN, for instance), officials undertake a particular job for a limited period and seldom develop a career within the organization. In the EC the civil service is divided into four major grades, recruitment to the service generally being through competitive examination. After recruitment a period of training follows. Each agency within the UN operates its own recruitment system. By far the majority of senior appointees assigned to the UN at any one time have served there for less than 15 years.

In both types of international organization, there is a danger that officials will promote the interests of their own country to the detriment of internationalism. In the UN a quota system operates by which each country is allocated a number of appointments in proportion to its UN budget contribution. As a result, some national governments are in a strong position to influence recruitment. In contrast, the statutes governing the EC's civil service forbid quotas, but an unofficial system has developed whereby a relatively high proportion of senior officials are from the more populous states. Such practices risk sacrificing meritocracy to nationalism.

Some countries do expect their nationals working at the UN to promote national interests. But the quota system—and this is one of the arguments for it—leads to administrators from at least the more prosperous countries competing with fellow nationals rather than with colleagues from other countries. Moreover, it is unlikely that decisions, whether biased or not, will be made in the bureaucracy alone. And since officials' national origins are no secret, partiality is fairly difficult to conceal. (E.B./E.C.P.)

Administrative law

Administrative law is the legal framework within which public administration is carried out. It derives from the need to create and develop a system of public administration under law, a concept that may be compared with the much older notion of justice under law. Since administra-

tion involves the exercise of power by the executive arm of government, administrative law is of constitutional and political, as well as juridical, importance.

The regulation of public administration

There is no universally accepted definition of administrative law, but rationally it may be held to cover the organization, powers, duties, and functions of public authorities of all kinds engaged in administration; their relations with one another and with citizens and non-governmental bodies; legal methods of controlling public administration; and the rights and liabilities of officials. Administrative law is to a large extent complemented by constitutional law, and the line between them is hard to draw. The organization of a national legislature, the structure of the courts, the characteristics of a cabinet, and the role of the head of state are generally regarded as matters of constitutional law, whereas the substantive and procedural provisions relating to central and local governments and judicial review of administration are reckoned matters of administrative law. But some matters, such as the responsibility of ministers, cannot be exclusively assigned to either administrative or constitutional law. Some French and American jurists regard administrative law as including parts of constitutional law.

The law relating to public health, education, housing, and other public services could logically be regarded as part of the corpus of administrative law; but because of its sheer bulk it is usually considered ancillary.

DEFINING PRINCIPLES

One of the principal objects of administrative law is to ensure efficient, economical, and just administration. A system of administrative law that impedes or frustrates administration would clearly be bad, and so, too, would be a system that results in injustice to the individual. But to judge whether administrative law helps or hinders effective administration or works in such a way as to deny justice to the individual involves an examination of the ends that public administration is supposed to serve, as well as the means that it employs.

In this connection only the broadest generalities can be attempted. It can be asserted that all states, irrespective of their economic and political system or of their stage of development, are seeking to achieve a high rate of economic growth and a higher average income per person. They are all pursuing the goals of modernization, urbanization, and industrialization. They are all trying to provide the major social services, especially education and public health, at as high a standard as possible. The level of popular expectation is much higher than in former ages. The government is expected not only to maintain order but also to achieve progress. There is a widespread belief that wise and well-directed government action can abolish poverty, prevent severe unemployment, raise the standard of living of the nation, and bring about rapid social development. People in all countries are far more aware than their forefathers were of the impact of government on their daily lives and of its potential for good and evil.

The growth in the functions of the state is to be found in the more-developed and in the less-developed countries; in both old and new states; in democratic, authoritarian, and totalitarian regimes; in the Communist countries of eastern Europe and in the mixed economies of the West. The movement is far from having reached its zenith. With each addition to the functions of the state, additional powers have been acquired by the administrative organs concerned, which may be central ministries, local, provincial, or regional governments, or special agencies created for a particular purpose.

Distinctions between public administration and private action. Activities such as traffic control, fire-protection services, policing, smoke abatement, the construction or repair of highways, the provision of currency, town and country planning, and the collection of customs and excise duties are usually carried out by governments, whose executive organs are assumed to represent the collective will of the community and to be acting for the common good. It is for this reason that they are given powers not normally conferred on private persons. They may be authorized to infringe citizens' property rights and restrict their freedom of action in many different ways, ranging from the quarantining of infectious persons to the instituting of criminal proceedings for nonpayment of taxes. To take another example, the postal laws of many countries favour the post office at the expense of the customer in a way unknown where common carriers are concerned. Again, a public authority involved in slum clearance or housing construction tends to be in a much stronger legal position than a private developer.

The result of the distinction between public administration and private action is that administrative law is quite different from private law regulating the actions, interests, and obligations of private persons. Civil servants do not generally serve under a contract of employment but have a special status. Taxes are not debts, nor are they governed by the law relating to the recovery of debts by private persons. In addition, relations between one executive organ and another, and between an executive organ and the public, are usually regulated by compulsory or permissive powers conferred upon the executive organs by the legislature.

Internal and external aspects of administrative law

The law regulating the internal aspects of administration (*e.g.*, relations between the government and its officials, a local authority and its committees, or a central department and a local authority) differs from that covering external relations (those between the administration and private persons or interests). In practice, internal and external aspects are often linked, and legal provisions of both kinds exist side by side in the same statute. Thus, a law dealing with education may modify the administrative organization of the education service and also regulate the relations between parents and the school authorities.

Another distinction exists between a command addressed by legislation to the citizen, requiring him to act or to refrain from acting in a certain way, and a direction addressed to the administrative authorities. When an administrative act takes the form of an unconditional command addressed to the citizen, a fine or penalty is usually attached for failure to comply. In some countries the enforcement is entrusted to the criminal courts, which can review the administrative act; in others the administrative act itself must be challenged in an administrative court.

The need for legal safeguards over public administration. Statutory directions addressed to the executive authorities may impose absolute duties, or they may confer discretionary powers authorizing a specified action in certain circumstances. Such legislation may give general directions for such activities as factory inspection, slum clearance, or town planning. The statute lays down the conditions under which it is lawful for the administration to act and confers on the authorities the appropriate powers, many of which involve a large element of discretion. Here the executive is not confined merely to carrying out the directions of the legislature; often it also shares in the lawmaking process by being empowered to issue regulations or ordinances dealing with matters not regulated by the statute. This may be regarded either as part of the ordinary process by which the legislature delegates its powers or as an inevitable feature of modern government, given that many matters are too technical, detailed, or subject to frequent change to be included in the main body of legislation—legislation being less easy to change than regulations.

Problems of executive power

Whatever the source of the executive's rule-making power, safeguards against misuse are necessary. For instance, the regulation must not exceed the delegated powers; its provisions must conform with the aims of the parent statute; prior consultation with interests likely to be affected should take place whenever practicable; and the regulations must not contravene relevant constitutional rules and legal standards. In some countries regulations are scrutinized by a type of watchdog known as the council of state before they come into force; in others, by the parliamentary assembly; and in yet others, by the ordinary courts.

In most countries the executive arm of government possesses certain powers not derived from legislation, customary law, or a written constitution. In the United Kingdom there are prerogative powers of the crown, nearly all of which are now exercised by ministers and which concern

such matters as making treaties, declaring war and peace, pardoning criminals, issuing passports, and conferring honours. In Italy, France, Belgium, and other continental European countries, certain acts concerning the higher interests of the state are recognized as *actes de gouvernement* and are thereby immune from control by any court or administrative tribunal. In the German Empire (1871–1918) the principle that an administrative act carried its own legal validity was accepted at the end of the 19th century by leading jurists. This led to the doctrine that administration was only loosely bound to the law. The doctrine was rejected in the Federal Republic of Germany (1949), however, and efforts were made to reduce the area in which the executive was free to act outside administrative law.

Bureaucracy and the role of administrative law. An inevitable consequence of the expansion of governmental functions has been the rise of bureaucracy. The number of officials of all kinds has greatly increased, and so too have the material resources allocated to their activities, while their powers have been enlarged in scope and depth. The rise of bureaucracy has occurred in countries ruled by all types of government, including the Communist countries, the dictatorships and Fascist regimes, and the political democracies. It is as conspicuous in the new states of Africa and Asia as among the highly developed countries of western Europe or North America. A large, strong, and well-trained civil service is essential in a modern state, irrespective of the political character of its regime or the nature of its economy.

Fear of the maladies that tend to afflict bureaucracy has produced a considerable volume of protest in some countries; and, even in those where opposition to the government or the party in power is not permitted, criticism and exposure of bureaucratic maladministration are generally encouraged.

Bureaucratic maladies are of different kinds. They include an overdevotion of officials to precedent, remoteness from the rest of the community, inaccessibility, arrogance in dealing with the general public, ineffective organization, waste of labour, procrastination, an excessive sense of self-importance, indifference to the feelings or convenience of citizens, an obsession with the binding authority of departmental decisions, inflexibility, abuse of power, and reluctance to admit error. Many of these defects can be prevented or cured by the application of good management techniques and by the careful training of personnel. A whole range of techniques is available for this purpose, including effective public relations, work-study programs, organization and management, operational research, and social surveys.

Distinction between political and legal control

Administrative law is valuable in controlling the bureaucracy. Under liberal-democratic systems of government, political and judicial control of administration are regarded as complementary, but distinct. The former is concerned with questions of policy and the responsibility of the executive for administration and expenditure. The latter is concerned with inquiring into particular cases of complaint. Administrative law does not include the control of policy by ministers or the head of state. Under Communist regimes, however, this distinction rarely exists. The control exercised by an elected council or a presidium over a similar body at the next lower level of government is regarded as a form of legal control over administration. Internal methods of control are also regarded as falling within the ambit of administrative law. These include an appeal from the decision of an official within the same organization or an appeal to a higher administrative unit. A distinctive feature of Communist countries is that there is no definition of the powers of governmental organs at different levels. They are all assumed to be unlimited in scope but always subject to an equally unlimited right of intervention and restraint by the corresponding organ at the next higher level of government. This contrasts with the explicit definition of powers at each level of government that is found in the Western-type democracies. The admixture of political and legal control results in administrative law having a loose and imprecise meaning in Communist countries.

JUDICIAL REVIEW OF ADMINISTRATION

Judicial review of administration is, in a sense, the heart of administrative law. It is certainly the most appropriate method of inquiring into the legal competence of a public authority. The aspects of an official decision or an administrative act that may be scrutinized by the judicial process are the competence of the public authority, the extent of a public authority's legal powers, the adequacy and fairness of the procedure, the evidence considered in arriving at the administrative decision and the motives underlying it, and the nature and scope of the discretionary power. An administrative act or decision can be invalidated on any of these grounds if the reviewing court or tribunal has a sufficiently wide jurisdiction. There is also the question of responsibility for damage caused by the public authority in the performance of its functions. Judicial review is less effective as a method of inquiring into the wisdom, expediency, or reasonableness of administrative acts, and courts and tribunals are unwilling to substitute their own decisions for that of the responsible authority.

Judicial review of administration varies internationally. Sweden and France, for instance, have gone as far as subjecting the exercise of all discretionary powers, other than those relating to foreign affairs and defense, to judicial review and potential limitation. Elsewhere, a preoccupation with procedure results in judicial review deciding only whether the correct procedure was observed rather than examining the substance of the decision.

It is of course impractical to subject every administrative act or decision to investigation, for this would entail unacceptable delay. The complainant must, therefore, always make out a prima facie case that maladministration has occurred.

Judicial review cannot compel the state to act in a particular way because the courts concerned cannot impose sanctions on the government, which itself controls the use of force. Such remedies as an injunction, an order for specific performance, or an order for mandamus will not lie against the central government. These inhibitions, however, are of less practical importance than might be supposed. Nevertheless, nearly all governments (even revolutionary ones) are eager to proclaim the lawfulness of the regime and seldom disregard the decisions of an authorized court or tribunal.

In judicial review of administration at a national level, a country's history, politics, and constitutional theory all play their part. There are, broadly, three major systems: the common-law model; the French, or council of state, model; and the procurator model.

The common-law system. *Origins.* The common-law system originated in England in the Middle Ages. In the 17th century relations between the courts and the executive developed into a constitutional struggle between the Stuart kings and the judges over the judges' right to decide questions affecting the royal power and even to pronounce an independent judgment in cases in which the king had an interest. Francis Bacon, in his essay *Of Judicature* (written in 1612), put forth the royalist point of view when he declared that the judges should be "lions, but yet lions under the throne." "It is a happy thing in a state," he wrote, "when kings and states do often consult with judges; and again, when judges do often consult with the king and state: the one, when there is matter of law intervenient in business of state; the other, when there is some consideration of state intervenient in matter of law." The subordination of the judicature to the royal will was strongly resisted by Chief Justice Sir Edward Coke, Bacon's great rival, who refused to comply with James I's wishes in a number of cases in which the royal prerogative was involved. The King harangued the judges more than once on their duty to respect the royal prerogative and power.

Independence of the judiciary

In the constitutional conflict that took place a generation later, the judges and the lawyers made common cause with Parliament against Charles I, and eventually the independence of the judges was established. Henceforth there was to be one system of law to which all would owe obedience. As a result, the executive possessed no inherent powers other than those subject to the rule of law inasmuch as legislation now had to emanate from the crown

in Parliament. In addition, the judges were expected to protect the subject against the executive. A more intangible consequence was the belief that "government" and "law" were often thought to be opposed to one another. The earlier conflict between crown and judges survived to become an antagonism between the legal profession and the executive, particularly the civil service.

These developments established the principle that the executive should never interfere with the judiciary in the exercise of its functions. This was, indeed, almost the only strict application in England of the doctrine of the separation of powers. On the other hand, it was regarded as right and proper that the judiciary should interfere with the executive whenever a minister or a department was shown to have acted illegally. In this way the concept of the rule of law came gradually to be identified with the idea that the judges, in ordinary legal proceedings in the ordinary courts, could pronounce upon the lawfulness of the activities of the executive. Any attempt to divide the seamless web of the law, any suggestion of a distinction between public and private law, appeared destructive of the law's universality and its power to keep the executive within bounds.

The principle that all public authorities are liable to have the lawfulness of their acts and decisions tested in the ordinary courts was applied everywhere the common law prevailed, including the United States, despite the much stricter interpretation given by the Founding Fathers there to the doctrine of the separation of powers—a doctrine embodied in the federal and state constitutions. A complete separation of powers was not considered feasible by the framers of the Constitution, and they therefore introduced checks and balances, whereby each of the three branches of government would be prevented from growing too powerful by the countervailing power of the others. This actually strengthened the power of the courts to review the actions of the executive. Elsewhere in the common-law world, the extended role of the courts in reviewing administration was adopted without any public debate concerning the separation of powers or the need to protect liberty by a system of checks and balances. This absence of an explicitly defined role for courts led, in the early post-World War II years in Britain, to real fears that the courts would be unable or unwilling to question the expanded powers of governmental bodies.

American checks and balances

Modification of the common-law system. The common-law system has been extensively modified in the course of the 20th century. Until recently it did not correspond to the realities of the situation in Britain because, prior to the Crown Proceedings Act, 1947, it was not possible to sue ministers and their departments in tort; government ministers in Britain are considered ministers of the crown, and an ancient legal doctrine holds that "the king can do no wrong." Moreover, the development of state-provided social services has been accompanied by the creation of a large number of administrative tribunals to determine disputes between a government department and a citizen. The jurisdiction of these tribunals is of a specialized and narrowly circumscribed character and relates to such functions as social insurance and social assistance, the National Health Service, rent control, assessment of property for local taxation, the compulsory acquisition of land by public authorities, and the registration of children's homes. Since 1958 a permanent Council on Tribunals appointed by the lord chancellor has exercised a general supervision over about 40 tribunal systems, but they remain an unsystematic and uncoordinated movement. However, they provide a method of administrative adjudication far cheaper, more informal, and more rapid than that offered by the courts; the members are persons possessing special knowledge and experience of the subject dealt with; they do not have to follow the strict and complex rules of evidence that prevail in the courts; and it is possible to introduce new social standards and moral considerations to guide their decisions. These tribunals have won general approval for the quality and impartiality of their work. An appeal on a question of law lies in most instances from the decision of an administrative tribunal to the High Court of Justice. There is still no comprehensive administrative jurisdiction in Britain permitting judicial review over the whole field of executive action and decision.

British administrative tribunals

In Australia a similar movement took place with the growth of a large number of administrative tribunals that regulate many different spheres of public administration, such as industrial conditions; the award of pensions, allowances, and other state grants; town planning; censorship of films; fair rents; the licensing of occupations calling for special skills or public responsibility; trade, transport, and marketing; the assessment of national taxes, local taxes, or duties; the protection of industrial design, patents, and copyrights; and compensation for interference with private-property rights in the public interest. As in Britain, the growth of these tribunals has been sporadic, and no attempt has been made to introduce a systematic or comprehensive administrative jurisdiction. A similar situation has developed in other Commonwealth countries.

In the United States the courts review administration much more comprehensively than in Britain. Nevertheless, much adjudication is now performed by public authorities other than the courts of law. The movement toward administrative tribunals began with the Interstate Commerce Act (1887), establishing the Interstate Commerce Commission to regulate railways and other carriers. This law introduced a new type of federal agency, outside the framework of the executive departments and largely independent of the president. Other regulatory commissions followed: the Federal Trade Commission, the Federal Communications Commission, the Securities and Exchange Commission, the National Labor Relations Board, and the Occupational Safety and Health Administration. These bodies have had administrative, legislative, and judicial functions delegated to them by Congress, and the doctrine of the separation of powers can no longer be successfully invoked to challenge the constitutionality of such legislation. The regulatory commissions are often described by American jurists as administrative tribunals.

American regulatory commissions

Thus, in the United States, as in other parts of the Anglo-American common-law world, the concept of the exclusive exercise by the ordinary courts of all judicial powers and of the absence of special administrative tribunals has been substantially modified by these developments.

The council of state system. *The French system.* In France the separation of powers was given a place of honour in the Declaration of the Rights of Man and of the Citizen (1789). In the French view, however, if a court were permitted to review an administrative act or decision, it would contravene the separation of powers as much as if the executive could override the decision of a court. Just as an appeal from a court lies to a higher court, the reasoning goes, so an appeal from an administrative authority should lie to a higher administrative authority. Only thus would the true separation of powers be observed.

French conception of separation of powers

Herein lies the explanation of administrative law as a system of law separate from the body of law administered in the courts. A law of August 1790 declared that the judiciary was distinct from and would always remain separated from the executive. It forbade judges, on pain of dismissal, to interfere in any way with the work of administrative bodies. In October 1790 a second law stated that under no circumstances should claims to annul acts of administrative bodies fall within the jurisdiction of the courts. Such claims should be brought before the king as head of the general administration.

The Conseil du Roi of the ancien régime, with its functions as legal adviser and administrative court, is generally considered to be the precursor of the Conseil d'État. The basic structure of the Conseil d'État was laid down by Napoleon, however. Among the functions accorded to it by the constitution of the year VIII (December 1799) was that of adjudicating in conflicts that might arise between the administration and the courts. It was also empowered to adjudicate any matters previously left to the minister's discretion that ought to be the subject of judicial decision. In 1806 a decree created a Judicial Committee of the Conseil to examine applications and report thereon to the General Assembly of the Conseil. These enactments laid the foundation of an administrative jurisdiction that was not clearly established until May 24, 1872, when a

law delegated to the Conseil d'État the judicial power to make binding decisions and recognized the Conseil as the court in which claims against the administration should be brought.

The Conseil d'État is and always has been part of the administration. It has for long had the task of giving legal advice to the government on bills, regulations, decrees, and administrative questions. It is this that long led foreign jurists into believing that, when sitting as a court, its decisions would inevitably be biased in favour of the executive. Nothing could be further from the truth, and today the Conseil is universally recognized as an independent court that provides French citizens with exceptionally good protection against maladministration. Suits that are directed against the French administration are heard in the Section du Contentieux, or Judicial Division, the successor of the Judicial Committee after restructuring in 1872.

French system of administrative courts

The Conseil d'État is the final authority in administrative disputes. Owing to the immense volume of work falling on it, the former prefectural councils, which served as administrative courts subordinate to the Conseil d'État, were transformed in 1953 into administrative tribunals of first instance, and the professional qualifications and career prospects of their members were improved. The great majority of cases go before these tribunals, and the Conseil d'État is the court of first and last instance only in those exceedingly rare cases when it is specially designated for that purpose.

If difficulty or doubt arises as to whether a case falls within the administrative jurisdiction or that of the ordinary courts, the question is resolved by the Tribunal des Conflits. This is a court specially established for the purpose, consisting of five judges from the Cour de Cassation (the highest civil court) and five from the Conseil d'État. The minister of justice, in his capacity as keeper of the seals (*garde des sceaux*), may sometimes preside and cast a tie-breaking vote.

Several other countries have followed France in establishing councils of state. Among them are Italy, Greece, Belgium, Spain, Turkey, Portugal, and Egypt. It must be stated, however, that in no other country has a council of state acquired such high status, powers, authority, or prestige as in France.

The German system. Germany traditionally has had no council of state, but it does have a fully articulated system of special administrative courts. In the states, or *Länder,* there are lower administrative courts and superior administrative courts, and for the federation there is the Federal Administrative Court, which acts mainly as a court of appeals from the superior administrative courts in the *Länder* and even from the lower administrative courts in certain circumstances. The Federal Administrative Court serves also as a court of first and last instance in disputes not involving questions of constitutionality between the federation and the *Länder* or between two or more *Länder;* it hears petitions by the federal Cabinet on declarations that an association is prohibited under the Basic Law of the Federal Republic, petitions against the federation in matters concerning the diplomatic or consular service, and cases concerning the business of the Federal Intelligence Service.

German system of administrative and civil jurisdictions

A *Land* administrative court possesses jurisdiction concerning the acts of the *Länder* administrative authorities and also complaints against officers of the federal government located in the *Länder.* Some of the highest federal organs are exempt from the *Länder* courts. Few cases go beyond the *Länder* supreme administrative courts.

Recourse to an administrative court is available for public law disputes unless the matter has been assigned to another court by federal legislation. (Public law governs the relationship between the state and executive in the exercise of their governmental authority and the individual—insofar as the relationship is not commercial.) The Administrative Courts Code holds that property claims arising from services for the common good and restitution claims arising from violation of duties under public law shall be heard by the ordinary courts. In other words, the German system is complicated by the rule that only the ordinary civil courts can award damages against an official or the executive arm of government. As a consequence, the distinction between the ordinary courts and the administrative courts depends on the remedy sought and not on the subject matter of the dispute or the nature of the parties. The jurisdiction of the administrative courts in Germany is therefore less comprehensive and clear-cut than in France.

The procurator system. The third system for ensuring administrative legality, called the Procuracy, was founded in Russia by Peter the Great in 1722, who intended it to be the "eye of the tsar." Catherine II issued a directive in 1764 stating that the procurator general and his staff were to supervise the execution of the laws in the provinces, ensure justice, and prevent abuses. This was designated general supervision. In 1864, however, the Procuracy was relieved of its responsibility for supervising administration, its functions being confined to judicial matters, such as acting as public prosecutor in all criminal cases and conducting them on behalf of the government and the law.

The Procuracy was abolished in November 1917 but revived in 1922. The Soviet constitution charged the procurator general with the general duty of supervising the observance of the law by all ministries and institutions subordinate to them as well as by individual officials and citizens. The procurator general was appointed by the Supreme Soviet for five years. He appointed subordinate procurators at all administrative levels, from union republic to district and town.

Disciplinary functions of the Soviet procurator

The functions actually performed by the procurator underwent many fluctuations and vicissitudes after 1922. The role as a public prosecutor continued, but it was exercised mainly to enforce party policies or programs on recalcitrant citizens rather than to punish public officers or authorities for breaches of the law. Sometimes the task of general supervision was emphasized; at other times it was abandoned in favour of more urgent tasks, as in World War II. Finally, the position of the Procuracy was laid down in a union law of May 24, 1955, and a decree of the Presidium of the Supreme Soviet made on April 7, 1956.

The procurator was not the president of a court or a tribunal but rather a watchdog of legality. His organization comprised a department for general supervision; a bureau of investigation for the supervision of preliminary inquiries in criminal matters; a department for the supervision of investigations carried out by the KGB (Committee for State Security); departments to supervise criminal and civil proceedings in the courts; a department to supervise prisons, compulsory-labour centres, and the like; and departments for statistics, administration, and research.

General supervision was defined by Soviet writers on administrative law as meaning supervision by the procurators over legality in administration. Procurators were expected to see that the laws were strictly observed, to oppose their violation by anyone whatsoever, to protect the citizens, and to ensure that they fulfilled their duties. The law of May 24, 1955, required the Procuracy to ensure that the regulations or decisions issued by ministries, departments, their subordinate establishments and enterprises, and cooperative and other public organizations strictly conformed to the constitution and laws of the Soviet Union and the republics as well as to the decrees of the Council of Ministers of the Soviet Union and the republics. They were also to ensure strict execution of the laws by officials and citizens. The procurator was concerned solely with the legality of administrative action.

After the mid-1960s a change took place in the role of the procurator from the handling of complaints by citizens against government bodies to the handling of "economic" complaints (*e.g.,* violations such as filing fraudulent plan fulfillment records) and the monitoring of economic performance. The decline in the number of individual citizens' grievances handled by the Procuracy, as well as the economic and coercive purposes for which it was used, make direct analogy between the procurator and an ombudsman misleading.

Since the Procuracy was not a court, it could not make a binding decision. This point was emphasized by Article 58 of the 1977 constitution, giving citizens the right to take complaints against administrative actions to the courts.

The normal procedure (apart from cases of dereliction involving a criminal prosecution) was for the procurator to protest against any illegality that he detected or that was brought to his notice or to initiate disciplinary action against an erring official. Every citizen had a right to lodge a complaint with the procurator, and denunciation and exposure of bureaucratic abuses were officially encouraged by the Communist Party. The fact that the Procuracy could not make a binding decision did not necessarily prevent it from being an effective organ for securing administration according to law. Neither the French Conseil d'État nor the courts in any country can enforce judgments against the central government, but this does not prevent the decisions of the Conseil or declaratory judgments of the courts from being observed almost as a matter of course.

The former Communist regimes of eastern Europe established procuracies based on the Soviet model. In Poland an additional institution to maintain administrative legality was the Supreme Chamber of Control, which was independent of the government and subordinate only to the legislature and the Council of State, a political body quite different from the French model. The functions of the Supreme Chamber of Control involved exercising general supervision over public administration and took into account legality, economy, and opportuneness.

THE OMBUDSMAN

The ombudsman is a part of the system of administrative law for scrutinizing the work of the executive. He is the appointee not of the executive but of the legislature. The ombudsman enjoys a large measure of independence and personal responsibility and is primarily a guardian of correct behaviour. His function is to safeguard the interests of citizens by ensuring administration according to law, discovering instances of maladministration, and eliminating defects in administration. Methods of enforcement include bringing pressure to bear on the responsible authority, publicizing a refusal to rectify injustice or a defective administrative practice, bringing the matter to the attention of the legislature, and instigating a criminal prosecution or disciplinary action.

When Sweden created the office of ombudsman in the constitution of 1809, the holder of that office was occupied with civil affairs and was appointed by the legislature. He was independent of both executive and judiciary and had full powers to inquire into the details of any administrative or executive act and into certain judicial activities if reported to him by individuals as an abuse of rights. He had effective authority to prosecute civil servants and other public officials—including, on occasion, ministers themselves.

Independence of the Swedish ombudsman

The Swedish ombudsman's responsibility now comprises civil affairs, including the judicature, the police, prisons, and the public administration, both central and local, but excluding ministers and the monarch. He can act as a public prosecutor (although he does not often do so); as a receiver of complaints from aggrieved citizens; or as an inspector of such institutions as jails, mental hospitals, homes for delinquent children, and retreats for alcoholics to discover if they are being administered in accordance with the law.

The institution of ombudsman was first adopted in other Scandinavian countries and then—especially from the 1960s—in many countries throughout the world, including New Zealand (1962), the United Kingdom (1967), Israel (1971), Portugal (1976), The Netherlands (1981), and Spain (1981). Australia, the United States, and Canada have ombudsmen at the state or provincial level, and in the United States several cities have municipal ombudsmen. In Britain there is an ombudsman to investigate complaints against local government, the National Health Service, and administration in Northern Ireland, in addition to the ombudsman operating at the national level. Some specialized ombudsmen have been appointed in the United States to safeguard the rights of prisoners to medical treatment. In Israel the police have an office of public complaints, and there is a military ombudsman; there is also a state controller, who issues annual reports on executive procedures.

There is no doubt about the value of the ombudsman in the states in which the institution has been established. Part of the ombudsman's usefulness lies in his ability to reassure citizens who believe they have been unjustly treated that careful inquiry into their complaints shows their suspicions to be groundless. In most countries the ombudsman has little positive power other than the right to inspect and to demand the fullest information. He may, however, recommend a particular interpretation of, or a particular modification of, the law. He can also recommend that the government pay compensation to a complainant.

ADMINISTRATIVE PROCEDURE

An orderly procedure, besides being efficient, allows responsibility to be fixed on a particular officer or body at each stage of the administrative process. It can safeguard the rights of citizens and protect the executive against the criticism of having acted in an arbitrary manner. It can ensure regularity and consistency in the handling of individual cases. Much depends, however, on the quality and purpose of the procedural requirements. Most countries possess only an uncodified mass of administrative law prescribing procedure. Much of it is to be found in the laws and regulations governing particular functions of government, such as taxation, public health, education, and town planning.

Scope of procedural rules

Rules of administrative procedure cover such matters as the setting of administrative machinery in motion; methods for lodging appeals; the rights of interested persons; the time limits that must be observed; the conditions to be satisfied by objectors; and the right of legal representation. The leading treatise on U.S. administrative law devotes many chapters to such procedural topics as rule making, requirement of opportunity to cross-examine and rebut, adjudication procedure, examiners, bias, evidence, official notice, findings, reasons, and opinions.

Some countries have a general code of administrative procedure embodied in legislation. Among them are Austria, Poland, Spain, and the United States.

In common-law systems, the doctrine of natural justice influences administrative procedure in two ways: (1) that a person may not be judge of his own cause, and (2) that a person shall not be dealt with to his material disadvantage, whether of person or property, or removed from or disqualified for office, without being given adequate notice of what is alleged against him and an opportunity to defend himself.

An indirect result of the second principle is the public hearing, widely used by government departments (and in the United States by regulatory commissions) in deciding matters involving individual or corporate rights. In the United Kingdom a public inquiry is now a common means of handling appeals to the Department of the Environment against the decisions of local authorities in such matters as planning applications and compulsory purchase of land.

In 1957 the Franks Committee was appointed by the British lord chancellor to study administrative tribunals and such procedures as the holding of a public inquiry. The committee declared that the work of administrative tribunals and of public inquiries should be characterized by openness, fairness, and impartiality, and their report applied these aims in great detail. The recommendations of the committee were largely accepted and resulted in the Tribunals and Enquiries Act of 1958.

(W.A.R./E.C.P./Ed.)

BIBLIOGRAPHY. The most comprehensive treatment of administration is ANDREW DUNSIRE, *Administration: The Word and the Science* (1973, reprinted 1981). Included among the classics in the field are FREDERICK WINSLOW TAYLOR, *The Principles of Scientific Management* (1911), available in numerous later editions; HENRI FAYOL, *General and Industrial Management*, rev. ed. (1984; originally published in French, 1917); and MAX WEBER, *Economy and Society: An Outline of Interpretive Sociology*, edited by GUENTHER ROTH and CLAUS WITTICH, 2 vol. (1978; originally published in German, 4th rev. ed., 1956). MARSHALL W. MEYER, *Change in Public Bureaucracies* (1979), is an important quantitative study of the process of change; and E.N. GLADDEN, *A History of Public Administration*, 2 vol.

(1972), is an informative survey of developments from the 11th century to the present day. For further study, useful information can be found in JAY M. SHAFRITZ, *The Facts on File Dictionary of Public Administration* (1985); and also in ROBERT D. MIEWALD, *The Bureaucratic State: An Annotated Bibliography* (1984).

The traditional approach to public administration and its principles are set forth in LUTHER H. GULICK and L. URWICK (eds.), *Papers on the Science of Administration* (1937, reprinted 1987); and L. URWICK, *The Elements of Administration*, 2nd ed. (1947). Challenges to the principles, as well as efforts to build a theory of decision making as central to administration, appear in CHESTER I. BARNARD, *The Functions of the Executive* (1938, reprinted 1979); HERBERT A. SIMON, *Administrative Behaviour: A Study of Decision-Making Processes in Administrative Organization*, 2nd ed. (1957, reissued 1965); and HERBERT A. SIMON, DONALD W. SMITHBURG, and VICTOR A. THOMPSON, *Public Administration* (1950, reprinted 1971). A thoughtful review of the evolution of public administration in its relation to society is provided in DWIGHT WALDO, *The Administrative State: A Study of the Political Theory of American Public Administration*, 2nd ed. (1984). A challenge to the traditional dichotomy between policy and administration is expressed cogently in various works of PAUL H. APPLEBY, most notably in his *Policy and Administration* (1949, reprinted 1975). Later developments of similar views are found in DAVID B. TRUMAN, *The Governmental Process: Political Interests and Public Opinion*, 2nd ed. (1971); EMMETTE S. REDFORD, *Democracy in the Administrative State* (1969); and HAROLD SEIDMAN and ROBERT GILMOUR, *Politics, Position and Power: From the Positive to the Regulatory State*, 4th ed. (1986). The nature and role of cost–benefit analysis is discussed in PETER SELF, *Econocrats and the Policy Process: The Politics and Philosophy of Cost-Benefit Analysis* (1975). The incremental approach to decision making is set out in CHARLES E. LINDBLOM, *The Intelligence of Democracy: Decision Making Through Mutual Adjustment* (1965); the problems involved in applying techniques such as PPBS and Programme Analysis and Review are discussed in AARON WILDAVSKY, *The Politics of the Budgetary Process*, 4th ed. (1984); and ANDREW GRAY and WILLIAM I. JENKINS, *Administrative Politics in British Government* (1985). The prophet of the human relations movement was Mary Parker Follett, some of whose writings are published in *Dynamic Administration: The Collected Papers of Mary Parker Follett*, new ed., edited by ELLIOT M. FOX and L. URWICK (1973, reissued 1982). The derivative movement, now called organization development, is treated in CHRIS ARGYRIS, *Integrating the Individual and the Organization* (1964); RENSIS LIKERT, *New Patterns of Management* (1961, reprinted 1987); and WARREN G. BENNIS, *Organization Development: Its Nature, Origins, and Prospects* (1969). A wide-ranging discussion of issues in policy analysis is provided in BRIAN W. HOGWOOD and LEWIS A. GUNN, *Policy Analysis for the Real World* (1984). A growing literature has developed since World War II in case studies of actual administrative experience. Pioneered and led by the American Inter-University Case Program, the use of cases has spread to many other countries. An example of the use of cases in comparative analysis is FREDERICK C. MOSHER (ed.), *Governmental Reorganizations: Cases and Commentary* (1967). JOHN E. ROUSE, JR., *Public Administration in American Society: A Guide to Information Sources* (1980), is a comprehensive annotated bibliography.

On the administrative systems of different countries, see BRIAN CHAPMAN, *The Profession of Government: The Public Service in Europe* (1959, reprinted 1980); F.F. RIDLEY (ed.), *Specialists and Generalists: A Comparative Study of the Professional Civil Servant at Home and Abroad* (1968); FRED W. RIGGS, *Administration in Developing Countries: The Theory of Prismatic Society* (1964); MORROE BERGER, *Bureaucracy and Society in Modern Egypt: A Study of the Higher Civil Service* (1957, reissued 1969); and JOSEPH LA PALOMBARA (ed.), *Bureaucracy and Political Development*, 2nd ed. (1967).

There are few general comparative studies on civil services. The important comparative works in the prewar period are HERMAN FINER, *Theory and Practice of Modern Government*, rev. ed. (1949, reprinted with a new introduction, 1970); ERNEST BARKER, *The Development of Public Services in Western Europe, 1660–1930* (1944, reissued 1966); LEONARD D. WHITE (ed.), *The Civil Service in the Modern State* (1930), and *The Civil Service Abroad: Great Britain, Canada, France, Germany* (1935). The postwar comparative studies include POUL MEYER, *The Administrative Organization: A Comparative Study of the Organization of Public Administration* (1957); EDWARD C. PAGE, *Political Authority and Bureaucratic Power: A Comparative Analysis* (1985); and B. GUY PETERS, *The Politics of Bureaucracy*, 2nd ed. (1984). For a wider discussion of public employment, including local and regional government officials and employees in public enterprises and health services, as well as a survey of the growth in public employment since the middle of the 19th century, see RICHARD ROSE *et al.*, *Public Employment in Western Nations* (1985). For a comparative public policy analysis, see ARNOLD J. HEIDENHEIMER, HUGH HECLO, and CAROLYN TEICH ADAMS, *Comparative Public Policy: The Politics of Social Choice in Europe and America*, 2nd ed. (1983). Two classic works that established the modern theory of the relations between civil servants and the state are RUDOLF GNEIST, *Der Rechtsstaat* (1872); and LÉON DUGUIT, *Law in the Modern State* (1919, reprinted 1970; originally published in French, 1913).

European countries with long traditions of civil service government have very considerable bibliographies on various aspects of public administration. Useful surveys of the administrative systems of the larger European nations are found in F.F. RIDLEY (ed.), *Government and Administration in Western Europe* (1979). The special aspect of control of the civil service is dealt with in CHARLES E. FREEDEMAN, *The Conseil d'État in Modern France* (1961; reprinted 1968). The political background to the struggle for civil servants' rights may be traced through PIERRE D'HUGUES, *La Guerre des fonctionnaires* (1912).

The overall pattern of the federal civil service and administration in the United States is dealt with in EDWARD S. CORWIN, *The President: Office and Powers, 1787–1957: History and Analysis of Practice and Opinion*, 5th rev. ed. (1984); and MARVER H. BERNSTEIN, *The Job of the Federal Executive* (1958, reprinted 1986). The special problem of the U.S. diplomatic service is the subject of a special enquiry presented in the COMMITTEE ON FOREIGN AFFAIRS PERSONNEL, *Personnel for the New Diplomacy: Report* (1962). The relations between the civil service and Congress are studied in considerable depth in JOSEPH P. HARRIS, *Congressional Control of Administration* (1964, reprinted 1980). LAURENCE E. LYNN, JR., *Managing the Public's Business: The Job of the Government Executive* (1981); HUGH HECLO, *A Government of Strangers: Executive Politics in Washington* (1977); and HERBERT KAUFMAN, *The Administrative Behavior of Federal Bureau Chiefs* (1981), focus on the difficulties experienced by appointed executives in managing their agencies.

General studies on the organization of executive power in former Communist countries include H. GORDON SKILLING, *The Governments of Communist East Europe* (1966); and GHITA IONESCU, *The Politics of the European Communist States* (1967). A detailed study of the organization of authority in the former Soviet Union may be found in MICHEL TATU, *Power in the Kremlin* (1969, originally published in French, 1967); this is contrasted with Chinese theory and practice in DONALD W. TREADGOLD (ed.), *Soviet and Chinese Communism: Similarities and Differences* (1967). The internal organization and structure of government in China itself are discussed in FRANZ SCHURMANN, *Ideology and Organization in Communist China*, 2nd ed. (1968); and A. DOAK BARNETT, *Cadres, Bureaucracy, and Political Power in Communist China* (1967). See also HARRY HARDING, *Organizing China: The Problem of Bureaucracy, 1949–1976* (1981).

The special problems of civil services in new states are outlined in KENNETH YOUNGER, *The Public Service in New States: A Study in Some Trained Manpower Problems* (1960, reprinted 1974); and in A.L. ADU, *The Civil Service in New African States* (1965), which can be compared with the same author's *Civil Service in Commonwealth Africa: Development and Transition* (1969). Interesting comparative, theoretical, and institutional studies on a broad front are found in RALPH BRAIBANTI (ed.), *Asian Bureaucratic Systems Emergent from the British Imperial Tradition* (1966); JOHN D. MONTGOMERY and WILLIAM J. SIFFIN (eds.), *Approaches to Development: Politics, Administration and Change* (1966); FERREL HEADY, *Public Administration: Comparative Perspective*, 3rd rev. ed. (1984); and JOSEPH LA PALOMBARA (ed.), *Bureaucracy and Political Development*, 2nd ed. (1963). Useful descriptions of the civil services of the European Communities and the United Nations can be found in CHARLES DEBBASCH (ed.), *La Politique de choix des fonctionnaires dans les pays européens* (1981).

There are no works covering the whole subject of administrative law in its differing forms in many countries. For the present system in the United Kingdom, see H.W.R. WADE, *Administrative Law*, 5th ed. (1982); DAVID FOULKES, *Administrative Law*, 6th ed. (1986); S.A. DE SMITH, *De Smith's Judicial Review of Administrative Action*, 4th ed., edited by J.M. EVANS (1980); and J.F. GARNER and B.L. JONES, *Garner's Administrative Law*, 6th ed. (1985). WILLIAM A. ROBSON, *Justice and Administrative Law: A Study of the British Constitution*, 3rd ed. (1951, reprinted 1970), is a standard work on the rise and purpose of administrative tribunals. PHILIP NORTON, *The Constitution in Flux* (1982), contains a useful introductory overview of the system of grievance redress in the United Kingdom. The leading work on American administrative law is KENNETH CULP DAVIS, *Administrative Law Treatise*, 2nd ed., 5 vol. (1978–84). The complex question

of the control of administrative discretion in the United States is examined by KENNETH CULP DAVIS, *Discretionary Justice: A Preliminary Inquiry* (1969, reprinted 1980). A readable account of the American system may be found in BERNARD SCHWARTZ, *An Introduction to American Administrative Law*, 2nd ed. (1962). See also WALTER GELLHORN *et al.*, *Administrative Law: Cases and Comments*, 8th ed. (1987). The French system is appraised in C.J. HAMSON, *Executive Discretion and Judicial Control: An Aspect of the French Conseil d'État* (1954, reprinted 1979); a reliable description of its principles is contained in L. NEVILLE BROWN and J.F. GARNER, *French Administrative Law*, 3rd ed. (1983). A comparison between the Anglo-American system and the French is given in BERNARD SCHWARTZ, *French Administrative Law and the Common-Law World* (1954). Leading treatises by eminent French jurists are GEORGES VEDEL and PIERRE DELVOLVÉ, *Droit administratif*, 9th ed. (1984); MARCEL WALINE, *Droit administratif*, 9th ed. (1963); and M. LONG, P. WEIL, and G. BRAIBANT, *Les Grands Arrêts de la jurisprudence administrative*, 8th ed. (1984). For Australia, see HARRY WHITMORE, *Principles of Australian Administrative Law*, 5th ed. (1980). On Germany, ERNST FORSTHOFF, *Lehrbuch des Verwaltungsrechts*, 10th rev. ed. (1973); and HANS J. WOLFF and OTTO BACHOF, *Verwaltungsrecht*, 3 vol. in various editions (1978), are reliable works. The procuracy can be studied in GLEN G. MORGAN, *Soviet Administrative Legality: The Role of the Attorney General's Office* (1962); LEON BOIM and GLENN G. MORGAN, *The Soviet Procuracy Protests, 1937–1973: A Collection of Translations* (1978); GORDON B. SMITH, *The Soviet Procuracy and the Supervision of Administration* (1978); and more briefly in LEONARD SCHAPIRO, *The Government and Politics of the Soviet Union*, new rev. ed. (1978). Wide-ranging books on the ombudsman are a symposium edited by DONALD C. ROWAT, *The Ombudsman: Citizen's Defender*, 2nd ed. (1968), and his *Ombudsman Plan: The Worldwide Spread of an Idea*, 2nd rev. ed. (1985); WALTER GELLHORN, *Ombudsmen and Others: Citizens' Protectors in Nine Countries* (1966), and *When Americans Complain: Governmental Grievance Procedures* (1966); and GERALD E. CAIDEN (ed.), *International Handbook of the Ombudsman*, 2 vol. (1983). An excellent work on Scandinavia is NILS HERLITZ, *Elements of Nordic Public Law*, (1969; originally published in Swedish, 1959).

(F.C.M./B.Ch./W.A.R./E.C.P.)

Public Opinion

There are many difficulties in the way of defining public opinion, the most important of which are discussed below. A simple definition is that public opinion is an aggregate of the individual views, attitudes, and beliefs about a particular topic, expressed by a significant proportion of a community.

This article is divided into the following sections:

Historical background 310
  Ancient times
  Middle Ages
  Early modern times
  Current conceptions of public opinion
The formation and change of public opinion 312
  Formation of attitudes
  Immediate environmental factors
  The mass media
  Interest groups
  Opinion leaders
  Complex influences
Public opinion and government 313
Public-opinion polling 314
  Methodology of opinion polling
  Criticisms and justifications of opinion surveys
Bibliography 316

HISTORICAL BACKGROUND

Ancient times. Although the term public opinion was not used until the 18th century, phenomena that closely resembled public opinion seem to have occurred in many historical epochs. One of the oldest written records from ancient Egypt, a poem entitled "The Dispute with His Soul of One Who Is Tired of Life," refers to an upheaval that apparently involved a complete reorientation of mass opinion:

> To whom shall I speak today?
> People are greedy . . . .
> Gentleness of spirit has perished.
> All the people are impudent . . . .
> People laugh at crimes of him who before
> Would have enraged the righteous . . . .
> There are no just men.
> The earth has been given over to evil doers.

Similar references to popular attitudes can be found in the histories of Babylonia and Assyria. The prophets of ancient Israel sometimes justified the policies of the government to the people and sometimes appealed to the people to oppose the government. In both cases they were concerned with swaying opinion. And in classical Greece it was observed by many that everything depended on the people, and the people were dependent on the word. Wealth, fame, and respect all could be given or taken away by persuading the populace.

Wide dissemination of news, which is usually necessary for the formation of public opinion, could be observed in classical Rome. Much of this took place through person-to-person channels. When the Roman statesman Cicero was in Cilicia in the year 51 BC, he asked his friend Caelius to keep him informed of what was happening in the capital. Caelius promised to do so: "If anything important of a political nature should occur . . . I will diligently describe to you its origin, the general opinion about it, and the prospects of future action that it opens up." Rome also had its wall newspapers, composed by Roman officials and posted in public places to inform the public about acts of the government and principal local events.

Public opinion in Rome

Middle Ages. During the Middle Ages in western Europe, the masses were encased in a rural, traditional society in which most activities and attitudes were dictated by a person's station in life; but phenomena much like public opinion could be observed among the religious, intellectual, and political elite. Religious disputations, the struggle between popes and the Holy Roman Empire, and the dynastic ambitions of princes all involved efforts to persuade, to create a following, and to line up the opinions of those who counted. In 1191 the English bishop William of Ely was attacked by his political opponents for hiring troubadours to extol his merits in public places, so that "people spoke of him as though his equal did not exist on earth." The propaganda battle between emperors and popes was waged largely through sermons, but handwritten literature also played a part.

From the end of the 13th century, the ranks of those who could be drawn into controversy regarding current affairs grew steadily. There was an increasing spread of education among the lay population. The rise of humanism in Italy saw the emergence of a group of writers and publicists whose services were eagerly sought by the princes who were consolidating national states. Some of these writers were used as advisers and diplomats; others were employed as publicists because of their ability to sway opinion. Such a one was the Italian Pietro Aretino (1492–1556), of whom it was said that he knew how to defame, to threaten, and to flatter better than all others and whose services were sought by both Charles V of Spain and Francis I of France. The Italian political philosopher Niccolò Machiavelli, a contemporary of Aretino, wrote that princes should not ignore popular opinion, particularly in regard to such matters as the distribution of offices.

Influence of printing on the growth of public opinion

The invention of printing from movable type in the 15th century and the Protestant Reformation in the 16th increased still further the numbers of people able to form opinions on contemporary issues. Martin Luther broke with the humanists by abandoning the use of classical Latin, intelligible only to the educated, and turned directly to the masses. "I will gladly leave to others the honour of doing great things," he wrote, "and will not be ashamed of preaching and writing in German for the unschooled layman." Luther's Ninety-five Theses, which were printed against his will and widely spread throughout Europe, were of a theological nature, but he also wrote on such subjects as the war against the Turks, the Peasants' Revolt, and the evils of usury. His vigorous expressions and the counterblasts from his many opponents, both lay and clerical, led to the formation of larger and larger groups holding opinions on important matters of the day.

Extensive attempts to create and influence public opinion were made during the Thirty Years' War (1618–48). A flood of propaganda tracts, many illustrated with woodcuts, emanated from both sides. Opinions were also swayed by means of speeches, sermons, and face-to-face discussions. Not surprisingly, both civil and religious authorities attempted to control the dissemination of unwelcome ideas by increasingly strict censorship. Pope Paul IV had the first Index of Prohibited Books drawn up in 1559. Charles IX of France decreed in 1563 that nothing could be printed without the special permission of the king.

More quietly, but more significantly, newspapers and news services were developing. Rudimentary private news services had been maintained by political authorities and wealthy merchants ever since classical times, but they were not available to the public. By 1500, however, it was possible to buy specialized news sheets in many of the principal cities of Europe. One of these, printed in 1514 or 1515, contains an extract of a merchant's letter telling of the Portuguese discovery of Brazil. The first regularly printed newspapers appeared in about 1600 and multiplied rapidly thereafter, although they were frequently bedevilled by censorship regulations. Regular postal services, started in France in 1464 and in the Austrian Empire in 1490, facilitated the spread of information enormously.

Bourses as sources of informed opinion

Early modern times. The great news centres of early modern times were the financial exchanges. With the introduction of a paid civil service and the employment of paid soldiers in the place of vassals, princes found it necessary to borrow money. The bankers, in turn, had to know a great deal about the credit of the princes, the state of their political fortunes, and their reputations with their subjects. All kinds of political and economic information flowed to the money-lending centres at Antwerp, Lyon, and Nürnberg, and this information gave rise to generally held opinions in the banking community. The *ditta di borsa*—the opinion on the bourse—is often referred to in documents of the period. Queen Elizabeth of England was regarded as especially well informed because Sir Thomas Gresham, the finance agent of the English crown, kept in constant touch with the Antwerp bourse.

Significantly, it was another financial official who first popularized the term public opinion in modern times. Jacques Necker, who was the finance minister of Louis XVI on the eve of the French Revolution, noted repeatedly in his writings that public credit depended upon the opinions of holders and buyers of government securities about the viability of the royal administration. He, too, was vitally concerned with the *ditta di borsa.* But he also remarked on the power of public opinion in other areas. "This public opinion," Necker wrote, "strengthens or weakens all human institutions." As he saw it, public opinion should be taken into account in political undertakings. Necker was not, however, concerned with the opinions of all Frenchmen. For him, the people who collectively shaped public opinion were those who could read and write, who lived in cities, who kept up with the day's news, and who had money to buy government securities—or in short, the bourgeoisie.

A public opinion that extended beyond the middle classes and embraced the urban masses took shape during the French Revolution. Observers of the Revolution were mystified, and often terrified, by this new phenomenon of public opinion, which seemed able to sweep aside the entrenched institutions of the time: the monarchy, the church, and the feudal system. Thinkers of the late 18th and early 19th centuries advanced a variety of definitions as to what public opinion actually was. One of the most detailed descriptions was given in 1799 by the German poet Christoph Wieland, who closely followed the stormy events in France and the rest of western Europe:

> I, for my part, understand by it an opinion that gradually takes root among a whole people, especially among those who have the most influence when they work together as a group. In this way it wins the upper hand to such an extent that one meets it everywhere. It is an opinion that without being noticed takes possession of most heads, and even in situations where it does not dare to express itself out loud can be recognized by a louder and louder muffled murmur. It then requires only some small opening that will allow it air, and it will break out with force. Then it can change all nations in a brief time and give whole parts of the world a new configuration.

A German philosopher of the time, Christian Garve, gave much more emphasis to the rational component:

> Public opinion as interpreted . . . by those French writers who are clearest on the subject is the agreement of many or of the majority of the citizens of a state with respect to judgments which every single individual has arrived at as a result of his own reflection or of his practical knowledge of a given matter.

The English philosopher Jeremy Bentham, who advanced the first detailed discussion of public opinion in English, was troubled by the difficulty of defining it and advised that the term be employed only in deference to common usage.

Current conceptions of public opinion. In spite of voluminous discussions of the subject, scholars still do not agree on a definition of public opinion. Members of a roundtable of the American Political Science Association that met in 1925 divided into three groups: those who did not believe that there was such a thing as public opinion; those who accepted its existence but doubted their ability to define it precisely; and those who could offer a definition. This last group could not, however, agree on the definition to be adopted. Although few scholars now question the existence of such a phenomenon as public opinion, differences in defining it have persisted to the present day.

These differences stem in part from the varying perspectives with which scholars have approached the study of public opinion and in part from the fact that the phenomenon is still not completely understood. Political scientists and some historians have tended to emphasize public opinion's role in the governmental process, paying particular attention to its influence on government policy. Some political scientists have regarded public opinion as equivalent to the national will. In this sense, there can be only one public opinion on an issue at any one time.

Sociologists usually give more emphasis to public opinion as a product of social interaction and communication. According to the sociological view, there can be no public opinion without communication among members of the public who are interested in a given issue. A large number of persons may hold quite similar views, but these will not coalesce into public opinion as long as each person remains ignorant of the opinions of the others. Communication may take place by means of the mass media of the press, radio, and television or through face-to-face discussions. Either way, people learn how others think about a given issue and may take the opinions of others into account in making up their own minds.

Sociological conceptions of public opinion

Sociologists suggest that there may be many different public opinions existing on a given issue at the same time. One body of opinion may be dominant or may be reflected in governmental policy, but this does not mean that other organized bodies of opinion do not exist. The sociological approach also sees the public-opinion phenomenon as extending to areas that are of little or no concern to government. Thus, fads and fashions are appropriate subject matter for students of public opinion, as are public attitudes toward movie stars or corporations.

It is often the case that opinions expressed in public may differ from those expressed in private and that only the former contribute to public opinion. Similarly, some attitudes—even though widely shared—may not be expressed at all. Thus, in a totalitarian state, a great many people may be opposed to the government but may fear to express their attitudes even to their families and friends. In such cases, an antigovernment public opinion fails to develop.

Private opinions, if expressed in public, may become a basis for public opinions. Until the 1930s, for example, there was an unwritten prohibition in the United States against discussions of venereal disease, although many individuals had private opinions about it. Then, when the subject began to be treated in the mass media and public opinion researchers began to ask questions about it, opinions that had formerly been private were expressed in public, and sentiment in favour of government action to stamp out venereal disease developed.

Some public-opinion survey specialists have preferred a definition that links public opinion directly to their polling procedures. Public opinion is therefore defined as being identical to what people's responses to a survey questionnaire would be. Other similar definitions have been to the effect that public opinion is whatever is discovered by public-opinion polls. This definition, while widely used in practice, has the disadvantage of implying that public opinion does not exist in places and times in which there are no opinion polls. A more generally applicable approach that embodies much the same reasoning is that public opinion on any matter may be conceived as the hypothetical result of some imaginary survey or vote.

Those who are primarily engaged in the manipulation of public opinion, notably professional politicians and public relations men, rarely stop to define it. The American journalist and political scientist Walter Lippmann has observed that there has been a tendency in democracies to make a mystery out of public opinion but that "there have been skilled organizers of opinion who understood the mystery well enough to create majorities on election day." (*Public Opinion,* 1922.) Public relations practitioners have concerned themselves less with public opinion in general than with the opinions of specified "publics" that may affect the fortunes of a client: employees, stockholders,

government officials, suppliers, and potential buyers, for example. Both politicians and public relations men are interested in influencing behaviour and thus in determining any attitudes and opinions that may affect that behaviour, whatever they may be called.

Nearly all scholars and manipulators of public opinion, regardless of the way they may define it, agree that at least four factors are involved in public opinion: there must be an issue; there must be a significant number of individuals who express opinions on the issue; there must be some kind of a consensus among at least some of these opinions; and this consensus must directly or indirectly exert influence.

THE FORMATION AND CHANGE OF PUBLIC OPINION

The democratic system itself defines a number of issues on which citizens are under pressure to form opinions. They are called upon to decide among various candidates in elections and, on occasion, to vote on constitutional amendments and various other propositions. Almost any matter on which the executive or legislature has to decide may become a public issue if a significant number of persons wish to make it one. The attitudes of these persons are often stimulated or reinforced by outside agencies—a crusading newspaper, a pressure group, or a governmental agency or official. Even matters that are not within the purview of any governmental agency may become public issues. No agency, for example, has the authority to determine how many children a family may have or how long a man's hair should be, but discussion on these subjects has at times been sufficiently intense to generate widespread opinion about them.

Elements involved in the formation of attitudes

Formation of attitudes. Once a public issue is identified, a certain number of people will begin to form attitudes about it. If the attitude is expressed to others by sufficient numbers of people, a public opinion on the topic begins to emerge. Not all people develop attitudes on public issues; some may not be interested, and others simply may not hear about them. The attitudes that are formed may be held for various reasons. Thus, four men may all be opposed to higher property taxes but for very different reasons. One man may not be against higher taxes in principle, but he opposes them because he is having trouble paying the mortgage on his house. This attitude serves an adjustment, or utilitarian, function in that it helps its holder to accommodate the immediate financial situation in which he finds himself. A second man may fight the tax because he does not want a certain social group, such as the poor or the unemployed, to derive any benefit from tax revenues. Such an attitude may be the result of a psychological insecurity, of a desire to keep the poor "in their place" in order to bolster his own sense of superiority toward underprivileged groups. For such a person, the attitude serves an ego-defensive function. A third man may resist the tax increase because he believes that governmental activities should be severely restricted. His attitude has a value-expressive function in that it reflects his overall philosophy. A fourth man opposes the increased tax because he is familiar with instances of governmental waste and is convinced that all necessary services could be rendered if officials spent the already available funds more rationally. His attitude is thus determined by knowledge or experience in that it is a reflection of what he has learned in the past. A fifth man, of course, might fight the tax for all four reasons. A seemingly homogeneous body of public opinion may thus be composed of individual opinions that are rooted in very different interests and values. If an attitude does not serve a function such as one of the above, it is unlikely to be formed: an attitude must be useful in some way to the person who holds it.

How many people will actually form opinions on a given issue, as well as what sort of opinions they form, depends partly on their own preexisting knowledge, attitudes, and values; partly on the personal situations in which they find themselves; and partly on a number of environmental factors. As far as preexisting knowledge and attitudes are concerned, it is often surprising to discover how many people are not informed about major issues and therefore have no attitudes toward them. In 1964, for instance, one in four Americans did not know that the government in China was Communist, and about the same proportion at any given time is ignorant of any major issue of public policy. Substantially the same situation has been found to exist in western European countries, and ignorance is probably even more widespread in nations with lower levels of education. Values are of considerable importance in determining whether people will form opinions on a particular topic. If people feel that their moral principles or personal philosophies are involved in an issue, they are more likely to take a favourable or opposing stand.

Immediate environmental factors. Environmental factors play an extremely important part in the formation of attitudes and opinions. Most pervasive is the influence of the immediate social environment: family, friends, neighbourhood, place of work, church, or school. People usually adjust their attitudes to conform with those that are most prevalent in the social groups to which they belong. If a person who considers himself a liberal is surrounded in his home or at his place of work by people who profess conservatism, he is more likely to switch his vote than is a liberal whose family and friends share his political views. Similarly, it was found in World War II that men transferred from one unit to another often adjusted their opinions to conform more closely with those in the unit to which they were transferred.

The influence of mass media

The mass media. The press, radio, and television are usually less important than the immediate social environment when it comes to the formation of attitudes, but they are still significant. They focus the attention on certain personalities and issues, and many people subsequently form opinions about these issues. Government officials have noted that their mail from the public tends to "follow the headlines"; whatever is featured in the press at a particular moment is likely to be the subject that most people write about. The mass media can also activate and reinforce latent attitudes. Political attitudes, for example, are likely to be activated and reinforced just before an election. Voters who may have only a mild preference for one party or candidate before the election campaign starts are often worked up by the mass media to a point where they not only take the trouble to vote but may contribute money or help a party organization in some other way.

The mass media play another extremely important role in letting individuals know what other people think and in giving leaders large audiences. In this way they make it possible for public opinion to include a large number of individuals and to spread over wider geographic areas. It appears in fact that in some European countries the growth of broadcasting, and especially television, has affected the operation of the parliamentary system. Before television, national elections were seen largely as contests between a number of candidates or parties for parliamentary seats. More recently, elections in such countries as Germany and Great Britain have appeared more as a personal struggle between the leaders of the principal parties concerned, since these leaders were featured on television and came to personify their parties. Television in France and the United States has been regarded as a powerful force strengthening the presidential system, since the president can easily appeal to a national audience over the heads of elected legislative representatives.

Even when the mass media are thinly spread, as in developing countries or in nations where the media are strictly controlled, word of mouth can sometimes perform the same functions as the press and broadcasting, although on a more limited scale. In developing countries, it is common for those who are literate to read from newspapers to those who are not, or for large numbers of persons to gather around the one village radio. Word of mouth in the marketplace or neighbourhood then carries the information farther. In countries where important news is suppressed by the government, a great deal of information is transmitted by rumour. Word of mouth thus helps public opinion to form in developing countries and encourages "underground" opinion in totalitarian countries, even though these processes are slower and usually involve fewer people than in countries where the media network is dense and uncontrolled.

The role of pressure groups

Interest groups. Pressure groups, or interest groups, also play an important part in the formation and spread of public opinion on issues of relevance to themselves. These groups may be concerned with political, economic, or ideological issues and often work through the mass media as well as by word of mouth in trying to influence attitudes. Some of the larger or more affluent interest groups in the United States, western Europe, and elsewhere make use of advertising and public relations to influence opinion. During the late 1960s and early 1970s, for example, hundreds of thousands of dollars were spent in the United States on advertising by opponents and proponents of U.S. policy in Southeast Asia, and lesser amounts were devoted to interest group advertising on other public issues. In Britain, much advertising space was purchased by opponents of Britain's proposed entry into the Common Market.

Opinion leaders. Opinion leaders play a major role in defining important issues and in influencing individual opinions regarding them. Political leaders, in particular, can turn a hitherto relatively unknown problem into a national issue if they decide to call attention to it. One of the ways in which opinion leaders rally opinion and smooth out the differences among those who are in basic agreement on a subject is by coining or popularizing symbols or slogans: Sir Winston Churchill popularized the phrase Cold War, and the Allies in World War I were fighting "a war to end all wars." Slogans are perhaps among the most useful tools that are available to the political leader. Once enunciated, symbols and slogans are frequently kept alive and communicated to large audiences by the mass media and may become the cornerstone of public opinion on any given issue.

Opinion leaders are not confined only to prominent figures in public life. There are likely to be persons in every social group to whom others in the immediate environment look for guidance on certain subjects. Thus, one person may be thought of by those in his own social group as especially qualified in the realm of local politics, another as a reliable guide in foreign affairs, and a third as an expert when it comes to buying a house. These local opinion leaders are generally unknown outside their own circle of friends and acquaintances, but their cumulative influence in the formation of public opinion is substantial.

Complex influences. Although a person's psychological makeup, his personal circumstances, and external factors such as pressure groups and opinion leaders all play a role in the formation of opinions, it is still not known exactly how public opinion on an issue takes shape. Many aspects of the public opinion process have yet to be explored. The same is true with regard to changes in public opinion. Some of these can be accounted for by changing events and circumstances, but others are more difficult to explain. It is known that public opinion on some subjects tends to follow events. Public attitudes toward other nations, for example, seem to depend largely on the relations between the governments of the two nations. Hostile attitudes do not cause poor relations; they are the result of them. People presumably change their attitudes when these attitudes do not correspond with their perception of prevailing circumstances and hence are not useful as guides to action. It is also frequently the case that an issue ceases to be important and simply fades from public attention, while new issues arise as the basis for new bodies of opinion. There are still, nevertheless, major changes in public opinion that are difficult to explain. During the second half of the 20th century in many parts of the world, attitudes toward religion, family, sex, international relations, social welfare, and the economy have undergone major shifts. There have been important issues claiming public attention in all these areas, but the changes in public opinion are difficult to relate to any major event or even to any complex of events.

PUBLIC OPINION AND GOVERNMENT

Many early thinkers saw public opinion as a powerful force that rulers must learn how to control. The 18th-century French philosopher Jean-Jacques Rousseau believed that all laws were based upon it but that this did not necessarily diminish the powers of government. It was Rousseau's opinion that "Whoever makes it his business to give laws to a people must know how to sway opinions and through them govern the passions of men." The 19th-century German philosopher G.W.F. Hegel described public opinion as containing both truth and falsehood together and added that it was the task of the great man to distinguish the one from the other. Jeremy Bentham saw the greatest difficulty of the legislator as being "in conciliating the public opinion, in correcting it when erroneous, and in giving it that bent which shall be most favourable to produce obedience to his mandates."

At the same time, Bentham and some other thinkers believed that public opinion was a useful check on the authority of rulers. Bentham demanded that all official acts be given publicity, so that an enlightened public opinion could pass on them, as would a tribunal: "To the pernicious exercise of the power of government it is the only check." The British jurist and historian James Bryce, writing a century later, maintained that if government was based on popular consent, this would give a nation great stability and strength: "It has no need to fear discussion and agitation. It can bend all its resources to the accomplishment of its collective ends." Bryce did not, however, believe that mass opinion could or should dominate details of governmental policy, since most people did not have the leisure or inclination to arrive at a position on every question. Rather, the masses would set the general tone for policy, their sentiments leading them to take a stand on the side of justice, honour, and peace.

Public opinion as a safeguard against abuse of authority

Those who worked to advance international understanding were particularly likely to invoke the power of public opinion. Both Bentham's "Plan for an Universal and Perpetual Peace" (1789) and Immanuel Kant's proposals in his essay "On Perpetual Peace" (1795) were based on the belief that public opinion is peace loving and that international peace can be sustained by it. Those who advocated establishment of the League of Nations after World War I also looked to world public opinion as the principal force that would sustain the League. Drafters of the charter of the United Nations Educational, Scientific and Cultural Organization (UNESCO) apparently had somewhat the same idea, noting that because the origins of war were to be found in the minds of the masses of men it was in the minds of men that the defenses of peace should be constructed.

Some scholars, while acknowledging the power of public opinion, warned that it could be a dangerous force. The 19th-century French writer Alexis de Tocqueville was concerned about the possible "tyranny of the majority" if government was in fact to be an expression of mass attitudes. Many other writers have expressed concern, often in a more extreme form, about the dangers of allowing government policy to be influenced too much by public opinion, which may well be uninformed, unthinking, and unstable. But whether public opinion is regarded as a constructive or a baneful force in a democracy, there are few politicians who are prepared to deny in public that government should follow public opinion.

An assessment of the significance of public opinion

In recent years, political scientists have been less concerned with what part public opinion should play in a democratic polity and have given more attention to establishing what part it does play in actuality. From the examination of numerous histories of policy formation, it is clear that no sweeping generalization can be made that will hold in all cases. The role of public opinion appears to vary from issue to issue, and the way it asserts itself differs from one democracy to another. The safest generalization that can be made is that public opinion does not influence the details of most policies but that it does set limits within which the policymaker must operate. That is, public officials will usually seek to satisfy a widespread demand, or will at least take it into account in their deliberations, and they will also try to avoid decisions that they believe will fly in the face of popular opinion. In addition, it has been observed that the relation between public opinion and public policy is two-way: policy influences opinion, as well as the reverse, and there is usually at least an initial tendency for the public to accept a decision once it is made. Public opinion seems to be particularly effective in influencing policymaking at the local level,

as officials appear to feel themselves constrained to yield to popular pressures for better roads, better schools, or more hospitals.

Public opinion at the national level seems to play a more limited role—partly because of the inability of most people to understand the complexities of most issues faced by government and partly because of the growth of executive power and the development of large governmental bureaucracies that serve as screens between the policymaker and the public. Representative government itself also tends to limit the power of public opinion to influence specific decisions, since ordinarily the public is given the choice only of approving or disapproving the election of a given official.

PUBLIC-OPINION POLLING

Public-opinion polling can provide a fairly exact analysis of the distribution of opinions on almost any issue within a given population. Assuming that the proper questions are asked, polling can also reveal something about the intensity with which opinions are held, about the reasons for these opinions, and about whether or not the issues have been discussed with others. Polling does not usually reveal whether or not the people holding an opinion can be thought of as constituting a cohesive group, and it is unlikely to provide very much information about the elites who may have played an important part in developing the opinion. But in spite of these deficiencies, polling is a valuable tool for estimating the state of public opinion on almost any subject.

Origins of opinion research

Opinion research developed from market research. Early market researchers picked small samples of the population and used these to obtain information on such questions as how many people read a given magazine or listen to the radio and what the public likes and dislikes in regard to various consumer goods. About 1930, both commercial researchers and scholars began to experiment with the use of these market research techniques to obtain information on opinions about political issues. In 1935 the U.S. statistician George Gallup began conducting nationwide surveys of opinions on current political and social issues in the United States. One of the first questions asked by the Gallup Poll (its full name is the American Institute of Public Opinion) was "are Federal expenditures for relief and recovery too great, too little, or about right?" To this, 60 percent of the sample replied that they were too great, only 9 percent thought they were too little, and 31 percent regarded them as about right.

From the 1930s on, the spread of opinion polls conducted by both commercial and academic practitioners continued at an accelerated pace in the United States and, to a more limited extent, elsewhere. In 1937, the *Public Opinion Quarterly,* which later became the official organ of the American Association for Public Opinion Research, began publication. State and local polls, some sponsored by newspapers, were started in many parts of the country, and opinion research centres were organized at several universities. Before and during World War II, opinion polls were extensively used by U.S. government agencies, notably the Department of Agriculture, the Treasury Department, and the War Department.

At the same time, opinion research was increasingly used in other parts of the world. Several affiliates of the American Institute of Public Opinion were organized in Europe and Australia in the late 1930s, and following World War II polling organizations appeared in numerous countries of Europe, Asia, and Latin America.

Polls have been successful in forecasting election results in nearly every case in which they have been used for this purpose. The two most notable failures were in the United States in 1948, when nearly all polls forecast a Republican victory and the Democrats won by a narrow margin, and in Great Britain in 1970, when all but one of the major polls predicted a Labour Party victory and the Conservative Party achieved a majority. In both cases the major polls did not show large deviations from the actual results but nevertheless picked the wrong winner. Professional opinion researchers point out that predicting elections will always be chancy because of the possibility of last-minute shifts and turnout problems; nevertheless, their record has been good over the years.

Although popular attention has been focussed on polls taken before major elections, most polling is devoted to other subjects, and university-based opinion researchers usually do not make election forecasts at all. Support for opinion studies comes largely from public agencies, foundations, and commercial firms, which are interested in such questions as how well people feel that their health, educational, and other needs are satisfied; how such problems as racial prejudice and drug addiction can be attacked; or how well a given industry is meeting public demands. Polls that are regularly published in newspapers or magazines usually have to do with some lively social issue—and elections are included only as one of many subjects of interest.

Methodology of opinion polling. The principal steps in opinion polling are the following: defining the "universe," choosing a sample, framing a questionnaire, interviewing persons in the sample, tabulating the results, and analyzing or interpreting the results.

Universe is the term used to denote whatever body of people is being studied. This is not always easy to define precisely. If, for example, one is making a study of college-student opinion, it is necessary to decide whether the universe should include full-time students only or whether it should include those who are not candidates for established degrees. The way in which these questions are answered will have an important bearing on the outcome of the survey and possibly on its usefulness.

Once the universe has been carefully defined a sample of the universe has to be picked. If possible, a "probability sample" should be chosen. Ideally, the best way to do this would be to assign a number to each person in the universe—or write his name on a slip of paper—place all the numbered or named slips in a container, mix thoroughly, and then pick a sample without looking at the names or numbers. In this way, each slip will have the same probability of being chosen. If each person is numbered, the same effect can be achieved by using tables of random numbers, which can often be purchased in book form. The random numbers are matched up with the numbered members of the universe, until a sample of the desired size is drawn. The numbering procedure is often not practicable, but a few universes are already assigned numbers—all the workers in a given factory, for instance, or all members of the armed forces. It is not surprising that some of the most reliable opinion surveys have been conducted by the military.

Another method, not quite so reliable statistically, is to include every "*n*th" member of the universe in the sample. Thus, if one wishes to study the attitudes of the subscribers to a certain magazine and the magazine has 10,000 subscribers, one could take every 10th name from the subscription list and end up with a sample of 1,000.

Neither of these methods is likely to be useful when the universe consists of a large population that has not been numbered and when the names of members of the universe are not listed in a card file somewhere. This is the situation that was faced by market and opinion researchers when they first started conducting large-scale surveys. They therefore adopted the simpler device of picking a "quota sample." In quota sampling, an effort is made to match up the characteristics of the sample with those of the universe, so that a small replica of the universe is achieved. If one knows, possibly on the basis of the most recent census, that there are 51 women to every 49 men in the universe, then the sample should reflect these proportions. The same principle should be applied with respect to age, income, education, occupation, religious preference, national origin, area of residence, and indeed every characteristic that might be relevant to the range of opinions being studied. Each interviewer is told the characteristics of the people that he must locate and interview.

Quota sampling

Most survey organizations prior to 1948 used quota samples, and some still do. The British organization that correctly forecast the outcome of the 1970 parliamentary election used a quota sample, but it is, nevertheless, a risky technique. In many countries census data are poor

or nonexistent. Even when there are reliable census data, some characteristics that may affect the opinions being studied cannot be taken into account. It is not known, for example, how many vegetarians there are in most populations or how many extroverts and introverts. Yet these characteristics may be related to opinions on certain subjects. Statisticians point out that in a quota sample it is impossible to give each member of the universe a known chance of being selected, and one cannot therefore calculate the range of error in the results that could be due to chance. In this type of sample, furthermore, interviewers have to use their judgment in selecting respondents, and their standards may vary. The great advantage of a quota sample is that it is rather inexpensive to design and interview. By contrast, selecting and interviewing a probability sample from a large population can be very expensive.

How large a sample should be depends on the precision that is desired. For many purposes, a sample of a few hundred is adequate—*if it is properly chosen.* A magazine, for instance, might poll a random sample of 200 of its subscribers and find that 18 percent wanted more fiction and 62 percent wanted more articles on current social issues. Even if each of these figures were wrong by as much as 10 percentage points, the poll would probably still be of value since it would give fairly accurate information about the way the subscribers ranked the types of content. An election poll, on the other hand, would have to be much more accurate than this, since leading candidates often split the vote rather evenly. For most purposes, a national sample of about 1,500 cases is adequate *unless* it is desired to make comparisons among rather small subgroups in the population or to compare one small group with a much larger one. In such cases a larger sample is required in order that a significant number of members of the minority group may be included.

Allowances for chance and errors

There are no hard-and-fast rules for interpreting poll results, since there are many possible sources for bias or error. Nevertheless, for a well-conducted poll the following "rule of thumb" allowances for error are helpful. When any group of people is compared with another and the sample size of the *smaller* group is about 100, the difference between the percentages based on the two groups should be greater than 14 if this difference is to be regarded as significant. If the smaller group is larger than 100, the allowance for error caused by chance decreases approximately as follows: for a group comprising 200 cases, allow 10 percentage points; for 400 cases, allow 7 percentage points; for 800, allow 5; for 1,000, allow 4; for 2,000, allow 3. Thus, if a national sample survey shows that 27 percent of college students favour a volunteer army, while 35 percent of adults who are not in college do, and there are only 200 students in the sample, the difference between the two groups may well be due to chance. If the difference were greater than 10 percentage points, then it is much more likely that the opinions of students actually differ from those who are not in college. Similar allowances have to be made when election polls are interpreted. The larger the sample and the larger the difference between the number of preferences expressed for each candidate, the greater the certainty with which the election result can be predicted.

Even larger variations than those due to chance may be caused by the way the questions are worded. If one poll asks "Are you in favour of increasing government aid to higher education?" while another poll asks "Are you in favour of the president's (or premier's) recommendation that government aid to higher education be increased?" the second question is likely to receive many more affirmative answers than the first. Similarly, the distribution of replies will often vary if an alternative is stated. The question might be phrased: "Are you in favour of increasing government aid to higher education, or do you think enough tax money is being spent on higher education now?" In this form, it is probable that the question would receive fewer affirmative responses than if only the first half were used. As a rule, relatively slight changes in question wording do not cause great variations in response when people hold very strong opinions, but if their opinions are not firm then slight differences may sway them one way or another. Opinion researchers therefore frequently ask exactly the same question over a period of years. In this way, the results from an earlier survey can more safely be compared with the results from a later one.

Questionnaire construction, as with sampling, requires a high degree of skill. The question must be clear to people of varying educational levels and backgrounds. Questions must not embarrass respondents. They must be arranged in a logical order and so on. Even experienced researchers find it necessary to pretest their questionnaires. They send out interviewers to interview a small sample with the preliminary questions, and they may revise these questions to ensure that they are unambiguous and are actually obtaining the information sought.

Poll questions may be of the "forced-choice" or "free-answer" type. In the former, a respondent is asked to reply "yes" or "no" or else may be given a list of prepared alternatives. Even so, many respondents are likely to reply "don't know" or to prefer an alternative that the researcher had not listed in advance. A free-answer question allows the respondent to state his opinion in his own words. For instance, "What do you think are the most important problems facing the country today?"

Importance of careful interviewing

Interviewing is also a skilled operation. An inexperienced interviewer may bias a respondent's answer by the way he asks a question. He may antagonize some respondents so that they refuse to go on with the interview. He may not record the replies to free-answer questions accurately, or he may not be sufficiently persistent in locating designated respondents. Most large polling organizations give interviewers special training before they are sent out on surveys or else contract with an interviewing service that has trained and experienced interviewers available. A good sample and a well-tested questionnaire are not sufficient to guarantee an accurate survey if interviewing is slipshod.

Tabulation is usually done by machine. To simplify this process, most questionnaires are "precoded," which is to say that numbers appear beside each question and each possible response. The answers given by respondents can thus be translated rapidly into a numerical form that can be used in a computer. In the case of free-answer questions, responses must usually be grouped into categories, each of which is also assigned a number. How the categories are defined may make a large difference in the way the results are presented. If a respondent mentions narcotics addiction as a major problem facing the country, for instance, this answer might be coded as a health problem or a crime problem or might be grouped with other replies dealing with drug abuse or alcoholism.

The final steps in a survey are the analysis and presentation of results. Some reports present only what are termed marginals—the proportion of respondents giving certain answers to each question. If 40 percent favour one candidate, 50 percent another, and 10 percent are undecided, these figures are marginals. Usually, however, a number of cross tabulations are also given. These may show, for instance, that Candidate A's support comes disproportionately from Jewish groups, and Candidate B's, from Irish groups. Sometimes a cross tabulation will substantially change the meaning of survey results. A poll may seem to show that one candidate is the favourite of black voters and another of white voters. But if the preferences of poor respondents and well-to-do respondents are analyzed separately, it may turn out that Candidate A is actually supported by most poor people and Candidate B by most well-to-do people. The most important factor in determining voting intention may thus not be whether a respondent is white or black but whether he is well-to-do or poor.

Assessing a survey's reliability

In judging the overall reliability of a survey, it is advisable to scrutinize at least eight factors: (1) the identity of the sponsor and the past record of reliability of the organization conducting the poll; (2) the exact wording of the questions used; (3) the care with which the population sampled has been defined; (4) the size of the sample and the method by which it was chosen; (5) the "completion rate," or proportion of the sample that actually responded (this is especially important in mail polls, in which frequently fewer than half of those in the sample respond); (6) the degree to which particular results are based on

the whole sample or on small parts of it; (7) the way in which the interviewing was done (whether by telephone, mail, or face-to-face); and (8) the time that the survey was taken (intervening events frequently make people change their opinions).

Criticisms and justifications of opinion surveys. There have been many criticisms of public-opinion polling. Among these are that people are asked to give opinions on matters about which they are not competent to judge, that polling interferes with the democratic process, and that survey research causes annoyance and invasion of privacy.

It is often pointed out that most members of the public are not familiar with the details of complex policies such as those governing tariffs or missile defense systems. Therefore, it is argued, opinion researchers should not ask questions about such subjects. The results at best could be meaningless and at worst misleading, since respondents may be reluctant to admit that they are ignorant. Critics also refer to the fact that many people hold inconsistent opinions, as shown by the polls themselves. The same person may favour larger government expenditures and simultaneously be opposed to higher taxes.

Poll takers usually acknowledge that these problems exist but maintain that they can be overcome by careful survey procedures and by proper interpretation of results. It is common for surveys to include "filter" questions, which help to separate those who are familiar with an issue from those who are not. Thus, the interviewer might first inquire: "Have you heard or read about the government's policy on the tariff?" Then the interviewer would ask only those who answered "yes" whether they were or were not in favour of the policy advocated by the government. Sometimes polls include factual questions that help to assess knowledge, such as "Can you tell me how the veto power in the United Nations Security Council works?" Furthermore, argue the researchers, if people are ignorant, or if they hold inconsistent opinions, this should be known. It is not possible to raise the level of information if areas of ignorance or inconsistency are not identified.

The bandwagon effect

Critics allege also that election polls create a "bandwagon effect"—that people want to be on the winning side and therefore switch their votes to the candidates whom the polls show to be ahead. They complain that surveys undermine representative democracy, since issues should be decided by elected representatives on the basis of the best judgment and expert testimony—not on the basis of popularity contests. They point out that some well-qualified candidates may decide not to run for office because the polls indicate that they have little chance of winning and that a candidate who is far behind in the polls has difficulty in raising funds for campaign expenditures since few contributors want to waste money on a lost cause. Other candidates find out from polls what the public wants and merely pander to popular preferences rather than run on their convictions about what is best for the country.

Those engaged in election research usually concede that polls may dissuade some candidates and also may inhibit campaign contributions. But they also point out that candidates and contributors would have to make their decisions on some basis anyway. If there were no polls, other and less accurate methods would be used to test public sentiment; columnists and political pundits would still make forecasts. As far as the bandwagon effect is concerned, careful studies have failed to show that it exists.

An abuse that is recognized by both critics and poll takers is the practice of leaking to the press partial or distorted results from private polls. A politician may contract privately with a research organization and may then release only results for those areas in which he is ahead, or he may release old results without stating the time when the poll was taken, or he may conceal the fact that a very small sample was used and that the results may have a large margin of error.

Finally, critics aver that the proliferation of opinion polls and market research surveys places an unfair burden on the public. People may be asked to respond to questionnaires that take an hour or more of their time. Pollers may tie up their telephones or occupy their doorsteps for long periods, sometimes asking questions about private matters that are not suitable subjects for public inquiry. But insofar as public resistance to polling is concerned, researchers point out that the "refusal" rate in most surveys is rather low. Most people, in fact, seem to enjoy answering the questions. They also note that, with the use of small samples, it is unlikely that any one individual will be approached very often.

Legislation to deal with these and other problems of poll taking has been proposed in the United States, Britain, and a number of other countries; however, little legislation has been adopted. Survey researchers maintain that such abuses as exist should be dealt with by the profession and by educating the public to evaluate and criticize poll results, rather than by government regulation.

BIBLIOGRAPHY. BERNARD HENNESSEY, *Public Opinion,* 5th ed. (1985) is a general text. The history of public opinion is traced in the appropriate articles in the *Encyclopaedia of the Social Sciences,* vol. 12 (1934, reissued 1967); and the *International Encyclopedia of the Social Sciences,* vol. 13 (1968); and in HAROLD D. LASSWELL, DANIEL LERNER, and HANS SPEIER (eds.), *Propaganda and Communication in World History,* 3 vol. (1979); and VINCENT PRICE, *Public Opinion* (1992).

The development of public opinion polling, academic survey research, and survey research for government is explored in MARTIN BULMER (ed.), *Essays on the History of British Sociological Research* (1985); JEAN M. CONVERSE, *Survey Research in the United States: Roots and Emergence, 1890–1960* (1987); and HERBERT H. HYMAN, *Taking Society's Measure* (1991). Journals include *Public Opinion Quarterly;* and *International Journal of Public Opinion Research* (quarterly).

Defenses of the polling process by some eminent practitioners, albeit with suggestions about how the polls might be improved, are found in GEORGE GALLUP and SAUL FORBES RAE, *The Pulse of Democracy* (1940, reissued 1968); ROBERT M. WORCESTER (ed.), *Political Opinion Polling: An International Review* (1983); LEO BOGART, *Polls and the Awareness of Public Opinion,* 2nd ed. (1988); ROBERT M. WORCESTER, *British Public Opinion* (1991); ALBERT H. CANTRIL, *The Opinion Connection* (1991); and DANIEL YANKELOVICH, *Coming to Public Judgment* (1991).

Academic critics include MICHAEL WHEELER, *Lies, Damn Lies, and Statistics* (1976); and BENJAMIN GINSBERG, *The Captive Public* (1986). RUSSELL J. DALTON, *Citizen Politics in Western Democracies* (1988), offers an excellent comparative study of the impact of public opinion on several Western political parties.

The most innovative and significant works using poll data are V.O. KEY, JR., *Public Opinion and American Democracy* (1961); RONALD INGLEHART, *The Silent Revolution: Changing Values and Political Styles Among Western Publics* (1977); ROBERT CAMERON MITCHELL and RICHARD T. CARSON, *Using Surveys to Value Public Goods* (1989); PAUL M. SNIDERMAN *et al., Reasoning and Choice: Explorations in Political Psychology* (1991); JOHN R. ZALLER, *The Nature and Origins of Mass Opinion* (1992); SHANTO IYENGAR and DONALD R. KINDER, *News That Matters* (1987), on how television news "primes" voters; ELISABETH NOELLE-NEUMANN, *The Spiral of Silence,* 2nd ed. (1993; originally published in German, 1980), which tries to show how perceptions of public opinion themselves shape what individuals say and do; PAUL M. SNIDERMAN and THOMAS PIAZZA, *The Scar of Race* (1993), which uses computer-assisted interviews to show how readily people can be talked out of the positions they have taken and why; THOMAS R. MARSHALL, *Public Opinion and the Supreme Court* (1989); and LAWRENCE R. JACOBS, *The Health of Nations: Public Opinion and the Making of American and British Health Policy* (1993).

Among the recent studies of American elections, which reveal a great deal about public opinion, are NORMAN H. NIE, SIDNEY VERBA, and JOHN R. PETROCIK, *The Changing American Voter,* enlarged ed. (1979); THOMAS E. PATTERSON, *The Mass Media Election* (1980); ERIC R.A.N. SMITH, *The Unchanging American Voter* (1989); and MARTIN P. WATTENBERG, *The Rise of Candidate-Centered Politics* (1991). PAUL R. ABRAMSON, *Political Attitudes in America* (1983); and ROBERT S. ERIKSON, NORMAN R. LUTTBEG, and KENT L. TEDIN, *American Public Opinion,* 5th ed. (1995), are also useful.

Expositions of the sampling method are found in the bibliography for the article STATISTICS. Detailed treatments of the conduct of surveys and key methodological issues can be found in WILLIAM A. BELSON, *The Design and Understanding of Survey Questions* (1981); HOWARD SCHUMAN and STANLEY PRESSER, *Questions and Answers in Attitude Surveys* (1981); CHARLES F. TURNER and ELIZABETH MARTIN (eds.), *Surveying Subjective Phenomena,* 2 vol. (1984); and D.A. DE VAUS, *Surveys in Social Research,* 3rd ed. (1991). NORMAN M. BRADBURN and SEYMOUR SUDMAN, *Polls & Surveys: Understanding What They Tell Us* (1988), offers a good introduction to the interpretation of survey results.

(W.P.D./Ed.)

Public Works

The concept of public works—government-sponsored fixed works constructed for public utility, convenience, or enjoyment—is one of the oldest elements of civilization. Networks of waterways and roads were the lifelines of the ancient empires. Modern public works are indispensable to industry, commerce, and even health, and they provide citizens with facilities and improvements that may be either unavailable or unaffordable through the private sector.

This article covers several of the most important types of civil-engineering projects that are constructed and maintained as public works. These include transportation works such as roads and highways, bridges, tunnels, airports, and lighthouses; hydraulic works such as canals, dams, and harbours; and environmental works such as pollution-control and solid-waste management systems. Several of these works are designed for the use of transportation vehicles, such as automobiles, aircraft, and ships; the history and technology of these vehicles is described in the article TRANSPORTATION.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 733 and 734, and the *Index.* (Ed.)

The article contains the following sections:

Roads and highways 318
  History 318
    Roads of antiquity
    The Middle Ages
    The birth of the modern road
    Roads in the age of the automobile
  The modern road 324
    Road engineering
    Road operation
Bridges 328
  The elements of bridge design 328
    Basic forms
    Materials
    Construction
    Performance in service
  The history of bridge design 331
    Early wood and stone bridges
    Iron and steel bridges, 1779–1929
    Concrete bridges
    Steel bridges after 1931
    Cable-stayed bridges
Airports 341
  Evolution of airports 342
  Modern airports 342
    Site selection
    Airfield layout and configuration
    Runway pavements
    Navigational aids, lighting, and marking
    Air traffic control
    Passenger terminal layout and design
    Cargo facilities
    Links to local ground transportation
    Airport capacity
    Drainage
    Noise
Canals and inland waterways 348
  History 348
    Ancient works
    Medieval revival
    The 16th to 18th century
    The 19th century
  Modern waterway engineering 351
    Characteristics of basic types
    Channels
    Locks
    Inland waterway craft
    Waterway maintenance
  Waterway systems 354
    Administration
    Major inland waterways and networks
    Economic significance
Dams 358
  History 358
    Ancient dams
    Forerunners of the modern dam
  The modern dam 359
    Basic problems in dam design
    Types of dams
Harbours and sea works 363
  Principles of maritime engineering 363
  Sea works for transportation 364
    Classical harbour works
    Breakwaters
    Docks and quays
    Roll-on, roll-off facilities
    Bulk terminals
    Dry docks
  Sea works for reclamation and conservancy 372
    Dredging
    The Delta Plan
    Desalinization
Lighthouses 374
  History of lighthouses 374
    Lighthouses of antiquity
    Medieval lighthouses
    The beginning of the modern era
  Modern lighthouses 375
    Construction
    Illuminants
    Optical equipment
    Intensity, visibility, and character of lights
    Sound signals
    Radio aids
    Automation
    Floating lights
    National lighthouse systems
Environmental works 380
  Municipal water supply 380
    Historical background
    Water sources
    Water requirements
    Water treatment
    Water distribution
  Water-pollution control 387
    Historical background
    Sources of pollution
    Sewage characteristics
    Wastewater collection systems
    Wastewater treatment and disposal
    Sludge treatment and disposal
  Solid-waste management 392
    Historical background
    Solid-waste characteristics
    Solid-waste collection
    Solid-waste treatment
  Hazardous-waste management 396
    Hazardous-waste characteristics
    Transport of hazardous waste
    Treatment, storage, and disposal
    Remedial action
Tunnels and underground excavations 397
  History 398
    Ancient tunnels
    From the Middle Ages to the present
  Tunneling techniques 399
    Basic tunneling system
    Modern soft-ground tunneling
    Modern rock tunneling
  Underground excavations and structures 406
    Rock chambers
    Shafts
    Immersed-tube tunnels
  Future trends in underground construction 410
    Environmental and economic factors
    Potential applications
    Improved technology
Bibliography 413

ROADS AND HIGHWAYS

The terms street, road, and highway define those traveled ways on which animals, people, and wheeled vehicles have moved throughout recorded history. The most ancient name for these arteries of travel seems to be the antecedent of the modern "way." Way stems from the Middle English *wey,* which in turn branches from the Latin *veho* ("I carry"), derived from the Sanskrit *vah* ("carry," "go," or "move"). The word highway goes back to the elevated Roman roads that had a mound or hill formed by earth from the side ditches thrown toward the centre, thus "high" "way." The word street originates with the Latin *strata* (initially, "paved") and later *strata via* ("a way paved with stones"). "Street" was used by the Anglo-Saxons for all the roads that they inherited from the Romans. By the Middle Ages, constructed roads were to be found only in the towns, and so "street" took on its modern limited application to town roads. The more recent word road, derived from the Old English word *rād* ("to ride") and the Middle English *rode* or *rade* ("a mounted journey"), is now used to indicate all vehicular ways.

Modern roads can be classified by type or function. The basic type is the conventional undivided two-way road. Beyond this are divided roads, expressways (divided roads with most side access controlled and some minor at-grade intersections), and freeways (expressways with side access fully controlled and no at-grade intersections). An access-controlled road with direct user charges is known as a tollway. In the United Kingdom freeways and expressways are referred to as motorways.

Divided roads, expressways, and freeways

Functional road types are local streets, which serve only adjacent properties and do not carry through traffic; collector, distributor, and feeder roads, which carry only through traffic from their own area; arterial roads, which carry through traffic from adjacent areas and are the major roads within a region or population centre; and highways, which are the major roads between regions or population centres.

The first half of this section traces the history of roads from earliest times to the present, exploring the factors that have influenced their development and suggesting that in many ways roads have directly reflected the conditions and attitudes of their times. The road is thus one of the oldest continuous and traceable metaphors for civilization and society. The second half of this section explains the factors behind the design, construction, and operation of a modern road. It is shown that a road must interact closely and carefully with the terrain and community through which it passes, with changing vehicle technology, with information technologies, and with the various abilities, deficiencies, and frailties of the individual driver.

History

ROADS OF ANTIQUITY

Ancient roads of the Mediterranean and Middle East. The first roads were paths made by animals and later adapted by humans. The earliest records of such paths have been found around some springs near Jericho and date from about 6000 BC. The first indications of constructed roads date from about 4000 BC and consist of stone-paved streets at Ur in modern-day Iraq and timber roads preserved in a swamp in Glastonbury, England. During the Bronze Age, the availability of metal tools made the construction of stone paving more feasible; at the same time, demand for paved roads rose with the use of wheeled vehicles, which were well established by 2000 BC.

Cretan stone roads. At about this time the Minoans on the island of Crete built a 30-mile (50 kilometre) road from Gortyna on the south coast over the mountains at an elevation of about 4,300 feet (1,300 metres) to Knossos on the north coast. Constructed of layers of stone, the roadway took account of the necessity of drainage by a crown throughout its length and even gutters along certain sections (see Figure 1). The pavement, which was about 12 feet (360 centimetres) wide, consisted of sandstone

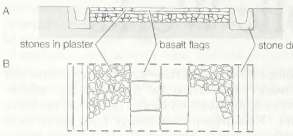


Figure 1: Ancient Cretan stone road, showing (A) cross section and (B) surface view.

Adapted from Hermann Schreiber, *History of Roads (Sinfonieder Strasse)*, copyright © 1959 Econ-Verlag GmbH, used by permission

bound by a clay-gypsum mortar. The surface of the central portion consisted of two rows of basalt slabs 2 inches (50 millimetres) thick. The centre of the roadway seems to have been used for foot traffic and the edges for animals and carts. It is the oldest existing paved road.

Roads of Persia and Babylon. The earliest long-distance road was a 1,500-mile route between the Persian Gulf and the Mediterranean Sea. It came into some use about 3500 BC, but it was operated in an organized way only from about 1200 BC by the Assyrians, who used it to join Susa, near the Persian Gulf, to the Mediterranean ports of Smyrna (İzmir) and Ephesus. More a track than a constructed road, the route was duplicated between 550 and 486 BC by the great Persian kings Cyrus II and Darius I in their famous Royal Road. Like its predecessor, the Persian Royal Road began at Susa, wound northwestward to Arbela, and thence proceeded westward through Nineveh to Harran, a major road junction and caravan centre. The main road then continued to twin termini at Smyrna and Ephesus. The Greek historian Herodotus, writing in about 475 BC, put the time for the journey from Susa to Ephesus at 93 days, although royal riders traversed the route in 20 days.

The Persian Royal Road

In Babylon about 615 BC the Chaldeans connected the city's temples to the royal palaces with a major Processional Way, a road in which burned bricks and carefully shaped stones were laid in bituminous mortar.

Egypt. Herodotus credits the Egyptians with building their first roads to provide a solid track upon which to haul the immense limestone blocks used in the pyramids, and archaeological evidence indicates that such road-building took place southwest of Cairo between 2600 and 2200 BC. The wheel arrived in Egypt at the relatively late date of about 1600 BC. There is little evidence of street surfacing in ancient Egyptian towns, though there is evidence of the use of paved processional roads leading to the temples. The ancient travel routes of Egypt ran from Thebes and Coptos on the central Nile east to the Red Sea and from Memphis (Cairo) across the land bridge to Asia Minor.

Greece. The early Greeks depended primarily on sea travel. There is evidence of the building of special roads for

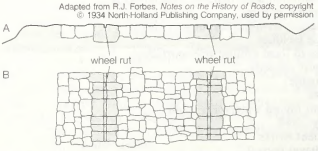


Figure 2: (A) Cross section and (B) surface view of an ancient Greek road.

religious purposes and transport about 800 BC, but there is little evidence of substantial road building for travel and transport prior to the Roman system. The Greeks did build a few ceremonial, or "sacred," roads, paved with shaped stone and containing wheel ruts about 55 inches (140 centimetres) apart (Figure 2).

Ancient roads of Europe. *The Amber Routes.* During the 2nd millennium BC, trade ways developed in Europe.

One route, for example, ran between Italy and Spain via Marseille and nearby Heraclea, close to present-day Avignon. Such ways were used for the movement of flints from Denmark, freestone from Belgium, salt from Austria, lead and tin from England, and amber from northern Europe. By about 1500 BC many of the ways in eastern and central Europe had linked together into an extensive trading network known as the Amber Routes. Four routes have been identified, the first from modern Hamburg southwestward by dual routes through Cologne and Frankfurt to Lyon and Marseille. The second also passed from Hamburg south to Passau on the Danube and then through the Brenner Pass to Venice. The third began at Samland on the East Prussian Coast (where amber is still found), crossed the Vistula River at Thorn, and thence continued southeastward through the Moravian Gate to Aquileia on the Adriatic. The fourth, the Baltic-Pontus road, followed the main eastern rivers, the Vistula, Saw, Sereth, Prut, Bug, and Dnieper.

While the Amber Routes were not roads in the modern sense, they were improved at river crossings, over mountain passes, and across wet and swampy areas. A few remnants of these roads survive today. They were constructed

Log roads

by laying two or three strings of logs in the direction of the road on a bed of branches and boughs up to 20 feet (6 metres) wide. This layer was then covered with a layer of transverse logs 9 to 12 feet in length laid side by side. In the best log roads, every fifth or sixth log was fastened to the underlying subsoil with pegs. There is evidence that the older log roads were built prior to 1500 BC. They were maintained in a level state by being covered with sand and gravel or sod. In addition, the Romans used side ditches to reduce the moisture content and increase the carrying capacity.

The Roman roads. The greatest systematic road builders of the ancient world were the Romans, who were very conscious of the military, economic, and administrative advantages of a good road system. The Romans drew their expertise mainly from the Etruscans—particularly in cement technology and street paving—though they probably also learned skills from the Greeks (masonry), Cretans, Carthaginians (pavement structure), Phoenicians, and Egyptians (surveying). Concrete made from cement was a major development that permitted many of Rome's construction advances.

The Romans began their road-making task in 334 BC and by the peak of the empire had built nearly 53,000 miles of road connecting their capital with the frontiers of their far-flung empire (see Figure 3, map). Twenty-nine great military roads, the *viae militares,* radiated from Rome. The most famous of these was the Appian Way. Begun in 312 BC, this road eventually followed the Mediterranean coast south to Capua and then turned eastward to Beneventum, where it divided into two branches, both reaching Brundisium (Brindisi). From Brundisium the Appian Way traversed the Adriatic coast to Hydruntum, a total of 410 miles from Rome.

The typical Roman road was bold in conception and construction. Where possible, it was built in a straight line from one sighting point to the next, regardless of obstacles, and was carried over marshes, lakes, ravines, and mountains. In its highest stage of development, it was constructed by excavating parallel trenches about 40 feet apart to provide longitudinal drainage—a hallmark of Roman road engineering. The foundation was then raised about three feet above ground level, employing material taken from the drains and from the adjacent cleared ground. As the importance of the road increased, this embankment was progressively covered with a light bedding of sand or mortar on which four main courses were constructed: (1) the *statumen* layer 10 to 24 inches (250 to 600 millimetres) thick, composed of stones at least 2 inches in size, (2) the *rudus,* a 9-inch-thick layer of concrete made from stones under 2 inches in size, (3) the *nucleus* layer, about 12 inches thick, using concrete made from small gravel and coarse sand, and, for very important roads, (4) the *summum dorsum,* a wearing surface of large stone slabs at least 6 inches deep. The total thickness thus varied from 3 to 6 feet. The width of the Appian Way in its ultimate development was 35 feet. The two-way, heavily crowned central carriageway was 15 feet wide. On each side it was flanked by curbs 2 feet wide and 18 inches high and paralleled by one-way side lanes 7 feet wide. This massive Roman road section, adopted about 300 BC,

Pavement of Roman roads



Figure 3: The road network of the Roman Empire; (inset) cross section of a typical Roman road.

set the standard of practice for the next 2,000 years. (See Figure 3, inset.)

The public transport of the Roman Empire was divided into two classes: (1) *cursus rapidi,* the express service, and (2) *agnarie,* the freight service. In addition, there was an enormous amount of travel by private individuals. The most widely used vehicles were the two-wheeled chariot drawn by two or four horses and its companion, the cart used in rural areas. A four-wheeled *raeda* in its passenger version corresponded to the stage coaches of a later period and in its cargo version to the freight wagons. Fast freight *raedae* were drawn by 8 horses in summer and 10 in winter and, by law, could not haul in excess of 750 pounds, or 330 kilograms. Speed of travel ranged from a low of about 15 miles per day for freight vehicles to 75 miles per day by speedy post drivers.

Ancient roads of South and East Asia. *India.* The Indus civilization in Sindh, Balochistān, and the Punjab probably flourished in the period 3250–2750 BC. Excavations indicate that the cities of this civilization paved their major streets with burned bricks cemented with bitumen. Great attention was devoted to drainage. The houses had drain pipes that carried the water to a street drain in the centre of the street, two to four feet deep and covered with slabs or bricks.

Evidence from archaeological and historical sources indicates that by AD 75 several methods of road construction were known in India. These included the brick pavement, the stone slab pavement, a kind of concrete as a foundation course or as an actual road surface, and the principles of grouting (filling crevices) with gypsum, lime, or bituminous mortar. Street paving seems to have been common in the towns in India at the beginning of the Common era, and the principles of drainage were well known. The crowning of the roadway and the use of ditches and gutters was common in the towns. Northern and western India in the period 300 to 150 BC had a network of well-built roads. The rulers of the Mauryan empire (4th century BC), which stretched from the Indus River to the Brahmaputra River and from the Himalayas to the Vindhya Range, generally recognized that the unity of a great empire depended on the quality of its roads. The Great Royal Road of the Mauryans began at the Himalayan border, ran through Taxila (near modern Rāwalpindi), crossed the five streams of the Punjab, proceeded by way of Jumna to Prayag (now Allahābād), and continued to the mouth of the Ganges River (see Figure 4). A "Ministry of Public Works" was responsible for construction, marking, and maintenance of the roads and rest houses and for the smooth running of ferries.

The Mauryan Royal Road

China's Imperial Highway. China had a road system that paralleled the Persian Royal Road and the Roman road network in time and purpose. Its major development began under Emperor Shih huang-ti about 220 BC. Many of the roads were wide, surfaced with stone, and lined with trees; steep mountains were traversed by stone-paved stairways with broad treads and low steps. By AD 700 the network had grown to some 40,000 kilometres. Traces of a key route near Sian are still visible.

The Silk Road. The trade route from China to Asia Minor and India, known as the Silk Road, had been in existence for 1,400 years at the time of Marco Polo's travels (*c.* AD 1270–90). It came into partial existence about 300 BC, when it was used to bring jade from Khotan (modern Ho-t'ien) to China. By 200 BC it was linked to the West, and by 100 BC it was carrying active trade between the two civilizations. At its zenith in AD 200 this road and its western connections over the Roman system constituted the longest road on earth. In Asia the road passed through Samarkand to the region of Fergana, where, near the city of Osh, a stone tower marked the symbolic watershed between East and West. From Fergana the road traversed the valley between the Tien Shan and Kunlun mountains through Kashgar, where it divided and skirted both sides of the Takla Makan Desert to join again at Ansi. The road then wound eastward to Chia-yü-kwan (Su-chou), where it passed through the westernmost gateway (the Jade Gate or Yumen) of the Great Wall of China. It then went southeast on the Imperial Highway to Sian and eastward to Shanghai on the Pacific Ocean. From Kashgar, trade routes to the south passed over the mountains to the great trading centre of Bactria and to northern Kashmir.

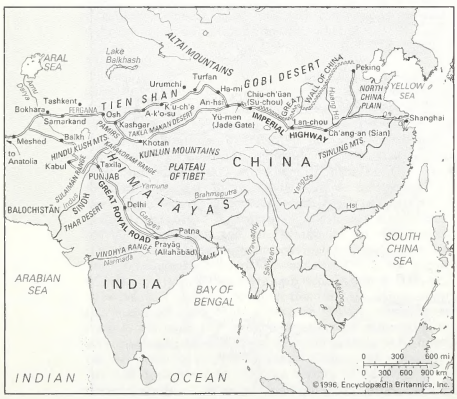


Figure 4: The route of the Silk Road circa the 13th century. Associated royal roads in South and East Asia are shown.

THE MIDDLE AGES

Europe and Asia. At the zenith of the Roman Empire, overland trade joined the cultures of Europe, North Africa, Asia Minor, China, and India. But the system of road transport was dependent on the Roman, Chinese, and Mauryan empires, and as these great empires declined in the early Christian era the trade routes became routes of invasion. Except in the Byzantine Empire, road networks fell into centuries of disrepair. Transport relied on pack trains, which could negotiate the badly maintained roads and sufficed to carry the reduced stream of commerce.

Revival of road building

The first signs of a road revival came during the reign of Charlemagne late in the 8th century. In the 9th century the Moors established an extensive street network in Córdoba, Spain. The Vikings operated the Varangian Road, a major trade route linking the Baltic and the Middle East via Russia. Further road revival was aided first by the need to service the regular round of trade fairs and then, in the 11th century, by a centralization of power and an increase in religious fervour.

Eventually a commercial revival set in. By the 12th century old cities were reviving and new ones were being built, especially in western Europe. Street paving became a reputable artisan activity, and by the 15th century well-maintained roads bringing food to the cities from their hinterlands became of critical importance. At the same time, wheeled vehicles increased in number and quality. There was an awakened interest in better overland travel, better protection of merchants and other travelers, and the improvement of roads. Public funds, chiefly derived from tolls, were committed to road upkeep. The corvée, or road-labour tax, made an even more substantial contribution. Long-distance overland commerce increased rapidly and included a restoration of the trade route between Europe and China through Central Asia that Marco Polo traveled in the late 13th century.

Inca roads of South America. Across the Atlantic, the period witnessed the rise of another notable road-building empire, that of the Incas. The Inca road system extended from Quito, Ecuador, through Cuzco, Peru, and as far south as Santiago, Chile. It included two parallel roadways, one along the coast about 2,250 miles in length, the other following the Andes about 3,400 miles in length with a number of cross connections. At its zenith, when the Spaniards arrived early in the 16th century, a network of some 14,000 miles of road served an area of about 750,000 square miles (1,940,000 square kilometres) in

which lived nearly 10 million people. The network was praised by 16th-century explorers as superior to that in contemporary Europe.

The Andes route was remarkable. The roadway was 25 feet wide and traversed the loftiest ranges. It included galleries cut into solid rock and retaining walls built up for hundreds of feet to support the roadway. Ravines and chasms were filled with solid masonry, suspension bridges with wool or fibre cables crossed the wider mountain streams, and stone surfacing was used in difficult areas. The steeper gradients were surmounted by steps cut in the rocks. Traffic consisted entirely of pack animals (llamas) and people on foot; the Inca lacked the wheel. Yet they operated a swift foot courier system and a visual signaling system along the roadway from watchtower to watchtower.

THE BIRTH OF THE MODERN ROAD

The master road builders. In Europe, gradual technological improvements in the 17th and 18th centuries saw increased commercial travel, improved vehicles, and the breeding of better horses. These factors created an incessant demand for better roads, and supply and invention both rose to meet that demand. In 1585 the Italian engineer Guido Toglietta wrote a thoughtful treatise on a pavement system using broken stone that represented a marked advance on the heavy Roman style. In 1607 Thomas Procter published the first English-language book on roads. The first highway engineering school in Europe, the School of Bridges and Highways, was founded in Paris in 1747. Late in the 18th century the Scottish political economist Adam Smith, in discussing conditions in England, wrote,

The use of broken stone

> Good roads, canals, and navigable rivers, by diminishing the expense of carriage, put the remote parts of the country more nearly upon a level with those in the neighbourhood of a town. They are upon that account the greatest of all improvements.

Up to this time roads had been built, with minor modifications, to the heavy Roman cross section, but in the last half of the 18th century the fathers of modern road building and road maintenance appeared in France and Britain.

Trésaguet. In France in 1764, Pierre-Marie-Jérôme Trésaguet, an engineer from an engineering family, became engineer of bridges and roads at Limoges and, in 1775, inspector general of roads and bridges for France. In that year he developed an entirely new type of relatively light road surface, based on the theory that the underlying natural formation, rather than the pavement, should support the load. His standard cross section, 18 feet wide, consisted of an eight-inch-thick course of uniform stones laid edgewise on the natural formation and covered by a two-inch layer of walnut-sized broken stone. The second layer was topped with a one-inch layer of smaller broken stone. In order to maintain surface levels, Trésaguet's pavement was placed in an excavated trench—a technique that made drainage a difficult problem.

Telford. Thomas Telford, born of poor parents in Dumfriesshire, in Scotland, in 1757, was apprenticed to a stone mason; intelligent and ambitious, he progressed to designing bridges and building roads. Telford placed great emphasis on two features: (1) maintaining a level roadway with a maximum gradient of 1 in 30 and (2) building a stone surface capable of carrying the heaviest anticipated loads. His roadways were 18 feet wide and built in three courses: (1) a lower layer, seven inches thick, consisting of good-quality stone carefully placed by hand (this was known as the Telford base), (2) a middle layer, also seven inches thick, consisting of broken stone of two-inch maximum size, and (3) a layer of gravel one inch thick.

McAdam. The greatest advance came from John Loudon McAdam, born in 1756 at Ayr in Scotland. McAdam began his road-building career in 1787 but reached major heights after 1804, when he was appointed general surveyor for Bristol, then the most important port city in England. The roads leading to Bristol were in poor condition, and in 1816 McAdam took control of the Bristol Turnpike. There he showed that traffic could be supported by a relatively thin layer of small, single-sized, angular pieces of broken stone placed and compacted on a well-drained natural formation and covered by an impermeable surface of smaller stones. He had no use for the masonry constructions of his predecessors and contemporaries.

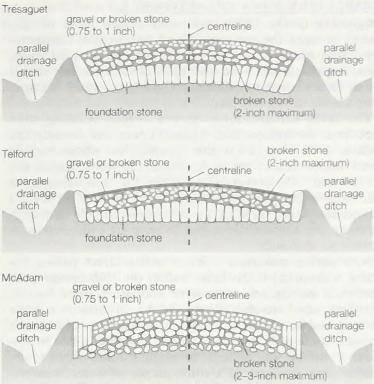


Figure 5: Cross sections of roads as designed by (top) Pierre Trésaguet, (middle) Thomas Telford, and (bottom) John McAdam.
Encyclopædia Britannica, Inc.

Drainage was essential to the success of McAdam's method, and he required the pavement to be elevated above the surrounding surface. The structural layer of broken stone was eight inches thick and used stone of two to three inches maximum size laid in layers and compacted by traffic—a process adequate for the traffic of the time. The top layer was two inches thick, using three-quarter-inch stone to fill surface voids between the large stones. Continuing maintenance was essential. The macadam road

Although McAdam drew on the successes and failures of others, his total structural reliance on broken stone represented the largest paradigm shift in the history of road pavements. The principles of the "macadam" road are still used today. McAdam's success was also due to his efficient administration and his strong view that road managers needed skill and motivation.

Early U.S. road systems. *The Lancaster Turnpike.* The first engineered and planned road in the United States was the Lancaster Turnpike, a privately constructed toll road built between 1793 and 1795. Connecting Philadelphia and Lancaster in Pennsylvania, its 62-mile length had a maximum grade of 7 percent and was surfaced with broken stone and gravel in a manner initially uninfluenced by the work of Telford and McAdam. However, pavement failures in 1796 led to the introduction of some of the new European methods.

The Cumberland Road. The Cumberland Road, also known as the National Pike, was an even more notable road-building feat. It had been advocated by both George Washington and Thomas Jefferson to aid western expansion and national unity. Work commenced in 1811, and the road opened for traffic between Cumberland and Wheeling, W.Va., in 1818. By 1838 it extended to Springfield, Ohio, and part of the way to Vandalia, Ill. Specification requirements called for a 66-foot right-of-way completely cleared. The roadway was to be covered 20 feet in width with stone 18 inches deep at the centre and 12 inches deep at the edge. The upper six inches was to consist of broken stone of three-inch maximum size and the lower stratum of stone of seven-inch maximum size. The road was constructed by the federal government, much of the finance being raised by land sales. Although maintenance was funded by tolls and federal appropriations, the road surface began to deteriorate in the 1820s. Federal funding ceased in 1838, and in 1841 the project was abandoned at Vandalia for political and practical reasons.

ROADS IN THE AGE OF THE AUTOMOBILE

Beginning in the 1840s, the rapid development of railroads brought the construction of lightweight Trésaguet-McAdam roads to a virtual halt. For the next 60 years, road improvements were essentially confined to city streets or to feeder roads to railheads. Other rural roads became impassable in wet weather.

The initial stimulus for a renewal of road building came not from the automobile, whose impact was scarcely felt before 1900, but from the bicycle, for whose benefit road improvement began in many countries during the 1880s and '90s. Nevertheless, while the requirements of the lightweight, low-speed bicycle were satisfied by the old "macadamized" surfaces, as the world entered the 20th century the automobile began to raise its own seemingly insatiable demands.

New paving materials. When urban street paving became widespread in the latter half of the 19th century, the common paving materials were hoof-sized stone blocks, similarly sized wooden blocks, bricks, McAdam's broken stones, and occasionally asphalt and concrete. McAdam's broken stone provided the cheapest pavement, but its unbound surface was difficult to maintain and was usually either slimy or dusty as a consequence of water, weather, and copious amounts of horse excrement. Thus, roads at the turn of the 20th century were largely inadequate for the demands about to be placed on them by the automobile and truck. As vehicle speeds increased rapidly, the available friction between road and tire became critical for accelerating, braking, and cornering. In addition, numerous pavement failures made it obvious that much stronger and tougher materials were required. The result was an ongoing search for a better pavement. Asphalt and concrete both offered promise.

Asphalt and concrete

Asphalt is a mixture of bitumen and stone, and concrete is a mixture of cement and stone. Asphalt footpaths were first laid in Paris in 1810, but the method was not perfected until after 1835. The first road use of asphalt occurred in 1824, when asphalt blocks were placed on the Champs-Élysées in Paris, but the first successful major application was made in 1858 on the nearby rue Saint-Honoré. The first successful concrete pavement was built in Inverness, Scot., in 1865. Neither technology, however, advanced far without the pressures of the car, and they both required the availability of powerful stone-crushing, mixing, and spreading equipment.

The impetus for the development of modern road asphalt came from the United States, which had few deposits of natural bitumen to draw upon and where engineers were therefore forced to study the principles behind the behaviour of this material. The first steps came in the 1860s, with the work of Belgian immigrant Edward de Smedt at Columbia University in New York City. De Smedt conducted his first tests in New Jersey in 1870 and by 1872 was producing the equivalent of a modern "well-graded" maximum-density asphalt. The first applications were in Battery Park and on Fifth Avenue in New York City in 1872. De Smedt went to Washington, D.C., in 1876 as part of President Ulysses S. Grant's desire to make that town "a Capital City worthy of a great Nation." Grant had appointed a commission to oversee road making, and it conducted its first trials on Pennsylvania Avenue in 1877. Sixty percent of the trials used de Smedt's new product and were great successes.

In 1887 de Smedt was followed as inspector of asphalts and cements by Clifford Richardson, who set about the task of codifying the specifications for asphalt mixes. Richardson basically developed two forms of asphalt: asphaltic concrete, which was strong and stiff and thus provided structural strength; and hot-rolled asphalt, which contained more bitumen and thus produced a far smoother and better surface for the car and bicycle.

One of the great convenient coincidences of asphalt development was that the automobile ran on gasoline, which at that time was simply a by-product of the distillation of kerosene from petroleum. Another by-product was bitumen. Until that time, most manufacturers had used coal tar (a by-product of the making of gas from coal) as the binder for road asphalt. As the demand for automobile fuel increased, however, so did the availability of bitumen and, hence, of good asphalt designed to the standards of de Smedt and Richardson. This gave U.S. road builders a major advantage over their European counterparts, who were still wedded to the virtues of the various natural asphalts, such as those from Neuchâtel, Switz., and the island of Trinidad.

Development of modern pavement

Richardson published a standard textbook on asphalt paving in 1905, and the practice did not change greatly thereafter. The biggest change was in the machinery available to produce, place, and finish the material rather than in the product itself. Toward the end of the century, there were major movements toward the use of recycled asphalt, chemical modifiers for improving bitumen properties, and small fibres for improving crack resistance. In addition, developments in testing and structural analysis made it possible to design an asphalt pavement as a sophisticated structural composite.

The first modern concrete roads were produced by Joseph Mitchell, a follower of Telford, who conducted three successful trials in England and Scotland in 1865–66. Like asphalt technology, concrete road building was largely developed by the turn of the 20th century and was restricted more by the available machinery than by the material. Problems were also encountered in producing a surface that could match the performance of the surface produced almost accidentally by hot-rolled asphalt. For the following century the two materials remained in intense competition, both offering a similar product at a similar cost, and there was little evidence that one would move far ahead of the other as they continued on their paths of gradual improvement. (The principles of modern pavement design are described below in *The modern road: Road engineering: Road design: Pavement*.)

Changes in finance. *From corvée to toll.* Through the millennia, responsibility for financing and building roads and highways had been both a local and a national responsibility in the nations of the world. It is notable that this responsibility had changed along with political attitudes toward road building and had not rested easily with any party. Many roads initially had been built to provide rulers with a means of conquest, control, and taxation; in periods of peace, the same rulers usually tried to pass the maintenance responsibilities on to local authorities, adjoining landowners, or the travelers who used the road. Local authorities and landowners usually fulfilled their responsibilities via the corvée, in which people were required to donate their labour to road work. Corvée was always unpopular and unproductive, but it was nevertheless more effective than attempts at direct taxation.

The last option, charging the traveler, gave rise to the toll road, a system that blossomed with the Industrial Revolution. Private turnpike trusts dominated British road building and maintenance throughout the 19th century, eventually covering 15 percent of the entire network. In the United States many toll roads were constructed in the first half of the 19th century under charters granted by the states.

From local to national funding. Thus, through the 19th century most road building was administered and financed on a local basis. British road building remained entirely local despite clear evidence that local responsibility was not providing adequate roads. The national government edged into the picture only through increased pressure from the cyclists, climaxed by the establishment in 1909 of a national Road Board authorized to construct and maintain new roads and to make advances to highway authorities to build new or improve old roads.

Except for the National Pike, early highway building in the United States was also carried on by local government. Congress made a number of land grants for the opening of wagon roads but exercised no control over the expenditure of funds—with the result that, as in Britain, little road building was accomplished.

Beginnings of the U.S. interstate system

In 1891 New Jersey enacted a law providing for state aid to the counties and established procedures for raising money at the township and county levels for road building. In 1893 Massachusetts established the first state highway commission. By 1913 most of the states had adopted

similar legislation, and by 1920 all states had their own road organization. However, there was little coordination among the states. National funding began in 1912 with the Post Office Appropriation Act, and the Federal Aid Road Act of 1916 established federal aid for highways as a national policy. The Bureau of Public Roads, established in the Department of Agriculture in 1893 to make "inquiries with regard to road management," was given responsibility for the program, and an apportionment formula based on area, population, and mileage of post roads in each state was adopted. Funds were allocated for construction costs, with the states being required to bear all maintenance costs. The location and selection of roads to be improved was left to the states, an arrangement that had some shortcomings.

Since 1892 a national Good Roads Movement had lobbied for a system of national roads joining the major population centres and contributing to the national economy. This point of view was recognized by the Federal Aid Highway Act of 1921, which required each state to designate a system of state highways not to exceed 7 percent of the total highway mileage in each state. Federal-aid funding was limited to this system, which was not to exceed three-sevenths of total highway mileage. Bureau of Public Roads approval of the system was required, and federal aid was limited to 50 percent of the estimated cost.

New highways. *The parkway.* The achievement of such a system in the automobile age required a new form of road. This grew from the parkway, which had many historical precedents but was introduced in its modern form in 1858 with the work of the landscape architects Frederick Law Olmsted and Calvert Vaux for Central Park in New York City. The concept was given further prominence by William Niles White of New York as a part of the Bronx River protection program of New York City and Westchester County. The 15-mile, four-lane single carriageway known as the Bronx River Parkway was built between 1916 and 1925. Protected on both sides by broad bands of parkland that limited access, the highway was located and designed so as to cause minimum disturbance to the landscape. Its use was restricted to passenger cars, and at-grade intersections were avoided. The success of the concept led to the creation of the Westchester County parkway system and the Long Island State Park Commission. More parkways were built in the New York area, including the Merritt Parkway (1934–40), which continued the Westchester Parkway System across Connecticut as a toll road providing divided roadways and limited access.

The freeway. The success of the parkway system led to the introduction of the freeway, which is a divided highway with no conflicting traffic movements and no access from adjoining properties. Between 1913 and 1921 a group called AVUS had built 10 kilometres (6 miles) of parkway through the Grunewald park in Berlin. Their successful experience led to the world's first full freeway being built from Cologne to Bonn between 1929 and 1932. In 1933 Adolf Hitler began construction of an integrated freeway network known as the *Reichsautobahnen,* or "national motor roads," beginning with the Frankfurt-Darmstadt-Mannheim-Heidelberg Autobahn. One purpose of the program was to alleviate unemployment, but the roads also appealed to German nationalism and had a strong militaristic intent. The entire system included three north-south routes and three east-west routes. The highway provided separate 7.5-metre (25-foot) carriageways divided by a median strip of 5 metres (16 feet). The roads were designed for large traffic volumes and speeds in excess of 150 kilometres per hour, bypassing cities and providing limited access. About 1,000 kilometres (600 miles) were completed by 1936, and 6,500 kilometres (4,000 miles) were in use when construction ceased in 1942.

The autobahn

The viability of the freeway concept in the United States was demonstrated by the Pennsylvania Turnpike. The Pennsylvania Turnpike Commission, established in 1937 to raise funds and build a toll road across the Appalachian Mountains, found an unusually favourable situation in the form of an abandoned railroad right-of-way, with many tunnels and excellent grades over much of the route that allowed the tollway to be completed in 1940 to freeway standards. The turnpike provided two 24-foot carriageways and a 10-foot median with no cross traffic at grade and with complete control of access and egress at 11 traffic interchanges. Its alignment and grades were designed for high volumes of high-speed traffic and its pavement to accommodate the heaviest trucks. The favourable public reaction to this new type of highway provided the impetus for the post-World War II toll-road boom in the United States, advanced the start of a major interstate highway program, and influenced highway developments elsewhere. The Pennsylvania Turnpike, originally running from Harrisburg to Pittsburgh, was later extended 100 miles east to Philadelphia and 67 miles west to the Ohio border, making it 327 miles long. An original feature of the turnpike, later widely copied, was the provision of restaurant and fueling facilities.

National and international highway systems. The Romans had realized that a coordinated system of roadways connecting the major areas of their empire would be of prime significance for both commercial and military purposes. In the modern era, the nations of Europe first introduced the concept of highway systems. In France, for example, the State Department of Roads and Bridges was organized in 1716, and by the middle of the 18th century the country was covered by an extensive network of roads built and maintained primarily by the national government. In 1797 the road system was divided into three classes of descending importance: (1) roads leading from Paris to the frontiers, (2) roads leading from frontier to frontier but not passing through Paris, and (3) roads connecting towns. By the early 1920s this general plan re-

Archive Photos

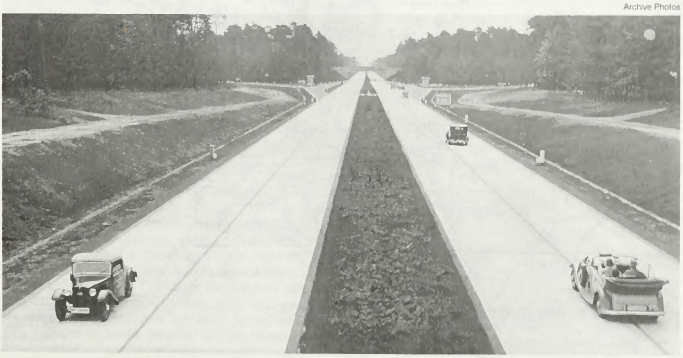


Figure 6: A section of the German autobahn shortly after its completion in the mid-1930s.

mained essentially the same except that a gradual change in class and responsibility had taken place. At that time the road system was divided into four classes: (1) national highways, improved and maintained by the national government, (2) regional highways, improved and maintained by the department under a road service bureau appointed by the Department Commission, (3) main local roads, connecting smaller cities and villages, built and maintained from funds of the communes supplemented by grants from the department, and (4) township roads, built and maintained by the communities alone.

The United Kingdom. While the British recognized the necessity for national support of highways and a national system as early as 1878, it was the Ministry of Transport Act of 1919 that first classified the roadway system into 23,230 miles of Class I roads and 14,737 miles of Class II roads. Fifty percent of the cost of Class I roads and 25 percent of the cost of Class II roads were to be borne by the national government. In the mid-1930s the need for a national through-traffic system was recognized, and the Trunk Roads Act of 1939, followed by the Trunk Roads Act of 1944, created a system of roadways for through traffic. The Special Roads Act of 1949 authorized existing or new roads to be classified as "motorways" that could be reserved for special classes of traffic. The Highways Act of 1959 swept away all previous highway legislation in England and Wales and replaced it with a comprehensive set of new laws.

The U.S. Interstate Highway System

The United States and Canada. The mammoth U.S. Interstate Highway System (formally, the National System of Interstate and Defense Highways) developed in response to strong public pressures in the 1950s for a better road system. These pressures culminated in the establishment by President Dwight Eisenhower of the Clay Committee in 1954. Following this committee's recommendations, the Federal Aid Highway Act and the Highway Revenue Act of 1956 provided funding for an accelerated program of construction. A federal gasoline tax was established, the funds from which, with other highway-user payments, were placed in a Highway Trust Fund. The federal-state ratio for funding construction of the Interstate System was changed to 90 percent federal and 10 percent state. It was expected that the system would be completed no later than 1971, but cost increases and planning delays extended this time by some 25 years. The system grew to a total length of more than 45,000 miles, connecting nearly all the major cities in the United States and carrying more than 20 percent of the nation's traffic on slightly more than 1 percent of the total road and street system.

The Canadian Highway Act of 1919 provided for a system of 40,000 kilometres (13,000 miles) of highways and provided for a federal allotment for construction not to exceed 40 percent of the cost. By the end of the century, more than 134,000 kilometres (83,000 miles) of highway had been built, of which approximately 16,000 kilometres (9,900 miles) were freeway.

The modern road

ROAD ENGINEERING

Since the beginning of the 20th century, as the automobile and truck have offered ever higher levels of mobility, vehicle ownership per head of population has increased. Road needs have been strongly influenced by this popularity and also by the mass movement of people to cities and thence to suburban fringes—a trend that has led to increasing travel needs and road congestion and to low-density cities, which are difficult to service by public transport. Often the building of new roads to alleviate such problems has encouraged further urban sprawl and yet more road travel. Long-term solutions require the provision of alternatives to car and truck transport, controls over land use, and the proper pricing of road travel. To this end, road managers must be concerned not merely with lines on maps but



Figure 7: Major national highways of the continental United States and southern Canada.

also with the number, type, speed, and loading of individual vehicles, the safety, comfort, and convenience of the traveling public, and the health and welfare of bystanders and adjoining property owners.

Ideally, the development of a major road system is an orderly, continuous process. The process follows several steps: assessing road needs and transport options; planning a system to meet those needs; designing an economically, socially, and environmentally acceptable set of roads; obtaining the required approval and financing; building, operating, and maintaining the system; and providing for future extensions and reconstruction.

Planning. Road needs are closely associated with the relative location of centres of population, commerce, industry, and transportation. Traffic between two centres is approximately proportional to their populations and inversely proportional to the distance between them. Estimating traffic on a route thus requires a prediction of future population growth and economic activity, an estimation of their effects on land use and travel needs, and a knowledge of any potential transport alternatives. The key variables defining road needs are the traffic volumes, tonnages, and speeds to be expected throughout the road's life.

Once the traffic demand has been estimated, it is necessary to predict the extent of the road works needed to handle that traffic. A starting point in these calculations is offered by surveys of the origins, destinations, and route choices of present traffic; computer models are then used to estimate future traffic volumes on each proposed route. Estimates of route choice are based on the understanding that most drivers select their estimate of the quickest, shortest, or cheapest route. Consideration in planning is also given to the effect of new traffic on existing streets, roads, and parking provisions.

Where feasible, the next step in planning a road system is to refine the selected route to a narrow corridor. The various alignment options are drawn, considering the local terrain and conditions. The economic, social, and environmental benefits and costs of these options are discussed with relevant official and community groups until an acceptable specific route is determined.

Road design. *Alignment and profile.* After a route has been selected, a three-dimensional road alignment and its associated cross-sectional profiles are produced. In order to reduce the amount of earth to be moved, the alignment is adjusted where practical so that the earth to be excavated is in balance with the embankments to be built. Computers allow many options to be explored and realistic views of the future road to be examined.

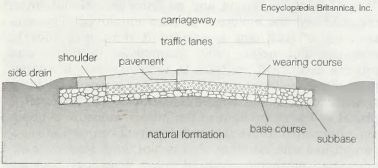


Figure 8: Schematic diagram of the elements of a modern roadway.

Parts of the road surface

In order fully to understand the design stage, a few standard terms must be defined (see Figure 8). A traffic lane is the portion of pavement allocated to a single line of vehicles; it is indicated on the pavement by painted longitudinal lines or embedded markers. The shoulder is a strip of pavement outside an outer lane; it is provided for emergency use by traffic and to protect the pavement edges from traffic damage. A set of adjoining lanes and shoulders is called a roadway or carriageway, while the pavement, shoulders, and bordering roadside up to adjacent property lines is known as the right-of-way.

In order to maintain quality and uniformity, design standards are established for each functional road type. The number of traffic lanes is directly determined by the combination of traffic volume and speed, since practical limits on vehicle spacing means that there is a maximum number of vehicles per hour that pass through a traffic lane. The width of lanes and shoulders, which must strike a balance between construction cost and driver comfort, allows the carriageway width to be determined. Standards also specify roadside barriers or give the clear transverse distances needed on either side of the carriageway in order to provide safety in the event that vehicles accidentally leave the carriageway. Thus it is possible to define the total right-of-way width needed for the entire road, although intersections will add further special demands.

Design standards also help to determine the actual alignment of the road by specifying, for each design speed, the minimum radius of horizontal curves, the maximum vertical gradient, the clearance under bridges, and the distance a driver must be able to see the pavement ahead in order to stop or turn aside.

Pavement. Road traffic is carried by the pavement, which in engineering terms is a horizontal structure supported by in situ natural material. In order to design this structure, existing records must be examined and subsurface explorations conducted. The engineering properties of the local rock and soil are established, particularly with respect to strength, stiffness, durability, susceptibility to moisture, and propensity to shrink and swell over time. The relevant properties are determined either by field tests (typically by measuring deflection under a loaded plate or the penetration of a rod), by empirical estimates based on the soil type, or by laboratory measurements. The material is tested in its weakest expected condition, usually at its highest probable moisture content. Probable performance under traffic is then determined. Soils unsuitable for the final pavement are identified for removal, suitable replacement materials are earmarked, the maximum slopes of embankments and cuttings are established, the degree of compaction to be achieved during construction is determined, and drainage needs are specified.

In a typical rural pavement, the base course provides the required supplement to the strength, stiffness, and durability of the natural formation. Its thickness ranges from 4 inches (10 centimetres) for very light traffic and a good natural formation to more than 40 inches (100 centimetres) for heavy traffic and a poor natural formation. The subbase is a protective layer and temporary working platform sometimes placed between the base course and the natural formation.

Structure of asphalt pavement

Pavements are called either flexible or rigid, according to their relative flexural stiffness. Flexible pavements have base courses of broken stone pieces either compacted into place in the style of McAdam or glued together with bitumen to form asphalt. In order to maintain workability, the stones are usually less than 1.5 inches in size and often less than 1 inch. Initially the bitumen must be heated to temperatures of 300°–400° F (150°–200° C) in order to make it fluid enough to mix with the stone. At the road site a paving machine places the hot mix in layers about twice the thickness of the stone size. The layers are then thoroughly rolled before the mix cools and solidifies. In order to avoid the expense of heating, increasing use has been made of bitumen emulsions or cutbacks, in which the bitumen binder is either treated with an emulsifier or thinned with a lighter petroleum fraction that evaporates after rolling. These treatments allow asphalts to be mixed and placed at ambient temperatures.

The surface layer of a flexible pavement protects the underlying base course from traffic and water while also providing adequate tire friction, generating minimal noise

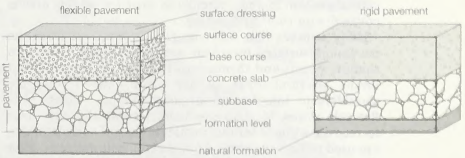


Figure 9: *Cross sections of modern pavements.* (Left) Flexible asphalt-based pavement; (B) rigid portland-cement concrete pavement.

in urban areas, and giving suitable light reflectance for night-time driving. Such surfaces are provided either by a bituminous film coated with stone (called a spray-and-chip seal) or by a thin asphalt layer. The spray-and-chip seal is used over McAdam-style base courses for light to moderate traffic volumes or to rehabilitate existing asphalt surfaces. It is relatively cheap, effective, and impermeable and lasts for about 10 years. Its main disadvantage is its high noise generation. Maintenance usually involves a further spray coating with bitumen. Asphalt surfacing is used with higher traffic volumes or in urban areas. Surfacing asphalt commonly contains smaller and more wear-resistant stones than the base course and employs relatively more bitumen. It is better able to resist horizontal forces and produces less noise than a spray-and-chip seal.

Rigid pavements are made of portland cement concrete. The concrete slab ranges in thickness from 6 to 14 inches. It is laid by a paving machine, often on a supporting layer that prevents the pressure caused by traffic from pumping water and natural formation material to the surface through joints and cracks. Concrete shrinks as it hardens, and this shrinkage is resisted by friction from the underlying layer, causing cracks to appear in the concrete. Cracking is usually controlled by adding steel reinforcement in order to enhance the tensile strength of the pavement and ensure that any cracking is fine and uniformly distributed. Transverse joints are sometimes also used for this purpose. Longitudinal joints are used at the edge of the construction run when the whole carriageway cannot be cast in one pass of the paving machine.

In places where the local natural material is substandard for use as a base course, it can be "stabilized" with relatively small quantities of lime, portland cement, pozzolana, or bitumen. The strength and stiffness of the mix are increased by the surface reactivity of the additive, which also reduces the material's permeability and hence its susceptibility to water. Special machines distribute the stabilizer into the upper 8 to 20 inches of soil.

In deciding whether to use a flexible, rigid, or stabilized pavement, engineers take into account lifetime cost, riding characteristics, traffic disruptions due to maintenance, ease and cost of repair, and the effect of climatic conditions. Often there is little to choose between rigid and flexible pavements.

The properties of the base course material are usually determined by laboratory tests, although field tests are sometimes conducted to check that the construction process has achieved the designer's intent. Designers typically consider the possibility of structural failure resulting from a single overload and also from damage accumulating under the passage of many routine loads. Both of these types of failure are almost entirely caused by trucks.

Drainage. Adequate drainage is the single most important element in pavement performance, and drainage systems can be extensive and expensive. Drainage involves handling existing watercourses, removing water from the pavement surface, and controlling underground water in the pavement structure. In designing the system, the engineer first selects the "design storm"—that is, the most severe flood that can be expected in a nominated period of time (as much as 100 years for a major road or as little as 5 years for a minor street carrying local traffic). The drainage system must be able to carry the storm water produced by this design storm without flooding the roadway or adjacent property. In areas where land use is changing from agricultural to residential or commercial, peak flows will increase notably as the surrounding area is covered with roofs and paving.

The "design storm"

Safety requires that water be rapidly removed from the pavement surface. In urban areas, the water runs into shallow gutters and thence into the inlets of underground drains. In rural areas, surface water flows beyond the shoulders to longitudinal drainage ditches, which have flat side slopes to enable vehicles leaving the pavement to recover without serious incident. Cut-off surface drains are used to prevent water from flowing without restriction down the slopes of cuttings and embankments.

Vertical drainage layers, formed from single-sized aggregate or special sheets called geofabrics and geomembranes, are used to prevent groundwater from seeping laterally into the pavement structure. In addition, a horizontal drainage layer is often inserted between base course and natural ground in order to remove water from the pavement structure and stop upward capillary movement of any natural groundwater. Underground drains can also be used to lower the groundwater level by both preventing water entry and removing water that does enter the pavement structure.

Financing. The full design of a proposed road is analyzed with respect to its costs and its economic, social, and environmental effects. It may also be subjected to public review. This step can be lengthy, as new roads are usually popular with the traveling public but sometimes cause distress in the communities through which they pass.

Local streets and collector roads are usually administered by local governments and financed by local taxes. Arterial roads and highways, however, need a wider administrative and financial input in order to guarantee route continuity and uniformity. Since the 1920s the financing of roads has been largely transferred to the road user. A variety of taxes is employed: on fuel and oil, on road usage, on vehicle purchase and ownership, on driver licensing, on truck mass and mass times distance traveled, on tire and accessory purchases, and on the economic benefits provided by roads (*e.g.,* higher property values or increased productivity). Fuel taxes usually provide the simplest source of revenue, but they are not necessarily intended solely for expenditure on roads. Many local roads are funded by property taxes.

Construction. After the road has been approved and financing found, surveyors define its three-dimensional location on the ground. Forming of the in-situ material to its required shape and installation of the underground drainage system can then begin. Imported pavement material is placed on the natural formation and may have water added; rollers are then used to compact the material to the required density. If possible, some traffic is permitted to operate over the completed earthwork in order to detect weak spots.

In countries where labour is inexpensive and less skilled, traditional manual methods of road construction are still commonplace. However, the developed world relies heavily on purpose-built construction plant. This can be divided into equipment for six major construction purposes: clearing, earthmoving, shaping, and compacting the natural formation; installing underground drainage; producing and handling the road-making aggregate; manufacturing asphalt and concrete; placing and compacting the pavement layers; and constructing bridges and culverts.

For clearing vegetation and undesirable materials from the roadway, the bulldozer is often employed. The construction of rock cuts is commonly done with shovels, draglines, and mobile drills. Shaping the formation and moving earth from cuttings to embankments is accomplished with bulldozers, graders, hauling scrapers, elevating graders, loaders, and large dump trucks. The material is placed in layers, brought to the proper moisture content, and compacted to the required density. Compaction is accomplished with tamping, sheeps-foot, grid, steel-wheeled, vibrating, and pneumatic-tired rollers. Backhoes, back actors, and trenchers are used for drainage work.

In order to avoid high haulage costs, the materials used for base course construction are preferably located near the construction site; it is economically impossible to use expensive materials for long lengths of road construction. The excavation process is the same as for rock cuts, although rippers may be used for obtaining lower-grade material. Crushers, screens, and washers produce stone of the right size, shape, and cleanliness.

The placement of paving material increasingly involves a paving machine for distributing the aggregate, asphalt, or concrete uniformly and to the required thickness, shape, and width (typically, one or two traffic lanes). The paving machine can slipform the edges of the course, thus avoiding the need for fixed side-forms. As it progresses down the road, it applies some preliminary compaction and also screeds and finishes the pavement surface. In modern machines, level control is by laser sighting.

The paving machine

In producing a spray-and-chip seal surface (or a bituminous surface treatment), a porous existing surface is covered with a film of hot, fluid bitumen that is sprayed in sufficient quantity to fill voids, cracks, and crevices without leaving excess bitumen on the surface. The surface is then sprayed with a more viscous hot bitumen, which is immediately covered with a layer of uniform-size stone chips spread from a dump truck. The roadway is then rolled to seat the stone in the sticky bitumen, and excess stone is later cleared by a rotary broom.

Maintenance. The life of a road structure depends on the quality of its maintenance and minor renovation. Maintenance keeps the roadway safe, provides good driving conditions, and prolongs the life of the pavement, thus protecting the road investment. Maintenance consists of activities concerned with the condition of the pavement, shoulders, drainage, traffic facilities, and right-of-way. It includes the prompt sealing of cracks and filling of potholes to prevent water entering through the surface, the removal of trash thrown on the wayside by the traveling public, and the care of pavement markings, signs, and signals. In rigorous winter climates, substantial effort is required to remove snow and ice from the pavement, to scatter salt for snow and ice removal, and to spread sand for better traction.

ROAD OPERATION

Traffic management. Road users are subject to traffic control via instructions and information provided by roadway markings, signs, and signals, and they are subject to legal control via the rules of the road (particularly those concerned with vehicular priority).

Traffic control. The marking of roadway surfaces with painted lines and raised permanent markers is commonplace and effective, despite high maintenance costs and visibility problems at night, in heavy traffic, and in rain or snow. A solid line is a warning or instruction not to cross, and a broken line is for guidance. Thus, solid lines indicate dangerous conditions (such as restricted sight distance where overtaking would be dangerous), pavement edges, stop lines, and turning lanes at intersections; broken lines indicate interior lane lines and centre lines on two-way roads where the sight distance is good. Lines are usually white, but yellow is used for centre lines in North America.

Traffic signs

Signs advise the driver of special regulations and provide information about hazards and navigation. They are classified as regulatory signs, which provide notice of traffic laws and regulations (*e.g.,* signs for speed limits and for stop, yield or give-way, and no entry); warning signs, which call attention to hazardous conditions (*e.g.,* sharp curves, steep grades, low vertical clearances, and slippery surfaces); and guide signs, which give route information (*e.g.,* numbers or designations, distances, directions, and points of interest).

Signs have standard shapes and colours—for instance, the red octagon used for the stop sign, the triangle for warning signs, the green rectangle with white lettering for freeway directional signs (commonly mounted over the roadway and of large size for easy reading at high speeds). Tourist signs are brown rectangles, and special shapes and colours are used for route markers. Many signs, such as the stop sign, are universally used, but there are some differences between the two common international systems based on either the American or the European practice. Basically, these differences are derived from a complete reliance on symbolic signs and a greater range of blue guidance signs in multilingual Europe.

Traffic signals are primarily used to control traffic in urban street systems—particularly at conventional intersections accommodating large traffic volumes, where they allocate right-of-way to the various traffic streams. They can also meter traffic entering access lanes onto busy freeways or to indicate the lanes to use on two-way roads. Simple traffic signals work on preset timing plans that vary with the time of day. More advanced traffic-actuated signals automatically monitor the traffic streams and allocate right-of-way accordingly. Signals can also be linked to a computer so that traffic traveling along a major route can receive a continuous wave of green signals, obtaining maximum traffic output from the system.

Legal control. Legal rules governing the movement of traffic are an essential part of order on the road. The rules may be divided into three categories. First are those applying to the vehicle and the driver, such as vehicle and driver registration, vehicle safety equipment and roadworthiness, accident reporting, financial liability, and truck weights and axle loads (to protect pavements and bridges from damage). Second are the movement rules for drivers and pedestrians, known as the rules of the road; these dictate which side of the road to use, maximum speeds, right-of-way, and turning requirements. Third are those regulations that apply to limited road sections, indicating speed limits, one-way operations, and turning controls.

GOMACO International

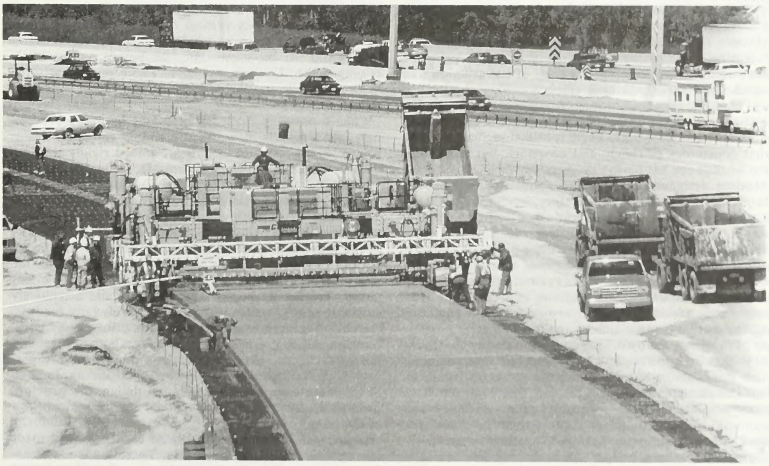


Figure 10: A slipform paving machine pouring concrete over steel reinforcement on a highway in Ontario in Canada.

Rules of the road

The important rules of the road are reasonably uniform throughout the world. For instance, in most countries drivers must give right-of-way to vehicles on their right. However, in practice the stop and yield (or give-way) signs have commonly supplanted the right-of-way rule. Speed limits vary greatly with jurisdiction, ranging from walking pace in a Dutch *woonerf,* or "shared" street, to unrestricted on a German autobahn. Speed limits are commonly reduced on roads approaching residential, shopping, or school areas and on dangerous road sections and sharp curves.

Special regulations are important for the efficient movement of traffic in specific segments of a street and road system. For instance, one-way streets in congested urban areas may provide safer driving conditions and increase the traffic-carrying capacity of the system. The provision of special turn arrows in traffic signals or the prohibition of turns at intersections contribute to safety, increase traffic throughput, and reduce conflict.

Safety. Traffic police (or road patrols or highway police) help improve road safety and traffic flow by enforcing driving regulations. They also regulate traffic at the scene of an accident and investigate accidents. Traffic enforcement has been aided by the use of technology—cameras, radar, video, and inductance loops—to detect and record traffic offenders automatically.

An important aspect of traffic regulation and accident prevention is the control of excessive speed, which contributes significantly to the number and severity of road crashes. Speed is commonly measured by radar devices or by pacing with a patrol car. In crash investigations, the speed of the cars is determined by the length of skid marks. Another key factor in road accidents is the influence of alcohol and drugs. Tests for intoxication are now widely conducted; the most common is the breath test, in which the driver blows into a device that analyzes the alcohol content of the breath and indicates the approximate blood alcohol level. Many authorities believe that 0.50 gram of alcohol per litre of blood is a realistic limit for ordinary motorists, but that zero levels should be demanded for critical operators such as drivers of public transport vehicles.

Road safety can also be built into the road. Divided roads are many times safer than two-way roads. Crash severity can be reduced by the use of "soft" signs and light poles and by guardrails and impact attenuators in front of fixed roadside objects such as bridge piers and the noses at the exit ramps of a freeway. Better road surfaces, alignments, signing, and marking improve driving conditions and increase road safety.

Nevertheless, about 90 percent of crashes are primarily due to human error. Many crashes have been attributed to simple inattention or failure to see warnings. Alcohol, fatigue, inexperience, aggression, and excessive risk taking are the most common crash causes involving behavioral changes in drivers. Lack of driving skills is rarely an issue; most drivers do not need training as much as they need education and experience. Meanwhile, road engineers must design road systems that attempt to reduce the frequency and impact of human error. (F.J.B./M.G.L.)

BRIDGES

A bridge is a structure that spans horizontally between supports, and its function is to carry vertical loads. The prototypical bridge is quite simple—two supports holding up a beam—yet the engineering problems that must be overcome even in this simple form are inherent in every bridge: the supports must be strong enough to hold the structure up, and the span between supports must be strong enough to carry the loads. Spans are generally made as short as possible; long spans are justified where good foundations are limited—for example, over estuaries with deep water.

Three goals of bridge design: efficiency, economy, elegance

All major bridges are built with the public's money. Therefore, bridge design that best serves the public interest has a threefold goal: to be as efficient, as economical, and as elegant as is safely possible. Efficiency is a scientific principle that puts a value on reducing materials while increasing performance. Economy is a social principle that puts value on reducing the costs of construction and maintenance while retaining efficiency. Finally, elegance is a symbolic or visual principle that puts value on the personal expression of the designer without compromising performance or economy. There is little disagreement over what constitutes efficiency and economy, but the definition of elegance has always been controversial.

Modern designers have written about elegance or aesthetics since the early 19th century, beginning with the Scottish engineer Thomas Telford. Bridges ultimately belong to the general public, which is the final arbiter of this issue, but in general there are three positions taken by professionals. The first principle holds that the structure of a bridge is the province of the engineer and that beauty is fully achieved only by the addition of architecture. The second idea, arguing from the standpoint of pure engineering, insists that bridges making the most efficient possible use of materials are by definition beautiful. The third case holds that architecture is not needed but that engineers must think about how to make the structure beautiful. This last principle recognizes the fact that engineers have many possible choices of roughly equal efficiency and economy and can therefore express their own aesthetic ideas without adding significantly to materials or cost.

Generally speaking, bridges can be divided into two categories: standard overpass bridges or unique-design bridges over rivers, chasms, or estuaries. This section describes features common to both types, but it concentrates on the unique bridges because of their greater technical, economic, and aesthetic interest.

The elements of bridge design

BASIC FORMS

There are six basic bridge forms: the beam, the truss, the arch, the suspension, the cantilever, and the cable-stay. Figure 11 illustrates these forms and indicates the way loads are carried by showing the approximate location of compression (where material is squeezed together) and tension (where material is stretched).

Compression and tension

Beam. The beam bridge is the most common bridge form. A beam carries vertical loads by bending. As the beam bridge bends, it undergoes horizontal compression on the top. At the same time, the bottom of the beam is subjected to horizontal tension. The supports carry the loads from the beam by compression vertically to the foundations.

When a bridge is made up of beams spanning between only two supports, it is called a simply supported beam bridge. If two or more beams are joined rigidly together over supports (as shown in Figure 11), the bridge becomes continuous.

Truss. A single-span truss bridge (as shown in Figure 11) is like a simply supported beam because it carries vertical loads by bending. Bending leads to compression in the top chords (or horizontal members), tension in the bottom chords, and either tension or compression in the vertical and diagonal members, depending on their orientation. Trusses are popular because they use a relatively small amount of material to carry relatively large loads.

Arch. The arch bridge carries loads primarily by compression, which exerts on the foundation both vertical and horizontal forces. Arch foundations must therefore prevent both vertical settling and horizontal sliding. In spite of the more complicated foundation design, the structure itself normally requires less material than a beam bridge of the same span.

Suspension. A suspension bridge carries vertical loads through curved cables in tension. These loads are transferred both to the towers, which carry them by vertical compression to the ground, and to the anchorages, which

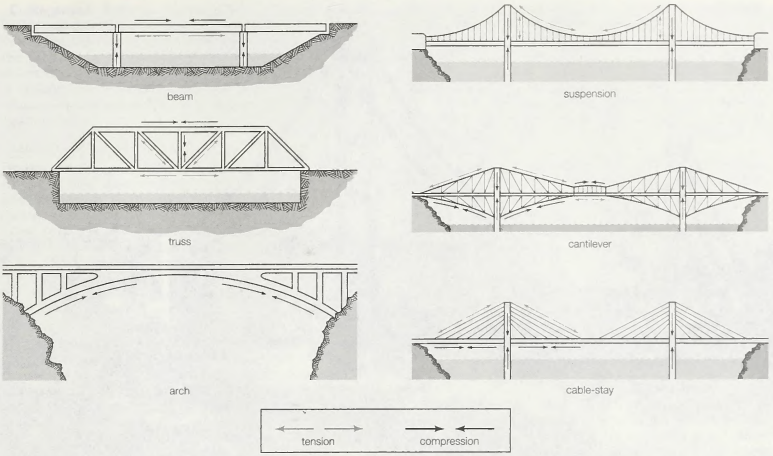


Figure 11: Six major bridge types, with forces of tension represented by light arrows and forces of compression by dark arrows.

must resist the inward and sometimes vertical pull of the cables. The suspension bridge can be viewed as an upside-down arch in tension with only the towers in compression. Because the deck is hung in the air, care must be taken to ensure that it does not move excessively under loading. The deck therefore must be either heavy or stiff or both.

Tension and compression in the cantilever bridge

Cantilever. A beam is said to be cantilevered when it projects outward, supported only at one end. A cantilever bridge is generally made with three spans, of which the outer spans are both anchored down at the shore and cantilever out over the channel to be crossed. The central span rests on the cantilevered arms extending from the outer spans (as in Figure 11); it carries vertical loads like a simply supported beam or a truss—that is, by tension forces in the lower chords and compression in the upper chords. The cantilevers carry their loads by tension in the upper chords and compression in the lower ones. Inner towers carry those forces by compression to the foundation, and outer towers carry the forces by tension to the far foundations.

Cable-stay. Cable-stayed bridges carry the vertical main-span loads by nearly straight diagonal cables in tension. The towers transfer the cable forces to the foundations through vertical compression. The tensile forces in the cables also put the deck into horizontal compression.

MATERIALS

The four primary materials used for bridges have been wood, stone, iron, and concrete. Of these, iron has had the greatest influence on modern bridges. From iron, steel is made, and steel is used to make reinforced and prestressed concrete. Modern bridges are almost exclusively built with steel, reinforced concrete, and prestressed concrete.

Wood and stone. Wood is relatively weak in both compression and tension, but it has almost always been widely available and inexpensive. Wood has been used effectively for small bridges that carry light loads, such as footbridges. Engineers now incorporate laminated wooden beams and arches into some modern bridges.

Stone is strong in compression but weak in tension. Its primary application has been in arches, piers, and abutments.

Iron and steel. The first iron used during the Industrial Revolution was cast iron, which is strong in compression but weak in tension. Wrought iron, on the other hand, is as strong in compression as cast iron, but it also has much greater tensile strength. Steel is an even further refinement of iron and is yet stronger, superior to any iron in both tension and compression. Steel can be made to varying strengths, some alloys being five times stronger than others. The engineer refers to these as high-strength steels.

Concrete. Concrete is an artificial stone made from a mixture of water, sand, gravel, and a binder such as cement. Like stone, it is strong in compression and weak in tension. Concrete with steel bars embedded in it is called reinforced concrete. Reinforcement allows for less concrete to be used because the steel carries all the tension; also, the concrete protects the steel from corrosion and fire.

Post-tensioned prestressing

Prestressed concrete is an important variation of reinforced concrete. A typical process, called post-tensioned prestressing, involves casting concrete beams with longitudinal holes for steel tendons—cables or bars—like reinforced concrete, but the holes for the tendons are curved upward from end to end, and the tendons, once fitted inside, are stretched and then anchored at the ends. The tendons, now under high tension, pull the two anchored ends together, putting the beam into compression. In addition, the curved tendons exert an upward force, and the designer can make this upward force counteract much of the downward load expected to be carried by the beam. Prestressed concrete reduces the amount of steel and concrete needed in a structure, leading to lighter designs that are often less expensive than designs of reinforced concrete.

CONSTRUCTION

Beam bridges. All bridges need to be secure at the foundations and abutments. In the case of a typical overpass beam bridge with one support in the middle, construction begins with the casting of concrete footings for the pier and abutments. Where the soil is especially weak, wooden or steel piles are driven to support the footings. After the concrete piers and abutments have hardened sufficiently, the erection of a concrete or steel superstructure begins. Steel beams are generally made in a factory, shipped to the site, and set in place by cranes. For short spans, steel beams are usually formed as a single unit. At the site, they are placed parallel to each other, with temporary forms between them so that a concrete deck can be cast on top. The beams usually have metal pieces welded on their top flanges, around which the concrete is poured. These pieces provide a connection between beam and slab, thus producing a composite structure.

For longer spans, steel beams are made in the form of plate girders. A plate girder is an I beam consisting of separate top and bottom flanges welded or bolted to a vertical web. While beams for short spans are usually



Figure 12: Construction of the Golden Gate Bridge, San Francisco, in 1936.
Golden Gate Bridge, Highway and Transportation District

of a constant depth, beams for longer spans are often haunched—that is, deeper at the supports and shallower at mid-span. Haunching stiffens the beam at the supports, thereby reducing bending at mid-span.

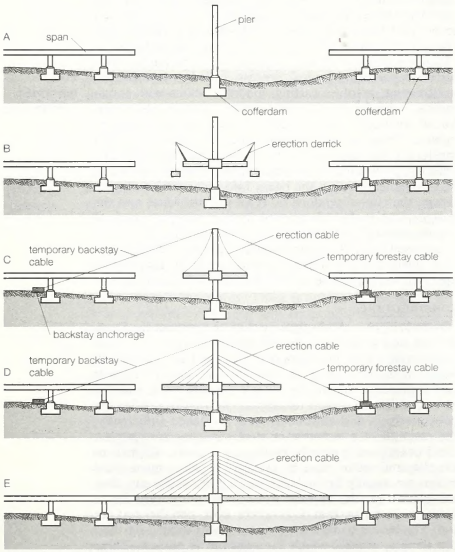


Figure 13: *Five steps in the cantilever-method construction of a cable-stayed bridge.*
(A) Erection of piers and support spans; (B) construction of work station on central pier and beginning of central span; (C) installation of temporary fore- and backstay cables, along with first suspension cables; (D) extension of central span; (E) completion of central span and removal of temporary cables.

Arch bridges. Arches are normally fabricated on-site. After the building of abutments (and piers, if the bridge is multiarch), a falsework is constructed. For a concrete arch, metal or wooden falsework and forms hold the cast concrete and are later removed. For steel arches, a cantilevering method is standard. Each side of an arch is built out toward the other, supported by temporary cables above or by falsework below until the ends meet. At this point the arch becomes self-supporting, and the cables or falsework are removed.

Suspension bridges. When bridges requiring piers are built over a body of water, foundations are made by sinking caissons into the riverbed and filling them with concrete. Cassions are large boxes or cylinders that have been made from wood, metal, or concrete. In the case of suspension bridges, towers are built atop the caissons. The first suspension-bridge towers were stone, but now they are either steel or concrete. Next, the anchorages are built on both ends, usually of reinforced concrete with embedded steel eyebars to which the cables will be fastened. An eyebar is a length of metal with a hole (or "eye") at the ends. Cables for the first suspension bridges were made of linked wrought-iron eyebars; now, however, cables are generally made of thousands of steel wires spun together at the construction site. Spinning is done by rope pulleys that carry each wire across the top of the towers to the opposite anchorage and back. The wires are then bundled and covered to prevent corrosion. When the cables are complete, suspenders are hung, and finally the deck is erected—usually by floating deck sections out on ships, hoisting them with cranes, and securing them to the suspenders.

Spinning of suspension cables

Cantilever bridges. Like suspension bridges, steel cantilever bridges generally carry heavy loads over water, so their construction begins with the sinking of caissons and the erection of towers and anchorages. For steel cantilever bridges, the steel frame is built out from the towers toward the centre and the abutments. When a shorter central span is required, it is usually floated out and raised into place. The deck is added last.

The cantilever method for erecting prestressed concrete bridges consists of building a concrete cantilever in short segments, prestressing each succeeding segment onto the earlier ones. Each new segment is supported by the previous segment while it is being cast, thus avoiding the need for falsework.

Cable-stayed bridges. Construction of cable-stayed bridges usually follows the cantilever method (see Figure 13). After the tower is built, one cable and a section of the deck are constructed in each direction. Each section of the deck is prestressed before continuing. The process is repeated until the deck sections meet in the middle, where they are connected. The ends are anchored at the abutments.

PERFORMANCE IN SERVICE

Bridges are designed, first, to carry their own permanent weight, or dead load; second, to carry traffic, or live loads; and, finally, to resist natural forces such as winds or earthquakes.

Live load and dead load. The primary function of a bridge is to carry traffic loads: heavy trucks, cars, and trains. Engineers must estimate the traffic loading. On short spans, it is possible that the maximum conceivable load will be achieved—that is to say, on spans of less than 100 feet (30 metres), four heavy trucks may cross at the same time, two in each direction. On longer spans of several thousand feet, the maximum conceivable load is such a remote possibility (imagine the Golden Gate Bridge with only heavy trucks crossing bumper-to-bumper in each direction at the same time) that the cost of designing for it is unreasonable. Therefore, engineers use probable loads as a basis for design.

In order to carry traffic, the structure must have some weight, and on short spans this dead load weight is usually less than the live loads. On longer spans, however, the dead load is greater than live loads, and, as spans get longer, it becomes more important to design forms that minimize dead load. In general, shorter spans are built with beams, hollow boxes, trusses, arches, and continuous versions of the same, while longer spans use cantilever, cable-stay, and suspension forms. As spans get longer, questions of shape, materials, and form become increasingly important. New forms have evolved to provide longer spans with more strength from less material.

Wind loads

Forces of nature. Dead and live weight are essentially vertical loads, whereas forces from nature may be either vertical or horizontal. Wind causes two important loads, one called static and the other dynamic. Static wind load is the horizontal pressure that tries to push a bridge sideways. Dynamic wind load gives rise to vertical motion, creating oscillations in any direction. Like the breaking of an overused violin string, oscillations are vibrations that can cause a bridge to fail. If a deck is thin and not properly shaped and supported, it may experience dangerous vertical or torsional (twisting) movements.

The expansion and contraction of bridge materials by heat and cold have been minimized by the use of expansion joints in the deck along with bearings at the abutments and at the tops of piers. Bearings allow the bridge to react to varying temperatures without causing detrimental stress to the material. In arches, engineers sometimes design hinges to reduce stresses caused by temperature movement.

Modern bridges must also withstand natural disasters such as hurricanes and earthquakes. In general, earthquakes are best withstood by structures that carry as light a dead weight as possible, because the horizontal forces that arise from ground accelerations are proportional to the weight of the structure. (This phenomenon is explained by the fundamental Newtonian law of force equals mass times acceleration.) For hurricanes, it is generally best that the bridge be aerodynamically designed to have little solid material facing the winds, so that they may pass through or around the bridge without setting up dangerous oscillations.

The history of bridge design

Modern bridges, the focus of this section, began with the introduction of industrially produced iron. They have evolved over the past 200 years as engineers have come to understand better the new possibilities inherent first in cast iron, then in wrought iron and structural steel, and finally in reinforced and prestressed concrete. These materials have led to bridge designs that have broken completely with the designs in wood or stone that characterized bridges before the Industrial Revolution.

Industrial strength has been an important factor in the evolution of bridges. Great Britain, the leading industrialized country of the early 19th century, built the most significant bridges of that time. Likewise, innovations arose in the United States from the late 19th century through the mid-20th century and in Japan and Germany in subsequent decades. Switzerland, with its highly industrialized society, has also been a fertile ground for advances in bridge building.

EARLY WOOD AND STONE BRIDGES

The ancient world. *Beam bridges.* The first bridges were simply supported beams, such as flat stones or tree trunks laid across a stream. For valleys and other wider

© A.G.E. FOTOSTOCK

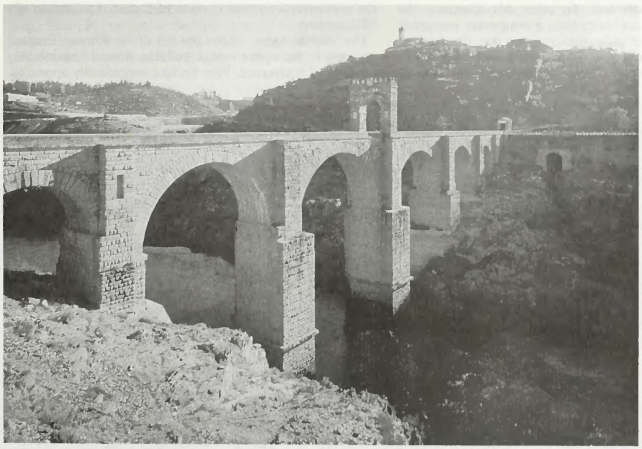


Figure 14: Roman masonry arch bridge, with spans up to 98 feet, built over the Tagus River at Alcántara, Spain, in the early 2nd century AD.

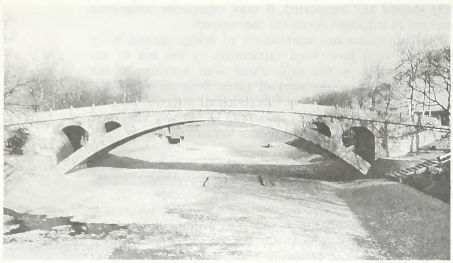


Figure 15: The Great Stone Bridge, built between 589 and 618 in Hopeh Province, China.
© Christopher Liu/ChinaStock Photo Library

channels—especially in East Asia and South America, where examples can still be found—ropes made of various grasses and vines tied together were hung in suspension for single-file crossing. Materials were free and abundant, and there were few labour costs, since the work was done by slaves, soldiers, or natives who used the bridges in daily life.

Clapper bridges

Some of the earliest known bridges are called clapper bridges (from Latin *claperius*, "pile of stones"). These bridges were built with long, thin slabs of stone to make a beam-type deck and with large rocks or blocklike piles of stones for piers. Postbridge in Devon, Eng., an early medieval clapper bridge, is an oft-visited example of this old type, which was common in much of the world, especially China.

Roman arch bridges. The Romans began organized bridge building to help their military campaigns. Engineers and skilled workmen formed guilds that were dispatched throughout the empire, and these guilds spread and exchanged building ideas and principles. The Romans also discovered a natural cement, called pozzolana, which they used for piers in rivers.

Roman bridges are famous for using the circular arch form, which allowed for spans much longer than stone beams and for bridges of more permanence than wood. Where several arches were necessary for longer bridges, the building of strong piers was critical. This was a problem when the piers could not be built on rock, as in a wide river with a soft bed. To solve this dilemma, the Romans developed the cofferdam, a temporary enclosure made from wooden piles driven into the riverbed to make a sheath, which was often sealed with clay. Concrete was then poured into the water within the ring of piles. Although most surviving Roman bridges were built on rock, the Sant'Angelo Bridge in Rome stands on cofferdam foundations built in the Tiber River more than 1,800 years ago.

The Romans built many wooden bridges, but none has survived, and their reputation rests on their masonry bridges. One beautiful example is the bridge over the Tagus River at Alcántara, Spain (see Figure 14). The arches, each spanning 98 feet, feature huge arch stones (voussoirs) weighing up to eight tons each. Typical of the best stone bridges, the voussoirs at Alcántara were so accurately shaped that no mortar was needed in the joints. This bridge has remained standing for nearly 2,000 years. Another surviving monument is the Pont du Gard aqueduct near Nîmes in southern France, completed in AD 14. This structure, almost 900 feet long, has three tiers of semicircular arches, with the top tier rising more than 150 feet above the river. The bottom piers form diamond-shaped points, called cutwaters, which offer less resistance to the flow of water.

Asian cantilever and arch bridges. In Asia, wooden cantilever bridges were popular. The basic design used piles driven into the riverbed and old boats filled with stones sunk between them to make cofferdam-like foundations. When the highest of the stone-filled boats reached above the low-water level, layers of logs were crisscrossed in such a way that, as they rose in height, they jutted farther out toward the adjacent piers. At the top, the Y-shaped, cantilevering piers were joined by long tree trunks. By crisscrossing the logs, the builders allowed water to pass through the piers, offering less resistance to floods than with a solid design. In this respect, these designs presaged some of the advantages of the early iron bridges.

In parts of China many bridges had to stand in the spongy silt of river valleys. As these bridges were subject to an unpredictable assortment of tension and compression, the Chinese created a flexible masonry-arch bridge. Using thin, curved slabs of stone, the bridges yielded to considerable deformation before failure.

The Great Stone Bridge

In the Great Stone Bridge in Chao-chou, Hopeh Province, China, built by Li Ch'un between 589 and 618, the single span of 123 feet has a rise of only 23 feet from the abutments to the crown (see Figure 15). This rise-to-span ratio of 1:5, much lower than the 1:2 ratio found in semicircular arches, produced a large thrust against the abutments. To reduce the weight, the builders made the spandrels (walls between the supporting vault and deck) open. The Great Stone Bridge thus employed a form rarely seen in Europe prior to the mid-18th century, and it anticipated the reinforced-concrete designs of Robert Maillart in the 20th century.

The Middle Ages. After the fall of the Roman Empire, progress in European bridge building slowed considerably until the Renaissance. Fine bridges sporadically appeared,

© Roger-Viollet

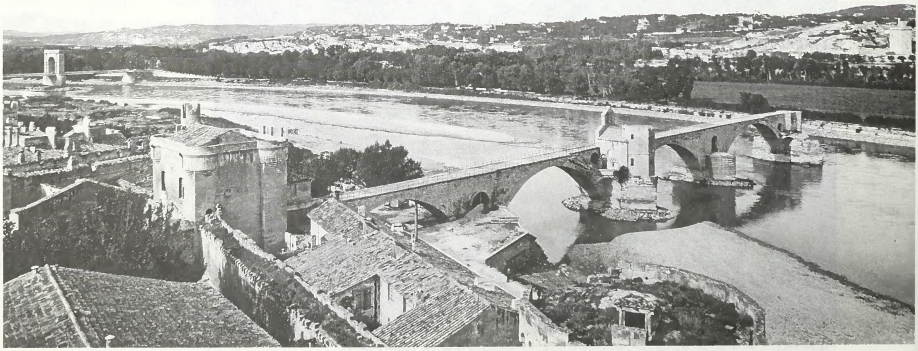


Figure 16: The Saint-Bénézet Bridge, built over the Rhône River at Avignon, Fr., in 1177–88.

however. Medieval bridges are particularly noted for the ogival, or pointed arch. With the pointed arch the tendency to sag at the crown is less dangerous, and there is less horizontal thrust at the abutments. Medieval bridges served many purposes. Chapels and shops were commonly built on them, and many were fortified with towers and ramparts. Some featured a drawbridge, a medieval innovation. The most famous bridge of that age was Old London Bridge, begun in the late 12th century under the direction of a priest, Peter of Colechurch, and completed in 1209, four years after his death. London Bridge was designed to have 19 pointed arches, each with a 24-foot span and resting on piers 20 feet wide. There were obstructions encountered in building the cofferdams, however, so that the arch spans eventually varied from 15 to 34 feet. The uneven quality of construction resulted in a frequent need for repair, but the bridge held a large jumble of houses and shops and survived more than 600 years before being replaced.

A more elegant bridge of the period was the Saint-Bénézet Bridge at Avignon, Fr. Begun in 1177, part of it still stands today (see Figure 16). Another medieval bridge of note is Monnow Bridge in Wales, which featured three separate ribs of stone under the arches. Rib construction reduced the quantity of material needed for the rest of the arch and lightened the load on the foundations.

The Renaissance and after. *Stone arch bridges.* During the Renaissance, the Italian architect Andrea Palladio took the principle of the truss, which previously had been used for roof supports, and designed several successful wooden bridges with spans up to 100 feet. Longer bridges, however, were still made of stone. Another Italian designer, Bartolommeo Ammannati, adapted the medieval ogival arch by concealing the angle at the crown and by starting the curves of the arches vertically in their springings from the piers. This elliptical shape of arch, in which the rise-to-span ratio was as low as 1:7, became known as basket-handled and has been adopted widely since. Ammannati's elegant Santa Trinità Bridge (1569) in Florence, with two elliptical arches, carried pedestrians and later automobiles until it was destroyed during World War II; it was afterward rebuilt with many of the original materials recovered from the riverbed Yet another Italian, Antonio da Ponte, designed the Rialto Bridge (1591) in Venice, an ornate arch made of two segments with a span of 89 feet and a rise of 21 feet. Antonio overcame the problem of soft, wet soil by having 6,000 timber piles driven straight down under each of the two abutments, upon which the masonry was placed in such a way that the bed joints of the stones were perpendicular to the line of thrust of the arch. This innovation of angling stone or concrete to the line of thrust has been continued into the present.

By the middle of the 18th century, bridge building in masonry reached its zenith. Jean-Rodolphe Perronet, builder of some of the finest bridges of his day, developed very flat arches supported on slender piers. His works included the Pont de Neuilly (1774), over the Seine, the Pont Sainte-Maxence (1785), over the Oise, and the beautiful Pont de la Concorde (1791), also over the Seine (see Figure 17). In Great Britain, William Edwards built what many people consider the most beautiful arch bridge in the British Isles—the Pontypridd Bridge (1750), over the Taff in Wales, with a lofty span of 140 feet. In London the young Swiss engineer Charles Labelye, entrusted with the building of the first bridge at Westminster, evolved a novel and ingenious method of sinking the foundations, employing huge timber caissons that were filled with masonry after they had been floated into position for each pier. The 12 semicircular arches of portland stone, rising in a graceful camber over the river, set a high standard of engineering and architectural achievement for the next generation and stood for a hundred years.

The bridges of Jean-Rodolphe Perronet

Also in London, John Rennie, engaged by private enterprise in 1811, built the first Waterloo Bridge, whose level-topped masonry arches were described by the Italian sculptor Antonio Canova as "the noblest bridge in the world." It was replaced by a modern bridge in 1937–45. Rennie subsequently designed the New London Bridge of multiple masonry arches. Completed in 1831, after Rennie's death, it was subsequently widened and was finally replaced in the 1960s.

Timber truss bridges. In the 18th century, designs with timber, especially trusses, reached new span lengths. In 1755 a Swiss builder, Hans Grubenmann, used trusses to support a covered timber bridge with spans of 171 and 193 feet over the Rhine at Schaffhausen. Many timber truss bridges were built in the United States. One of the best long-span truss designs was developed by Theodore Burr, of Torrington, Conn., and based on a drawing by Palladio; a truss strengthened by an arch, it set a new pattern for covered bridges in the United States. Burr's McCall's Ferry Bridge (1815; on the Susquehanna River near Lancaster, Pa.) had a record-breaking span of 360 feet. Another successful design was the "lattice truss," patented by Ithiel Town in 1820, in which top and bottom chords were made of horizontal timbers connected by a network of diagonal planks.

Early trusses were built without precise knowledge of how the loads are carried by each part of the truss. The first engineer to analyze correctly the stresses in a truss was Squire Whipple, an American who designed hundreds of small truss bridges and published his theories in 1869. Understanding precisely how loads were carried led to a reduction in materials, which by then were shifting from wood and stone to iron and steel.

(H.S.-Sm./D.P.B./P.N.B.)

© Luc Boegly/Archipress

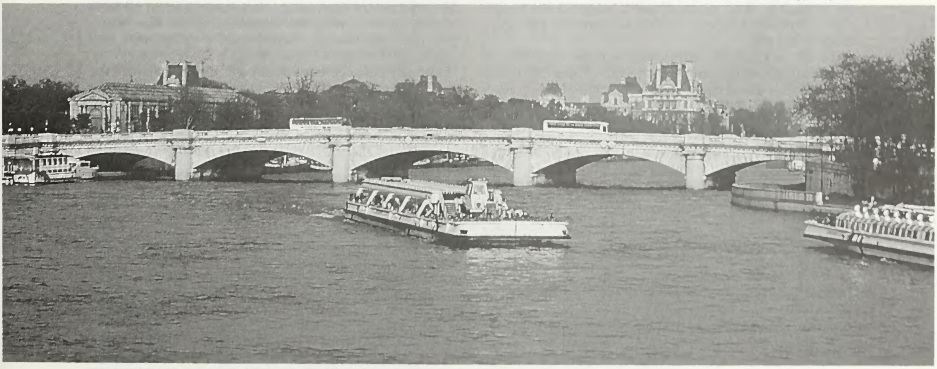


Figure 17: The Pont de la Concorde, a masonry arch bridge spanning the Seine River in Paris, Fr., completed in 1791 by Jean-Rodolphe Perronet.

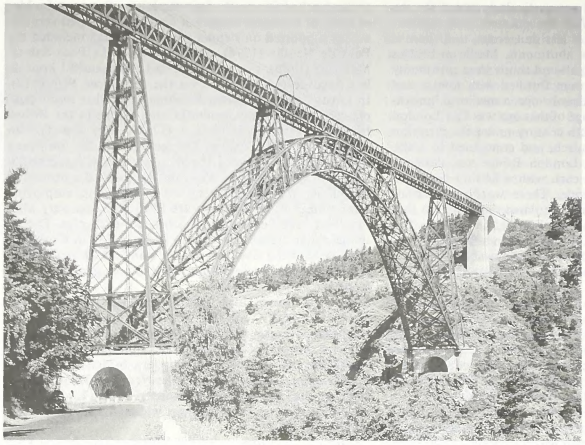


Figure 18: *The Garabit Viaduct, over the Truyère River near Saint-Flour, Fr.*
Spanning 541 feet, this wrought-iron railroad bridge was designed by Gustave Eiffel and completed in 1884.
© A.F. Kersting

IRON AND STEEL BRIDGES, 1779–1929

Iron. *Early designs.* During the Industrial Revolution the timber and masonry tradition was eclipsed by the use of iron, which was stronger than stone and usually less costly. The first bridge built solely of iron spanned the River Severn near Coalbrookdale, Eng. Designed by Thomas Pritchard and built in 1779 by Abraham Darby, the Coalbrookdale Bridge, constructed of cast-iron pieces, is a ribbed arch whose nearly semicircular 100-foot span imitates stone construction by exploiting the strength of cast iron in compression. In 1795 the Severn region was wracked by disastrous floods, and the Coalbrookdale Bridge, lacking the wide flat surfaces of stone structures, allowed the floodwaters to pass through it. It was the only bridge in the region to survive—a fact noted by the Scottish engineer Thomas Telford, who then began to create a series of iron bridges that were judged to be technically the best of their time. The 1814 Craigellachie Bridge, over the River Spey in Scotland, is the oldest surviving metal bridge of Telford's. Its 150-foot arch has a flat, nearly parabolic profile made up of two curved arches connected by X-bracing. The roadway has a slight vertical curve and is supported by thin diagonal members that carry loads to the arch.

The Coalbrookdale Bridge

The use of relatively economical wrought iron freed up the imaginations of designers, and one of the first results was Telford's use of chain suspension cables to carry loads by tension. His eyebar cables consisted of wrought-iron bars of 20 to 30 feet with holes at each end. Each eye matched the eye on another bar, and the two were linked by iron pins. The first of these major chain-suspension bridges and the finest of its day was Telford's bridge over the Menai Strait in northwestern Wales. At the time of its completion in 1826, its 580-foot span was the world's longest. In 1893 its timber deck was replaced with a steel deck, and in 1940 steel chains replaced the corroded wrought-iron ones. The bridge is still in service today.

Railway bridges. The rise of the locomotive as a mode of transportation during the 19th century spurred the design of new bridges and bridge forms strong enough to handle both the increased weight and the dynamic loads of trains. The most significant of these early railway bridges was Robert Stephenson's Britannia Bridge, also over the Menai Straits. Completed in 1850, Stephenson's design was the first to employ the hollow box girder. The hollow box gave the deck the extra stiffness of a truss, but it was easier to build and required less engineering precision—at the cost, however, of extra material. The wrought-iron boxes through which the trains ran were originally to be carried by chain suspension cables, but, during the building, extensive theoretical work and testing indicated that the cables were not needed; thus the towers stand strangely useless. For the Royal Albert Bridge (1859) over the River Tamar at Saltash, Eng., designer Isambard Kingdom Brunel used a combination of tubular arch and chain cable. The arches rise above the deck and, in conjunction with the chain suspenders, gives the bridge in profile what appear to be a set of eyes. The bridge at Saltash also carries trains, and its two main spans of 455 feet are comparable in length to the Britannia's 460-foot spans.

Among the most important railway bridges of the latter 19th century were those of Gustave Eiffel. Between 1867 and 1869 Eiffel constructed four viaducts of trussed-girder design along the rail line between Gannat and Commentry, west of Vichy in France. The most striking of these, at Rouzat, features wrought-iron towers that for the first time visibly reflect the need for lateral stiffness to counter the influence of horizontal wind loads. Lateral stiffness is achieved by curving the towers out at the base where they meet the masonry foundations, a design style that culminated in Eiffel's famous Parisian tower of 1889.

The bridges of Gustave Eiffel

Eiffel also designed two major arch bridges that were the longest-spanning structures of their type at the time. The first, the 1877 Pia Maria Bridge over the Duoro River near Oporto, Port., is a 522-foot, crescent-shaped span that rises 140 feet at its crown. Again, a wide spreading of the arches at their base gives this structure greater lateral stiffness. The crowning achievement of the crescent-arch form in the 19th century was represented by the completion in 1884 of Eiffel's 541-foot Garabit Viaduct over the Truyère River near Saint-Flour, Fr. (See Figure 18.) Unlike the bridge at Duoro, the Garabit arch is separated visually from the thin horizontal girder. Both arches were designed with hinges at their supports so that the crescent shape widens from points at the supports to a deep but light truss at the crown. The hinged design served to facilitate construction and also to produce the powerful visual image intended by Eiffel.

Suspension bridges. In the United States, engineer John Roebling established a factory in 1841 for making rope out of iron wire, which he initially sold to replace the hempen rope used for hoisting cars over the portage rail-

way in central Pennsylvania. Later Roebling used wire ropes as suspension cables for bridges, and he developed the technique for spinning the cables in place rather than making a prefabricated cable that needed to be lifted into place. In 1855 Roebling completed an 821-foot-span railway bridge over the Niagara River in western New York state. Wind loads were not yet understood in any theoretical sense, but Roebling recognized the practical need to prevent vertical oscillations. He therefore added numerous wire stays, which extended like a giant spiderweb in various directions from the deck to the valley below and to the towers above. The Niagara Bridge confounded nearly all the engineering judgment of the day, which held that suspension bridges could not sustain railway traffic. Although the trains were required to slow down to a speed of only three miles (five kilometres) per hour and repairs were frequent, the bridge was in service for 42 years, and it was replaced only because newer trains had become too heavy for it.

The Niagara Bridge

Roebling's Cincinnati Bridge (now called the John A. Roebling Bridge) over the Ohio River was a prototype for his masterful Brooklyn Bridge (see below *Steel: Suspension bridges*). When this 1,057-foot-span, iron-wire cable suspension bridge was completed in 1866, it was the longest spanning bridge in the world. Roebling's mature style showed itself in the structure's impressive stone towers and its thin suspended span, with stays radiating from the tower tops to control deck oscillations from wind loads.

Steel. *Railway bridges.* Between the American Civil War and World War I, railroads reached their peak in the United States and elsewhere, increasing the need for bridges that could withstand these heavier loads. New processes for making steel gave rise to many important bridges, such as the Eads Bridge over the Mississippi River at St. Louis, the Forth Bridge over the Firth of Forth in Scotland, the Hell Gate Bridge and Bayonne Bridge in New York City, and the Sydney Harbour Bridge in Australia.

The 1874 Eads Bridge was the first major bridge built entirely of steel, excluding the pier foundations. Designed by James Buchanan Eads, it has three arch spans, of which the two sides are each 502 feet and the middle is 520 feet. The Eads bridge was given added strength by its firm foundations, for which pneumatic caissons, instead of cofferdams, were used for the first time in the United States. Another innovation carried out by Eads, based on a proposal by Telford, was the construction of arches by the cantilevering method. The arches were held up by cables supported by temporary towers above the piers, all of which were removed when the arches became self-supporting.

The Forth Bridge over the Firth of Forth in Scotland, designed by Benjamin Baker, has two cantilevered spans of 1,710 feet, which made it the world's longest bridge upon its completion in 1890. (See Figure 19.) The steel structure rises 342 feet above the masonry piers. Although from an approaching standpoint it appears dense and massive, in profile it exhibits a surprising lightness. Baker designed the bridge with an artist's temperament. In his writings he criticized the Britannia Bridge for its towers, which Stephenson admitted had been left in place only in case the bridge needed suspension chains and not out of structural necessity. The Forth Bridge, on the other hand, is pure structure; nothing has been added for aesthetic appearance that does not have a structural function. For more than a century the bridge has carried a railway, and indeed it was one of the last great bridges built for that purpose in the 19th century.

The Forth Bridge



Figure 19: *The Forth Bridge, over the Firth of Forth in Scotland.* Designed by Benjamin Baker and completed in 1890, this steel cantilever bridge has two spans of 1,710 feet and rises 342 feet above its masonry piers.

© A.F. Kersting

The Hell Gate Bridge, completed by Gustav Lindenthal in 1916, also had an aesthetic intention. It was made to look massive by its stone towers and by the increased spacing of the two chords at the support, yet structurally the towers serve no purpose; the lower chord of the arch is actually hinged at the abutments, and all of the load is carried to the foundations by that lower chord. Nevertheless, the bridge has an imposing presence, and its arch of 978 feet was the world's longest at the time.

Similar in arch form to Hell Gate is the 1931 Bayonne Bridge, designed by Lindenthal's former associate, Othmar Ammann. Spanning the Kill van Kull between Staten Island, N.Y., and Bayonne, N.J., the Bayonne Bridge, though longer than the Hell Gate Bridge at 1,652 feet, is significantly lighter. The main span for the Hell Gate required 87 million pounds (39 million kilograms) of steel, compared with 37 million pounds for the Bayonne. Part of the reason is the lower live loads; for the Hell Gate, train loading was taken at 24,000 pounds per foot (36,000 kilograms per metre) of bridge length, whereas for the Bayonne the car loading was 7,000 pounds per foot (10,000 kilograms per metre). But the decrease is also due to an effort to make the arch more graceful as well as

© Lance Nelson/The Stock Market



Figure 20: The Sydney Harbour Bridge, a 1,650-foot steel arch linking Sydney and North Sydney, N.S.W., Australia.

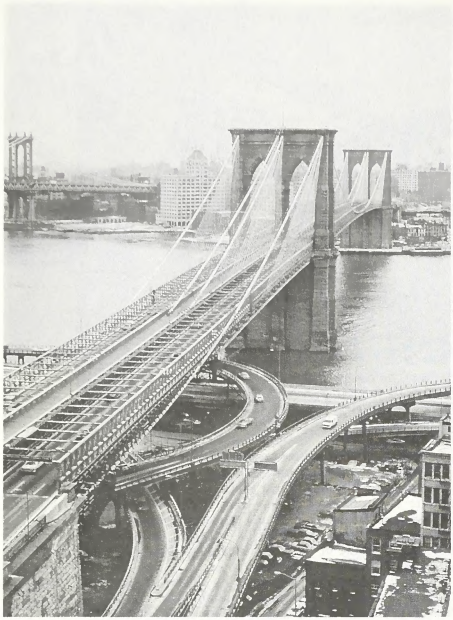


Figure 21: *The Brooklyn Bridge, spanning the East River, New York City.*
Designed by John Roebling and completed in 1883, this steel-cable suspension bridge has a span of 1,595 feet.
Envision/© Simon Feldman

more economical. Massive-looking stone-faced abutments were designed for the sake of appearance but then were never built, leaving a rather useless tangle of light steel latticework at the abutments. Nevertheless, from a distance the Bayonne Bridge shows a lightness and delicacy that bespeaks structural integrity.

Across the world in Sydney Harbour, N.S.W., Australia, Sir Ralph Freeman designed a steel arch bridge with a span of 1,650 feet that was begun in 1924 and completed in 1932 (see Figure 20). Because of the deep waters in the harbour, temporary supports were impractical, so the steel arch was assembled by cantilevering out from each bank and meeting in the middle. A high-strength silicon steel was used, making it the heaviest steelwork of its kind. The Sydney Harbour Bridge is a two-hinged arch, with its deck 172 feet above the water. It carries four railroad tracks, a roadway 57 feet wide, and two walkways. On each bank it is supported by a pair of large stone towers that, like those of the Hell Gate, disguise the fact that almost the entire load is carried by the lower arch chord.

The Port Authority of New York and New Jersey



Figure 22: *The George Washington Bridge, spanning the Hudson River between New Jersey and Manhattan Island, New York.*
Designed by Othmar Ammann and completed in 1931, the 3,500-foot-span suspension bridge is shown in its original configuration, before a second deck was added.

The Brooklyn Bridge

Suspension bridges. John Roebling died in 1869, shortly after work began on the Brooklyn Bridge, but the project was taken over and seen to completion by his son, Washington Roebling. Technically, the bridge overcame many obstacles through the use of huge pneumatic caissons, into which compressed air was pumped so that men could work in the dry; but, more important, it was the first suspension bridge on which steel wire was used for the cables. Every wire was galvanized to safeguard against rust, and the four cables, each nearly 16 inches (40 centimetres) in diameter, took 26 months to spin back and forth over the East River. After many political and technical difficulties and at least 27 fatal accidents, the 1,595-foot-span bridge was completed in 1883 to such fanfare that within 24 hours an estimated quarter-million people crossed over it, using a central elevated walkway that John Roebling had designed for the purpose of giving pedestrians a dramatic view of the city. (See Figure 21.)

By the turn of the 20th century, the increased need for passage from Manhattan to Brooklyn over the East River resulted in plans for two more long-span, wire-cable, steel suspension bridges, the Williamsburg and Manhattan bridges. The Williamsburg Bridge, designed by L.L. Buck with a span of just over 1,600 feet, became the longest cable-suspension span in the world upon completion in 1903. Its deck truss is a bulky lattice structure with a depth of 40 feet, and the towers are of steel rather than masonry. The truss in effect replaced Roebling's stays as stiffeners for the deck. The 1909 Manhattan Bridge has a span of 1,470 feet. Its fixed steel towers spread laterally at the base, and a 24.5-foot-deep truss is used for the deck. Of greater significance than the deck construction, however, was the first application of deflection theory, during the design of these two bridges, in calculating how the horizontal deck and curved cables worked together to carry loads. First published in 1888 by the Austrian academic Josef Melan, deflection theory explains how deck and cables deflect together under gravity loads, so that, as spans become longer and the suspended structure heavier, the required stiffness of the deck actually decreases. Deflection theory especially influenced design in the 1930s, as engineers attempted to reduce the ratio of girder depth to span length in order to achieve a lighter, more graceful, appearance without compromising safety. Up to 1930, no long-span suspension bridge had a ratio of girder depth to span length that was higher than 1:84.

Ralph Modjeski's Philadelphia-Camden Bridge (now called the Benjamin Franklin Bridge), over the Delaware River, is another wire-cable steel suspension bridge; when completed in 1926, it was the world's longest span at 1,750 feet. However, it was soon exceeded by the Ambassador Bridge (1929) in Detroit and the George Washington Bridge (1931) in New York. The Ambassador links the United States and Canada over the Detroit River. Because of heavy traffic on the river, a wide clearance was necessary. The steel suspension bridge designed by Jonathan Jones has a span of 1,850 feet and a total length, including approach spans, of more than 9,000 feet. The design of the Ambassador Bridge originally called for using heat-treated steel wires for the cables. Normally wires were cold-drawn, a method in which steel is drawn through successively smaller holes in dies, reducing its diameter yet raising its ultimate tensile strength. Extensive laboratory tests showed that heat-treated wires had a slightly higher ultimate strength, but during the construction of the Ambassador Bridge several of them broke, and, to the contractors' credit, all the cables spun thus far were

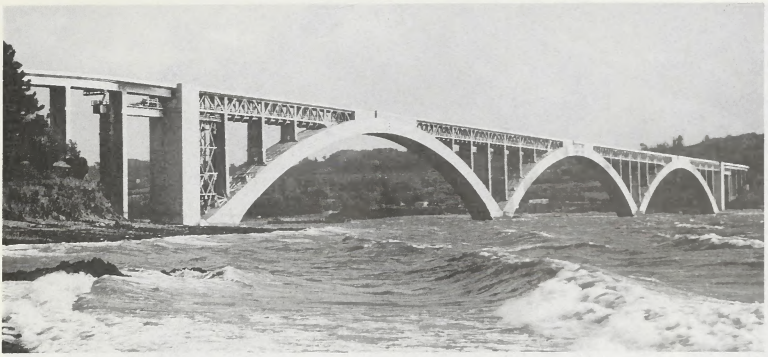


Figure 23: *The Plougastel Bridge, over the Elorn Estuary near Brest, Fr.*
Each hollow-box arch, made of reinforced concrete, has a span of 585 feet. The bridge was designed by Eugène Freyssinet and completed in 1930.
Freyssinet International

immediately replaced with cold-drawn wire. The example illustrates the limitations of laboratory testing as opposed to studies of actual working conditions.

The George Washington Bridge

The George Washington Bridge, a steel suspension bridge designed by Ammann, was significant first for its span length of 3,500 feet and second for its theoretical innovations. After studying deflection theory, Ammann concluded that no stiffness was needed in the deck at all, as it would be stabilized by the great weight of the bridge itself. Indeed, the George Washington Bridge is the heaviest single-span suspension bridge built to date, and its original ratio of girder depth to span was an astonishing 1:350. Originally the 635-foot-high towers were to have a masonry facade, but a shortage of money during the Great Depression precluded this, and the steel framework stands alone (see Figure 22). Ammann designed the bridge to carry a maximum of 8,000 pounds per foot, even though the maximum conceivable load on the bridge was estimated at 46,000 pounds per foot, thus illustrating the principle that longer bridges need not be designed for maximum load. In 1962 the addition of a second deck for traffic resulted in the construction of a deck truss, giving the bridge its current ratio of girder depth to span of 1:120.

CONCRETE BRIDGES

Reinforced concrete. *Early bridges.* During the 19th century, low-cost production of iron and steel, when added to the invention of portland cement in 1824, led to the development of reinforced concrete. In 1867 a French gardener, Joseph Monier, patented a method of strengthening thin concrete flowerpots by embedding iron wire mesh into the concrete. Monier later applied his ideas to patents for buildings and bridges. In 1879 another Frenchman, François Hennebique, set out to fireproof a metal-frame house in Belgium, and his decision to cover the iron beams with concrete led him to develop a structural system wherein the metal bars (replacing iron beams) carried tension and the concrete carried compression. By the end of the century reinforced concrete had become an economical substitute for stone, since it was generally cheaper to produce concrete than to quarry stones. In addition to its price and load-carrying advantages, reinforced concrete could be molded into a variety of shapes, allowing for much aesthetic expression on the part of the engineer without significantly increasing materials or cost.

The most prolific designers first using reinforced concrete were Hennebique and the German engineer G.A. Wayss, who bought the Monier patents. Hennebique's Vienne River Bridge at Châtellerault, Fr., built in 1899, was the longest-spanning reinforced arch bridge of the 19th century. Built low to the river—typical of many reinforced-concrete bridges whose goal of safe passage across a small river is not affected by heavy boat traffic—the Châtellerault bridge has three arches, the centre spanning just over 160 feet. In 1904 the Isar River Bridge at Grünewald, Ger., designed by Emil Morsch for Wayss's firm, became the longest reinforced-concrete span in the world at 230 feet.

Freyssinet's concrete arches

The longest-spanning concrete arches of the 1920s were designed by the French engineer Eugène Freyssinet. In his bridge over the Seine at Saint-Pierre-du-Vauvray (1922), two thin, hollow arches rise 82 feet at mid-span and are connected by nine crossbeams. The arches curve over the deck, which is suspended by thin steel wires lightly coated with mortar and hanging down in a triangular formation. The 435-foot span, then a record for reinforced concrete, thus has a light appearance. The bridge was destroyed during World War II but was rebuilt in 1946 using the same form.

In 1930 Freyssinet completed his most renowned work, the Plougastel Bridge over the Elorn Estuary near Brest, Fr. (See Figure 23.) This bridge featured three 585-foot hollow-box arch spans, then the longest concrete spans in the world. Because of the great scale of this structure, Freyssinet studied the creep, or movement under stress, of concrete. This led him to his general idea for prestressing (see below *Prestressed concrete*).

In 1943 the Plougastel was eclipsed in length by the Sandö Bridge over the Ångerman River in Sweden. The Sandö Bridge is a thin, single-ribbed, reinforced-concrete arch with a span of 866 feet, rising 131 feet above the river.

Maillart's innovations. Swiss engineer Robert Maillart's use of reinforced concrete, beginning in 1901, effected a revolution in structural art. Maillart, all of whose main bridges are in Switzerland, was the first 20th-century designer to break completely with the masonry tradition and put concrete into forms technically appropriate to its properties yet visually surprising. For his 1901 bridge over the Inn River at Zuoz, he designed a curved arch and a flat roadway connected by longitudinal walls that turned the complete structure into a hollow-box girder with a span of 125 feet and with hinges at the abutments and the crown. This was the first concrete hollow-box to be constructed. The arch at Zuoz is thickened at the bottom, and all of the load to the abutments is carried at these thick points. The walls near the abutments, therefore, are technically superfluous. For his 1905 bridge over the Vorder Rhine at Tavanasa, with a span of 167 feet, Maillart cut out the spandrel walls to achieve a technically superior form that was also visually new. As at Zuoz, the concrete arches of the Tavanasa bridge were connected by hinges to both abutments and to each other at the crown, thus allowing the arch to rise freely without internal stress when the temperature rose and to drop when the temperature went down. By contrast, Hennebique's bridge at Châtellerault

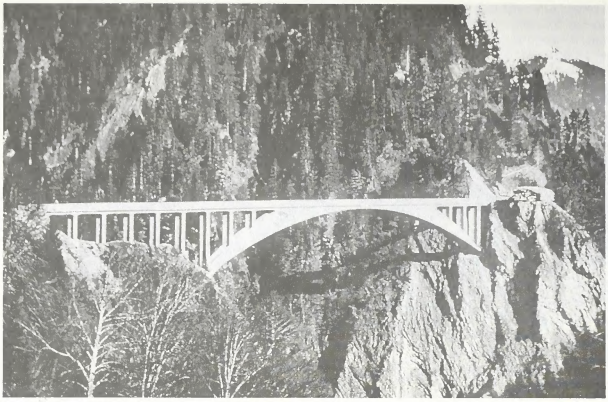


Figure 24: *The Salginatobel Bridge, near Schiers, Graubünden canton, Switz.*
Designed by Robert Maillart, the hollow-box, reinforced concrete arch, with a span of 295 feet, was built in 1930.
© Bildagentur Baumann

did not have hinges, and the arches cracked severely at the abutments and crown. The Tavanasa bridge was unfortunately destroyed by an avalanche in 1927.

The Salginatobel Bridge

Maillart's Valtschielbach Bridge of 1926, a deck-stiffened arch with a 142-foot span, demonstrated that the arch can be extremely thin as long as the deck beam is stiff. The arch at Valtschielbach increases in thickness from a mere 9 inches at the crown to just over 11 inches at the supports. Thin vertical slabs, or cross-walls, connect the arch to the deck, allowing the deck to stiffen the arch and thus permitting the arch to be thin. Such technical insight revealed Maillart's deep understanding of how to work with reinforced concrete—an understanding that culminated in a series of masterpieces beginning with the 1930 Salginatobel Bridge (see Figure 24), which, as with the others already mentioned, is located in the Swiss canton of Graubünden. The form of the Salginatobel Bridge is similar to the Tavanasa yet modified to account for a longer central span of 295 feet, which is needed to cross the deep ravine below. Maillart's hollow-box, three-hinged arch design not only was the least costly of the 19 designs proposed but also was considered by the district engineer to be the most elegant. The stone abutments of earlier Maillart bridges were dispensed with at Salginatobel, as the rocky walls of the ravine that meet the arch are sufficient to carry the load.

Other notable bridges by Maillart are the bridge over the Thur at Felsegg (1933), the Schwandbach Bridge near Hinterfultigen (1933), and the Töss River footbridge near Wulflingen (1934). The Felsegg bridge has a 226-foot span and features for the first time two parallel arches, both three-hinged. Like the Salginatobel Bridge, the Felsegg bridge features X-shaped abutment hinges of reinforced concrete (invented by Freyssinet), which were more economical than steel hinges. The Schwandbach Bridge, with a span of 123 feet, is a deck-stiffened arch with a horizontally curved roadway. The true character of reinforced concrete is most apparent in this bridge, as the inner edge of the slab-arch follows the horizontal curve of the highway, while the outer edge of the arch is straight. Vertical trapezoidal cross-walls integrate the deck with the arch, and the result is one of the most acclaimed bridges in concrete. The Töss footbridge is a deck-stiffened arch with a span of 125 feet. The deck is curved vertically at the crown and countercurved at the riverbanks, integrating the structure into the setting.

Maillart's great contribution to bridge design was that, while he kept within the traditional discipline of engineering, always striving to use less material and keep costs down, he continually played with the forms in order to achieve maximum aesthetic expression. Some of his last bridges—at Vessy, Liesberg, and Lachen—illustrate his mature vision for the possibilities of structural art. Over the Arve River at Vessy in 1935, Maillart designed a three-hinged, hollow-box arch in which the thin cross-walls taper at mid-height, forming an X shape. This striking design, giving life to the structure, is both a natural form and a playful expression. Also in 1935, a beam bridge over the Birs River at Liesberg employed haunching of the beams, a tapering outward at the base of the thin columns, and a curved top edge becoming less deep near the abutments. For a skewed railway overpass at Lachen in 1940, Maillart used two separate three-hinged arches that sprang from different levels of the abutment, creating a dynamic interplay of shapes.

First application of prestressed concrete

Prestressed concrete. *Eugène Freyssinet.* The idea of prestressing concrete was first applied by Freyssinet in his effort to save the Le Veurdre Bridge over the Allier River near Vichy, Fr. A year after its completion in 1910, Freyssinet noted the three-arch bridge had been moving downward at an alarming rate. A flat concrete arch, under its own dead load, generates huge compressive forces that cause the structure to shorten over time and, hence, move eventually downward. This "creep" may eventually cause the arch to collapse. Freyssinet's solution was to jack apart the arch halves at the crown, lifting the arch and putting the concrete into additional compression against the abutments and then casting new concrete into the spaces at the crown. By 1928, experience with the Le Veurdre Bridge led Freyssinet to propose the more common method of prestressing, using high-strength steel to put concrete into compression.

Freyssinet's major prestressed works came after the reinforced-concrete Plougastel Bridge and included a series of bridges over the Marne River following World War II. The Luzancy Bridge (1946), with a span of 180 feet, demonstrates the lightness and beauty that can be achieved using prestressed concrete for a single-span beam bridge.

The first major bridge made of prestressed concrete in the United States, the Walnut Lane Bridge (1950) in Philadelphia, was designed by Gustave Magnel and features three simply supported girder spans with a centre span of 160 feet and two end spans of 74 feet. Although it was plain in appearance, a local art jury responsible for final approval found that the slim lines of the bridge were elegant enough not to require a stone facade.

Ulrich Finsterwalder. During the years after World War II, a German engineer and builder, Ulrich Finsterwalder,



Figure 25: *The Ganter Bridge, Valais canton, Switz.*
The main span is 571 feet and the highest column 492 feet; the triangular concrete walls above the roadway contain prestressed cable-stays. The bridge was designed by Christian Menn and completed in 1980.
© Van Hoorick/PRISMA

developed the cantilever method of construction with prestressed concrete. Finsterwalder's Bendorf Bridge over the Rhine at Koblenz, Ger., was completed in 1962 with thin piers and a centre span of 673 feet. The double cantilevering method saved money through the absence of scaffolding in the water and also by allowing for reduced girder depth and consequent reduction of material where the ends of the deck meet in the centre. The resulting girder has the appearance of a very shallow arch, elegant in profile. Another fine bridge by Finsterwalder is the Mangfall Bridge (1959) south of Munich, a high bridge with a central span of 354 feet and two side spans of 295 feet. The Mangfall Bridge features the first latticed truss walls made of prestressed concrete, and it also has a two-tier deck allowing pedestrians to walk below the roadway and take in a spectacular view of the valley. Finsterwalder successfully sought to show that prestressed concrete could compete directly with steel not only in cost but also in reduction of depth.

Christian Menn. The technical and aesthetic possibilities of prestressed concrete were most fully realized in Switzerland with the bridges of Christian Menn. Menn's early arch bridges were influenced by Maillart, but, with prestressing, he was able to build longer-spanning bridges and use new forms. The Reichenau Bridge (1964) over the Rhine, a deck-stiffened arch with a span of 328 feet, shows Menn's characteristic use of a wide, prestressed concrete deck slab cantilevering laterally from both sides of a single box. For the high, curving Felsenau Viaduct (1974) over the Aare River in Bern, spans of up to 512 feet were built using the cantilever method from double piers. The trapezoidal box girder, only 36 feet wide at the top, haunches at the supports and carries an 85-foot-wide turnpike. More impressive yet is the high, curving Ganter Bridge (1980), crossing a deep valley in the canton of Valais (see Figure 25). The Ganter is both a cable-stayed and a prestressed cantilever girder bridge, with the highest column rising 492 feet and with a central span of 571 feet. The form is unique: the cable-stays are flat and covered by thin concrete slabs, making the bridge look very much like a Maillart bridge upside-down.

The Ganter Bridge

STEEL BRIDGES AFTER 1931

Long-span suspension bridges. The success of the George Washington Bridge—especially its extremely small ratio of girder depth to span—had a great influence on suspension bridge design in the 1930s. Its revolutionary design led to the building of several major bridges, such as the Golden Gate (1937), the Deer Isle (1939), and the Bronx-Whitestone (1939). The Golden Gate Bridge, built over the entrance to San Francisco Bay under the direction of Joseph Strauss, was upon its completion the world's longest span at 4,200 feet; its towers rise 746 feet above the water. Deer Isle Bridge in Maine, U.S., was designed by David Steinman with only plate girders to stiffen the deck, which was 25 feet wide yet had a central span of 1,080 feet. Likewise, the deck for Othmar Ammann's Bronx-Whitestone Bridge in New York was originally stiffened only by plate girders; its span reached 2,300 feet. Both the Deer Isle and the Bronx-Whitestone bridges later oscillated in wind and had to be modified following the Tacoma Narrows disaster.

Tacoma Narrows. In 1940 the first Tacoma Narrows Bridge opened over Puget Sound in Washington state, U.S. Spanning 2,800 feet, its deck, also stiffened by plate girders, had a depth of only 8 feet. This gave it a ratio of girder depth to span of 1:350, identical to that of the George Washington Bridge. Unfortunately, at Tacoma Narrows, just four months after the bridge's completion, the deck tore apart and collapsed under a moderate wind. At that

Copyright Smithsonian Institution

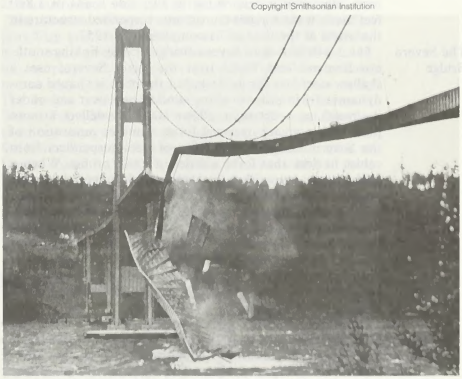


Figure 26: Collapse of the Tacoma Narrows Bridge, Washington state, in 1940.



Figure 27: The Konohana Bridge in Ōsaka, Japan, a single-cable suspension bridge with a span of 984 feet.
© Bon Color Photo Agency Co., Ltd.

time bridges normally were designed to withstand gales of 120 miles per hour, yet the wind at Tacoma was only 42 miles per hour. Motion pictures taken of the disaster show the deck rolling up and down and twisting wildly. These two motions, vertical and torsional, occurred because the deck had been provided with little vertical and almost no torsional stiffness. Engineers had overlooked the wind-induced failures of bridges in the 19th century and had designed extremely thin decks without fully understanding their aerodynamic behaviour. After the Tacoma bridge failed, however, engineers added trusses to the Bronx-Whitestone bridge, cable-stays to Deer Isle, and further bracing to the stiffening truss at Golden Gate. In turn, the diagonal stays used to strengthen the Deer Isle Bridge led engineer Norman Sollenberger to design the San Marcos Bridge (1951) in El Salvador with inclined suspenders, thus forming a cable truss between cables and deck—the first of its kind.

Lessons of the disaster. The disaster at Tacoma caused engineers to rethink their concepts of the vertical motion of suspension bridge decks under horizontal wind loads. Part of the problem at Tacoma was the construction of a plate girder with solid steel plates, eight feet deep on each side, through which the wind could not pass. For this reason, the new Tacoma Narrows Bridge (1950), as well as Ammann's 4,260-foot-span Verrazano Narrows Bridge in New York (1964), were built with open trusses for the deck in order to allow wind passage. The 3,800-foot-span Mackinac Bridge in Michigan, U.S., designed by Steinman, also used a deep truss; its two side spans of 1,800 feet made it the longest continuous suspended structure in the world at the time of its completion in 1957.

The Severn Bridge

The 3,240-foot-span Severn Bridge (1966), linking southern England and Wales over the River Severn, uses a shallow steel box for its deck, but the deck is shaped aerodynamically in order to allow wind to pass over and under it—much as a cutwater allows water to deflect around piers with a greatly reduced force. Another innovation of the Severn Bridge was the use of steel suspenders from cables to deck that form a series of Vs in profile. When a bridge starts to oscillate in heavy wind, it tends to move longitudinally as well as up and down, and the inclined suspenders of the Severn Bridge act to dampen the longitudinal movement. The design ideas used on the Severn Bridge were repeated on the Bosporus Bridge (1973) at Istanbul and on the Humber Bridge (1981) over the River Humber in England. The Humber Bridge in its turn became the longest-spanning bridge in the world, with a main span of 4,626 feet.

Truss bridges. Although trusses are used mostly as secondary elements in arch, suspension, or cantilever designs, several important simply supported truss bridges have achieved significant length. The Astoria Bridge (1966) over the Columbia River in Oregon, U.S., is a continuous three-span steel truss with a centre span of 1,232 feet, and the Tenmon Bridge (1966) at Kumamoto, Japan, has a centre span of 984 feet.

In 1976 the New River Gorge Bridge, the world's longest-spanning steel arch, was completed in Fayette county, West Virginia, U.S. Designed by Michael Baker, the two-hinged arch truss carries four lanes of traffic 876 feet above the river and has a span of 1,700 feet.

Japanese long-span bridges. In the 1970s the Japanese, working primarily with steel, began to build a series of long-span bridges using several forms that by the year 2000 would include nine of the world's 20 longest spans.

Ōsaka Harbour. In 1974 the Minato Bridge, linking the city of Ōsaka with neighbouring Amagasaki, became one of the world's longest-spanning cantilever truss bridges, at 1,673 feet. In 1989 two other impressive and innovative bridges were completed for the purpose of carrying major highways over the port facilities of Ōsaka Harbour. The Konohana suspension bridge carries a four-lane highway on a slender, steel box-beam deck only 10 feet deep. The bridge is self-anchored—that is, the deck has been put into horizontal compression, like that on a cable-stayed bridge, so that there is no force of horizontal tension pulling from the ground at the anchorages. Spanning 984 feet, it is the first major suspension bridge to use a single cable. The towers are delta-shaped, with diagonal suspenders running from the cable down the centre of the deck. On the same road as the Konohana is the Ajigawa cable-stayed bridge, with a span of 1,148 feet and an elegantly thin deck just over 10 feet deep.

Island bridges. The Kanmon Bridge (1975), linking Honshu and Kyushu over the Shimonoseki Strait, was the first major island bridge in Japan. At about this time the Honshu-Shikoku bridge project was undertaken in order to connect these two main islands. Part of this project, the Seto Great Bridge, opened in 1988.

The Seto Great Bridge

The Seto Great Bridge is actually a series of bridges and viaducts that connect five small islands between the city of Kojima, on the main island of Honshu, and Sakaide, on the island of Shikoku. It consists of a 3,084-foot suspension bridge, a curving viaduct, two identical 1,378-foot cable-stayed bridges, an S-plan structure including a continuous truss viaduct with a main span of 804 feet, and finally two nearly identical suspension bridges with spans of 3,248 feet and 3,609 feet separated by a common water anchorage. The double deck is a strong 43-foot-deep continuous truss that carries cars and trucks on the top deck and trains on the lower deck. The Seto Great Bridge thus includes the first suspension bridges designed to carry railroad traffic since John Roebling's Niagara Bridge. The primary achievement, however, is its scale, with an overall length of six miles.

The Seto Great Bridge forms the middle of three routes intended to make up the Honshu-Shikoku project. The eastern route, between Kōbe (Honshu) and Naruto (Shikoku), includes plans for the world's longest suspension bridge,

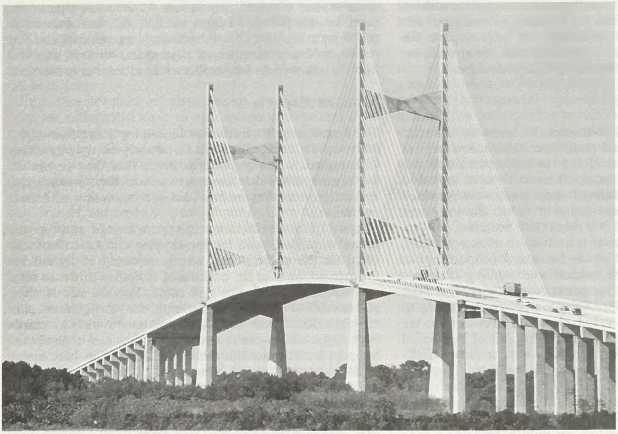


Figure 28: *The Dames Point Bridge, spanning the St. Johns River at Jacksonville, Fla.* Designed by Howard Needles and completed in 1987, the cable-stayed bridge has a span of 1,300 feet.
The Florida Times-Union

crossing the Akashi Strait with a main span of 6,530 feet and side spans of 3,150 feet. The western route, between Onomichi (Honshu) and Imbari (Shikoku), includes the Ikuchi Bridge, which upon its completion in 1991 became the world's longest-spanning cable-stayed bridge, with a main span of 1,608 feet. The two towers of the Ikuchi Bridge are delta-shaped, with two inclined planes of fan-arranged stays. Also on the Onomichi-Imbari route are plans for an even larger cable-stayed structure, the Tartara Bridge, designed with a main span of 2,920 feet and side spans of 886 feet and 1,050 feet.

CABLE-STAYED BRIDGES

German designs. Beginning in the 1950s, with the growing acceptance of cable-stayed bridges, there came into being a type of structure that could not easily be classified by construction material. Cable-stayed bridges offered a variety of possibilities to the designer regarding not only the materials for deck and cables but also the geometric arrangement of the cables. Early examples, such as the Strömsund Bridge in Sweden (1956), used just two cables fastened at nearly the same point high on the tower and fanning out to support the deck at widely separated points.

The Oberkasseler Bridge

By contrast, the Oberkasseler Bridge, built over the Rhine River in Düsseldorf, Ger., in 1973, used a single tower in the middle of its twin 846-foot spans; the four cables were placed in a harp or parallel arrangement, being equally spaced both up the tower and along the centre line of the deck. The Bonn-Nord Bridge in Bonn, Ger. (1966), was the first major cable-stayed bridge to use a large number of thinner cables instead of relatively few but heavier ones—the technical advantage being that, with more cables, a thinner deck might be used. Such multicable arrangements subsequently became quite common. The box girder deck of the Bonn-Nord, as with most cable-stayed bridges built during the 1950s and '60s, was made of steel. From the 1970s, however, concrete decks were used more frequently.

U.S. designs. Cable-stayed bridges in the United States reflected trends in both cable arrangement and deck material. The Pasco-Kennewick Bridge (1978) over the Columbia River in Washington state supported its centre span of 981 feet from two double concrete towers, the cables fanning down to the concrete deck on either side of the roadway. Designed by Arvid Grant in collaboration with the German firm of Leonhardt and Andra, its cost was not significantly different from those of other proposals with more conventional designs. The same designers produced the East Huntington Bridge across the Ohio River between Ohio and West Virginia in 1985. The East Huntington has a major span of 900 feet and a minor span of 608 feet. The single concrete tower is shaped like a long triangle in the traverse direction, and the cable arrangement is of the fan type; but, while the Pasco-Kennewick Bridge has two parallel sets of cables, at East Huntington there is but one set, fanning out from a single plane at the tower into two planes at the composite steel and concrete deck, so that, as one moves from pure profile to a longitudinal view, the cables do not align visually.

The Sunshine Skyway Bridge (1987), designed by Eugene Figg and Jean Mueller over Tampa Bay in Florida, has a main prestressed-concrete span of 1,200 feet. It, too, employs a single plane of cables, but these remain in one plane that fans out down the centre of the deck. The longest cable-stayed bridge in the United States is Dames Point Bridge (1987), designed by Howard Needles in consultation with Ulrich Finsterwalder and crossing the St. Johns River in Jacksonville, Fla. The main span at Dames Point is 1,300 feet, with side spans of 660 feet. From H-shaped towers of reinforced concrete, two planes of stays in harp formation support reinforced-concrete girders. The towers are carefully shaped to avoid a stiff appearance. (D.P.B./P.N.B.)

AIRPORTS

Aerodromes, airports, and airfields

According to the International Civil Aviation Organization, an aerodrome is a defined area on land or water (including buildings, installations, and equipment) that is intended to be used either wholly or in part for the arrival, departure, and surface movement of aircraft. Aerodrome is now largely a technical term and is no longer in common use. More usually the terms airport and airfield are employed, although there is some lack of precision in their meanings. Airport is used particularly to denote areas where air-transport passengers are carried, especially where movement occurs on a considerable scale and always where full-time customs facilities are available. Air-

field commonly refers to small aerodromes (often without paved runways) or facilities for handling air-transport aircraft or their passengers and cargo. At large airports the term often designates the operational areas for aircraft, including the runways, taxiways, aprons, and strips.

Evolution of airports

The requirements for aerodromes, or airports, have increased in complexity and scale since the earliest days of flying. Before World War II the landing and takeoff distance of most passenger-transport aircraft was at most 650 yards (600 metres). Additional clear areas were provided for blind landings or bad-weather runs but the total area involved rarely exceeded 500 acres (200 hectares).

It was not until the general introduction of heavy monoplanes for transport, such as the Douglas DC-3, during the late 1930s that extensive takeoff and landing distances were needed. Even then, the prewar airfields at New York City (La Guardia), London (Croydon), Paris (Le Bourget), and Berlin (Tempelhof) were laid out on sites close to the city centres. Because even transport aircraft of the period were relatively light, paved runways were a rarity. Croydon, Tempelhof, and Le Bourget, for example, all operated from grass strips only. Early airports were also major centres of leisure activity, often attracting more visitors than passengers. In 1939 La Guardia Airport attracted almost a quarter-million visitors per month, reaching a peak of 7,000 in one day, compared with a maximum daily throughput of only 3,000 passengers. In 1929 Berlin's airport reported 750,000 visitors and boasted a restaurant that could seat 3,000 people on the roof of the passenger terminal. The status of prewar airports as major social centres was reflected in their design, especially where the requirements of catering, observation decks, and parking were paramount. Indeed, the requirements of aircraft and passengers were not at all dominant at early airfields.

Much long-distance air transport was handled by the large seaplanes known as flying boats or clippers. These aircraft, though slow and of limited range, offered a level of comfort that was necessary for long-distance travel. Air terminal facilities were necessarily constructed close to large, open stretches of water. La Guardia Airport and Santos Dumont Airport in Rio de Janeiro are examples of airports that still operate on sites originally chosen for their ability to handle large seaplanes. The large facilities at Southampton Water in the United Kingdom have now disappeared, but the artificial lake at Linate Airport near Milan, Italy, is still to be found close to the present administration facilities.

Facilities for seaplanes

The vast majority of airfields throughout the world are still relatively simple facilities. Even now, many have unpaved runways or at most lightly paved runways with tiny terminal or administration buildings, a rudimentary control tower, and crude landing aids. Such facilities can deal only with light aircraft and a negligible flow of passengers or freight. Heavy air traffic is now almost entirely handled by sophisticated airport facilities that can accommodate the needs of crew, passengers, and freight and the great range of aircraft types that have evolved to meet the needs of modern air transport and general aviation.

Approximately 87 airports around the world now handle more than 10 million passengers each year; half of these are in the United States. Some 20 airports regularly move more than 30 million passengers on a yearly basis—Chicago's O'Hare International Airport and the Hartsfield Atlanta International Airport in Georgia each handle more than 70 million. The Memphis (Tenn.) International Airport, the home airport of FedEx cargo service, is the world's largest cargo shipper, with 5.25 million tons handled in 1999. In order to meet the increasing demand for air travel, large transport aircraft powered by multiple jet and turboprop engines have been built. Such aircraft require extensive ground facilities, runways, taxiways, firefighting and rescue services, passenger- and cargo-handling facilities, access to car parking and public transport, lighting, navigational and approach aids, and various support facilities such as catering, meteorology, and governmental inspection. In order to be attractively convenient, the modern airport must be located sufficiently close to the main centres of world population. At the same time, they must be adequately distant, so that the environmental problems associated with the noise of large aircraft and the activities of large numbers of passengers, workers, and visitors do not become intolerable to the cities that are served.

Modern airports

The largest airports in the world employ more than 100,000 workers each. They are immensely complex entities with regard to the physical facilities that they comprise, the organizations that are active within their boundaries, and the services that are provided in conjunction with their operation.

Physical facilities include runways, taxiways, aprons, and strips, which are used for the landing and takeoff of aircraft, for the maneuvering and positioning of aircraft on the ground, and for the parking of aircraft in order to load and discharge passengers and cargo. For the safe landing and takeoff of aircraft, lighting and radio navigational aids are provided. These are supplemented by airfield markings, signs and signals, and air traffic control facilities. Support facilities on the airside of the field include meteorology, fire and rescue, power and other utilities, aircraft maintenance, and airport maintenance. Landside facilities are the passenger and cargo terminals and the access system, which includes parking, roads, public transport facilities, and loading and unloading areas.

Airside and landside facilities

Many organizations are involved in the operation of a modern airport. Overall management is usually in the control of an organization, authority, or company that holds a license to operate the facility. This license is granted subject to a judgment by the national civil aviation authorities that the managing body is fit and competent to run an airport within national and, if applicable, international laws governing safety and operations. While overall responsibility for efficient, safe, and legal operation lies with the airport management, many of the individual services at an airport are provided by other organizations. Such organizations include airlines; air traffic control authorities; ground handling companies; fixed-base operators; concessionaires; security organizations; governmental agencies responsible for customs, immigration, health control, and police; support companies providing flight catering, fueling, aircraft engineering, and maintenance; aero clubs; and flying schools. Since the early 1980s, when privatization began to sweep through civil aviation, terminal-operation companies have also become more frequent, such as those that own terminals in Birmingham, Eng.; Brussels; and Toronto.

Airport services related to the aircraft are frequently referred to as airside. Many of these services are concentrated on the apron, or ramp, which is that part of the operational surface adjacent to the terminals where aircraft are maneuvered or parked. They include the apron handling of aircraft, airside passenger transfer to the aircraft, the handling of baggage and cargo, aircraft fueling, catering and cabin cleaning, engine starting, deicing, ground power and air-conditioning, and minor maintenance engineering. Other airside services are runway inspection, lighting and navigational aids, fire fighting and rescue, airside maintenance, and air traffic control. Among the landside services are those related to ground passenger handling; these include check-in, security, customs and immigration, baggage delivery, information, catering, cleaning and maintenance, shops and concessionary facilities, automobile rental, ground transportation, porters, special help for the elderly and handicapped, automobile parking, and public transportation (including taxis). In addition, because airports employ such a large number of workers, extensive provision must be made for their daily requirements.

SITE SELECTION

Aeronautical and environmental factors. Selecting a site for a new airport, or evaluating how well an existing site can be expanded to provide a new major airport, is a complex process. A balance must be achieved between aeronautical and air-transport requirements and the impact of the airport on its environment.

From an aeronautical viewpoint, the basic requirement of an airport is that it have a relatively flat area of land sufficiently large to accommodate the runways and other facilities and that this area be in a locality free from such obstructions to air navigation as mountains and tall buildings.

From the viewpoint of air-transport needs, airport sites must be sufficiently close to population centres that they are considered reasonably accessible to their users. Environmental considerations, on the other hand, dictate that the site should be far enough away from urban centres that noise and other deleterious effects on the population should be kept to acceptable levels. Furthermore, the airport should not destroy areas of natural beauty or other significance. These two sets of requirements, the aeronautical and the environmental, almost inevitably clash, the conflict becoming more severe as the scale of the envisaged airport increases.

Optimal sizes of airports

The most modest airport facility—with a single runway, an apron, and a building that serves simultaneously as terminal, administration area, and control tower—can comfortably be built on a site as small as 75 acres, since it requires only a flat, well-drained area sufficient to accommodate a short runway and its surrounding safety strip. Larger and more modern airport facilities, on the other hand, require multiple runways of extended length, extensive terminal apron areas, and large expanses of land devoted to parking and landside access roads. For such an airport, a minimum area of 3,000 acres is likely to be required. Several major airports—such as Dallas–Fort Worth International Airport in Texas, King Abdul Aziz International Airport near Jiddah, Saudi Arabia, and Charles de Gaulle Airport near Paris—are built on sites well in excess of this figure.

The selection process. The site-selection process for large airports can take many months; in some notable cases it has extended over many years. The procedure is complicated by the number of factors that must be taken into account. First, the operational capability of the site is assessed, particularly with respect to weather conditions such as wind, snow, ice, fog, and low visibility and also with respect to obstructions to air navigation around the airport, particularly on the approach and takeoff paths. The location of the facility in relation to air-traffic-controlled airspace is also operationally important. In addition, there must be an evaluation of the capacity of the available land to accommodate the expected configuration of runways and other facilities. Flat or very gently undulating land is necessary, because runways must be constructed according to restrictions on maximum allowable slopes—which, in turn, are governed by aircraft performance on landing and takeoff.

Ground access to the airport is also considered. An evaluation is made of the distance from population centres, the regional highway infrastructure, public transport facilities (including railways), and the availability of land for parking. Development costs are also estimated, taking into account the nature of the terrain, soil and rock conditions, drainage requirements, and local land values.

The environmental consequences of an airport development rank very high in any site-selection procedure. The impact of aircraft noise on the neighbouring population is often the most significant environmental factor, but in many countries account must also be taken of the impact on the flora and fauna of the area, pollution through chemical runoff into local groundwater, the presence of endangered species or significant cultural sites, and even undesirable changes in land use. Many governments now require that environmental analyses of airport development projects include evaluations of population reloca-

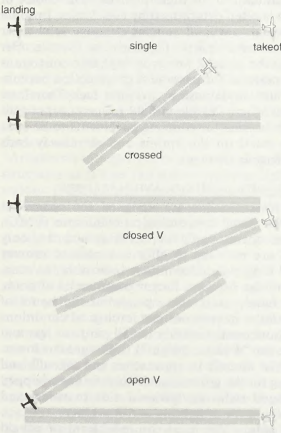


Figure 29: Four single runway configurations.

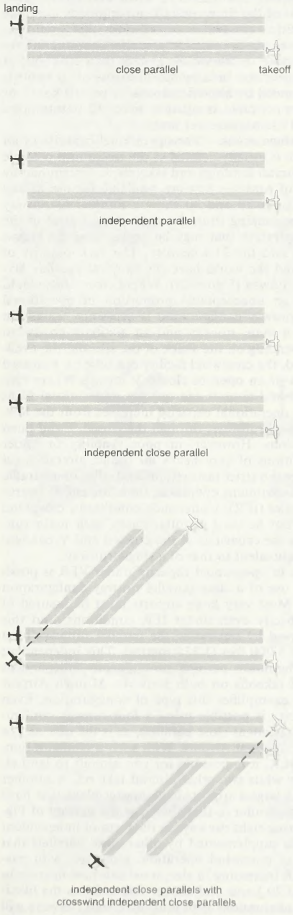


Figure 30: Four parallel runway configurations.

tion, changes in employment patterns, and distortion of existing regional land use and transportation planning.

AIRFIELD LAYOUT AND CONFIGURATION

Operational requirements. It is obvious even to the most casual observer that there is a large variation in the appearance and layout of airport facilities. Simple airports designed to accommodate light aircraft are essentially similar, but, as airports become larger and more complex, accommodating more passengers and cargo, their individual requirements affect their layouts and ensure that each becomes recognizably different.

The principal determinants of airport layout are the number of runways and their orientation, the shape of the available site, and constraints at the site both on the ground and in the air. The location and orientation of runways is governed in turn by the need to avoid obstacles, particularly during landing and takeoff procedures. For the largest airports, obstacles to air navigation must be considered up to 10 miles (15 kilometres) from the runways. Runway configurations must also ensure that, for 95 percent of the time, aircraft can approach and take off without either crosswinds or tailwinds that would inhibit operations. At the smallest airports, light aircraft are unable to operate in crosswinds greater than 10 knots; at all airports, operation in tailwinds in excess of 10 knots is not recommended by aircraft manufacturers (10 knots, or nautical miles per hour, is equal to about 12 statute miles per hour or 19 kilometres per hour).

Runway configurations. The operational capacity of an airport, which is usually defined as the maximum possible number of aircraft landings and takeoffs, is determined by the number of runways that are available for use at any one time. Figures 29 and 30 show a number of runway configurations ranging from the most basic layout to the sort of configuration that may be expected at the largest airports well into the 21st century. The vast majority of airports around the world have the simplest possible layout, a single runway (Figure 29). Where crosswinds would be high for an unacceptable proportion of operational time, a two-runway configuration is necessary, usually in the form of a main runway and an auxiliary crosswind runway. Depending on the shape of the site and the availability of land, the crosswind facility can take on a crossed configuration or an open or closed V layout. Where visibility is good and aircraft can operate under visual flight rules (VFR), operational capacity increases from the lowest level, crossed runways, through the closed V and open V configurations. However, in poor visibility or under certain conditions of very heavy air traffic, aircraft must operate under the strict instructions and rules of air traffic control and instrument operation; these are called instrument flight rules (IFR). Under such conditions, crosswind runways cannot be used simultaneously with main runways, so that the capacities of the crossed and V configurations are equivalent to that of a single runway.

The single runway

An increase in operational capacity under VFR is possible with the use of a close parallel runway configuration (Figure 30). Most very large airports must be assured of adequate capacity even under IFR conditions, and this can be achieved by separating the parallel runways by a minimum of 5,000 feet (1,525 metres). This independent parallel configuration permits simultaneous independent landings and takeoffs on both runways. Munich Airport in Germany exemplifies this type of configuration. Even greater capacity is possible using a four-runway configuration of independent close parallels, as is the case at Los Angeles International Airport. With such a configuration, even under IFR, it is possible for two aircraft to land simultaneously while two other aircraft take off. A number of the world's largest airports have master plans that have configurations similar to that shown at the bottom of Figure 30, featuring eight runways in the form of independent close parallels supplemented by other close parallels that are capable of crosswind operation. However, with passenger aircraft increasing in size, most can now operate in crosswinds of 20 knots and above. This reduces the likelihood that configurations with four crosswind runways will ever be constructed.

RUNWAY PAVEMENTS

Until the introduction of heavy monoplane aircraft in the latter part of the 1930s, civil air-transport aircraft were able to operate from grass runways with takeoff distances of less than 650 yards. The advent of heavy aircraft such as the DC-3 required the provision of paved runways; at the same time, takeoff distances increased to more than 1,000 yards. The length requirements for runways continued to increase into the mid-1970s, when large civilian aircraft such as the Douglas DC-8 and some models of the Boeing 747 required almost 4,000 yards of runway at sea level. (Even longer runways were necessary at higher elevations or where high ambient air temperatures occurred during operations.) The trend toward increasing runway lengths caused many problems at existing civilian airports, where runways had to be extended in order to accommodate the new aircraft. Ultimately, pressure by airport operators and the development of turbofan jet engines arrested and finally reversed the trend. Since the 1970s, runway length requirements have actually decreased, and the takeoff and climb performance of civilian aircraft has improved substantially. This has brought a dual benefit in reducing the area of land required by an airport and also in reducing the area around the airport that is adversely affected by noise on takeoff.

Increasing length of runways

At all but the smallest airports, pavements are now provided for runways, taxiways, aprons, and any other areas where aircraft are maneuvered. Pavements must be designed in such a way that they can bear the loads imposed by aircraft without failure. A pavement must be smooth and stable under conditions of loading during its expected or economic life. It should be free from dust and other particles that could be blown up and ingested into engines, and it must be capable of spreading and transmitting an aircraft's weight to the existing subsoil (or subgrade) in a manner that precludes subsoil failure. Another function of the pavement is to prevent weakening of the subsoil by moisture intrusion, especially from rainfall and frost.

Airfield pavements are of two types, rigid and flexible. Rigid pavements are constructed of portland cement concrete slabs resting on a prepared subbase of granular material or directly on a granular subgrade. Load is transmitted through the slabs to the underlying subgrade by flexure of the slabs. Flexible pavements are constructed of several thicknesses of asphalt or bituminous concrete layers overlying a base of granular material on a prepared subgrade. They spread the concentrated aircraft wheel loads throughout their depth until the load at the base of the pavement is less than the strength of the in situ soil. At all depths the strength of the pavement should be at least equal to the loads placed upon it by aircraft wheels. The choice of pavement type is often determined by economics. In some parts of the world, portland cement concrete is cheaper than asphalt; in other parts, the converse is true. For certain parts of the airfield, however, asphaltic concrete is an unsuitable material for pavement construction because of its vulnerability to damage by aviation fuel. Therefore, even at airports where flexible airfield pavements are generally in use, it is usual for concrete pavements to be used where aircraft stand on the aprons and at runway ends where fuel spillage is frequent.

NAVIGATIONAL AIDS, LIGHTING, AND MARKING

Only the simplest airfields are designed for operations conducted under visual meteorological conditions (VMC). These facilities operate only in daylight, and the only guidance they are required to offer is a painted runway centreline and large painted numbers indicating the magnetic bearing of the runway. Larger commercial airports, on the other hand, must also operate in the hours of darkness and under instrument meteorological conditions (IMC), when horizontal visibility is 650 yards or less and the cloud base (or "decision height") is 65 yards or lower. In order to assist aircraft in approaches and takeoffs and in maneuvering on the ground, such airports are equipped with sophisticated radio navigational aids (navaids) and visual aids in the form of lighting and marking.

Navigational aids. The most common form of navaid used for the approach phase of aircraft descent is the

The instrument landing system

instrument landing system (ILS). This is a radio signal that is beamed along the centreline of the runway and at the correct angle of approach (usually 3° above the horizontal). The beam is intercepted by an approaching aircraft up to 15 miles from the threshold of the runway. Information is given concerning position above and below the glide slope and deviation to the right or left of centreline; consequently, the pilot is able to determine from cockpit instruments a deviation of the aircraft from the proper approach.

Additional approach information is given visually to the pilot in the form of lighting approach aids. Two systems of approach aids are in use: the visual approach slope indicator system (VASIS) and the more modern precision approach path indicator (PAPI). Both work on the principle of guiding lights that show white when the pilot is above the proper glide slope and red when below.

Airfield lighting. Visual guidance to approaching aircraft is also provided by approach lighting systems, a configuration of high-intensity white lights running along the centreline of the runway and extending up to 650 yards beyond the threshold. At airfields where aircraft operate in very poor visibility, touchdown-zone lighting is provided over the first 1,000 yards from the runway threshold. These lights, set in patterns flush with the runway pavement, provide guidance up to the final moment of touchdown.

The runway itself is strongly delineated by a variety of guidance light systems. The threshold is designated by a line of green lights, and the edges and centreline are delineated by white lights that shine toward the maneuvering aircraft at regular intervals. The pilot is warned of the approaching runway end by a line of red lights at the end of the usable pavement. Taxiways are delineated by blue edge lights and by green centreline lights that also appear at regular intervals.

Runway markings. Considerable additional visual guidance is given to pilots by painted markings on the runway. The form of marking indicates at a glance whether radio instrument guidance is available at any particular airfield. On precision instrument runways, the runway edges are indicated by painted lines, and distances along the runway from the threshold are indicated by pavement markings. In addition, touchdown-zone markings are painted on the pavement immediately after the threshold, providing vital visual guidance during the moments immediately before touchdown when all lighting may be obscured by fog.

AIR TRAFFIC CONTROL

In the vicinity of airports—especially large airports, where in peak conditions as many as three landing or takeoff operations may occur every minute—the control of aircraft in the air is a difficult but extremely important operation. Aircraft require very large amounts of airspace, but at the same time the risk of collision must be set at very low, almost negligible, levels. Because aircraft are concentrated in the airspace around airports, acceptable levels of collision risk can be achieved only by strict adherence to procedures that are set out and monitored by air traffic control authorities.

Flight information regions

An aircraft in flight follows en route air traffic control instructions as it flies through successive flight information regions (FIRs). Upon approaching an airport at which a landing is to be made, the aircraft passes into the terminal control area (TCA). Within this area, there may be a greatly increased density of air traffic, and this is closely monitored on radar by TCA controllers, who continually instruct pilots on how to navigate within the area. The aircraft is then brought into the final approach pattern, at which point control passes to the approach controller, who monitors the aircraft to the runway itself. Once on the runway, the pilot is given instructions on ground maneuvers by the ground controller, whose responsibility is to avoid conflicting movements of aircraft in the operational area of the airfield. The ground controller gives the pilot instructions on reaching the apron stand position via the appropriate turnoffs and taxiways. Final positioning may be the responsibility of an apron controller. Departing aircraft go through a reverse procedure, whereby control is passed from ground control to departure control to terminal control area and, finally, to en route control.

PASSENGER TERMINAL LAYOUT AND DESIGN

Passenger requirements. As passenger throughput at airports increases, the passenger terminal becomes a more important element of the airport, attaining a dominant status in the largest facilities. The passenger terminal may amount to less than 10 percent of the total investment in a small airport, but at large airports terminals often account for more than 70 percent of infrastructural investment. The design that is ultimately adopted depends principally on the passenger volumes to be served and the type of passenger involved.

Passengers are frequently classified as business or leisure, scheduled or charter, originating or destined, and transfer or transit. Business travelers tend to pay significantly higher fares, and airlines usually wish to provide a high quality of service in order to attract such traffic. The passenger terminal at Heathrow Airport near London, for example, was designed to a very high standard of space and decor to attract just this type of passenger. Scheduled and charter passengers, meanwhile, tend to have very different needs in the terminal, especially at check-in and in the provision of ground transportation. Palma Airport, on the Spanish island of Majorca, has a landside that is designed to accommodate large numbers of charter tourists arriving and departing the airport by bus.

Transiting and transferring passengers

Some airports have a very high percentage of passengers who are either transiting the airport (*i.e.*, continuing on the

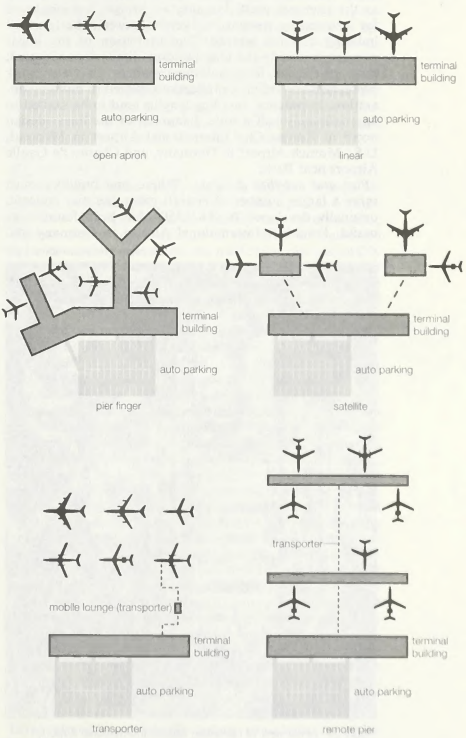


Figure 31: Six possible design concepts for airline passenger terminals.

same flight) or transferring to another flight. At Hartsfield Atlanta International Airport in Georgia and at Chicago's O'Hare International Airport, for example, two-thirds of all passengers transfer to other flights and have no wish to visit the cities where the airports are sited. These passengers have special needs but usually only on the airside of the terminal. There is no need to provide parking or ground transportation to the city for such passengers; they will, however, need transit lounges and other areas such as transit check-in desks.

Airports that receive a large number of transferring and transiting passengers are referred to as hubbing airports. At a hub, aircraft arrive in waves, and passengers transfer between aircraft during the periods when these waves are on the ground. By using a "hub-and-spoke" network, airlines are able to increase the load factors on aircraft and to provide more frequent departures for passengers—at the cost, however, of inconvenient interchange at the hub.

Terminal designs. The various concepts by which airport passenger terminals can be designed are shown in Figure 31.

Open-apron and linear designs. The oldest and simplest layout is the open apron design, in which aircraft park on the apron immediately adjacent to the terminal and passengers walk across the apron to board the aircraft by mobile steps. Frequently, the aircraft maneuver in and out of the parking positions under their own power. As airports grow, however, it is impossible to have large numbers of passengers walking across the apron. In this case, it is common to have terminals designed to the linear concept, with aircraft parked at gates immediately adjacent to the terminal itself. Usually, air bridges are employed for transferring passengers directly between the terminal building and the aircraft. The limitation of the linear concept is usually the long building dimensions required; these can mean long walking distances for transferring passengers and other complications related to building operation. In practice, building lengths tend to be limited to approximately half a mile. Examples of the linear design occur at Kansas City International Airport in Missouri, U.S., Munich Airport in Germany, and Charles de Gaulle Airport near Paris.

Pier and satellite designs. Where one building must serve a larger number of aircraft gates, the pier concept, originally developed in the 1950s, has been found very useful. Frankfurt International Airport in Germany and Schiphol Airport near Amsterdam still use such terminals. In the late 1970s, pier designs at Chicago's O'Hare and Atlanta's Hartsfield successfully handled in excess of 45 million mainly domestic passengers per year. However, as the number of aircraft gates grows, the distances that a passenger may have to travel within a pier-type terminal become exceedingly long, passenger circulation volumes become very large, and the terminal itself can become uncomfortable and unattractive to use. In order to cut down walking distances, some terminals, beginning in the 1960s, were designed on the satellite concept. Frequently, passengers are carried out to the satellites by some form of automated people mover or automatic train. Some satellite designs were very successful—for example, at Orlando and Tampa in Florida, U.S.—but to some degree the concept has fallen out of favour, having been found difficult to adapt to the changing size of aircraft and wasteful of apron space. Los Angeles International Airport originally had all its aircraft served at satellite buildings, but during the 1980s all satellites were converted to pier structures.

Inadequacy of the satellite design

Hartsfield Atlanta International Airport

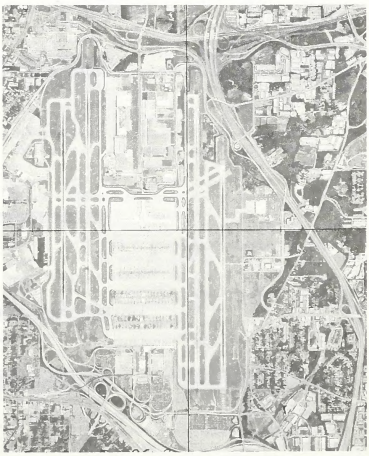


Figure 32: Aerial view of Hartsfield Atlanta International Airport, Atlanta, Ga., showing the independent close parallel runway configuration and the remote pier terminal design.

Transporter designs. In the early 1960s the transporter concept originated as a method of reducing aircraft maneuvering on the apron and of eliminating the need for passengers to climb up and down stairways in order to enter or exit the aircraft. In a concept derived from much older designs (such as that at Linate in Milan, where ordinary apron buses are used), passengers are brought directly to the aircraft by a specialized transporter vehicle. Mobile lounges used at Dulles International Airport near Washington, D.C., at Jiddah's King Abdul Aziz International Airport, and at Mirabel International Airport near Montreal have bodies that can be raised and lowered to suit the exact height of the terminal floor and the aircraft sill. However, passenger loading and unloading times are lengthened, causing turnaround delays, and aircraft are more likely to be damaged by the heavy lounges. For such reasons, this type of design has not proved popular with either passengers or airlines.

Remote-pier designs. The remote pier was introduced at Atlanta's Hartsfield in the early 1980s. In this concept, passengers are brought out to a remote pier by an automatic people mover and there embark or disembark in the conventional manner. The system has proved very efficient for handling transfer passengers, but the long distances involved in the terminal layout necessitate the use of a sophisticated people-mover system. The design of the terminal at Stansted, the third London airport, incorporates this concept.

Unit terminals. The term unit terminal is used wherever an airport passenger terminal system comprises more than one terminal. Unit terminals may be made up of a number of terminals of similar design (*e.g.,* Dallas–Fort Worth and Kansas City in the United States), terminals of different design (*e.g.,* London's Heathrow, Pearson International Airport near Toronto, John F. Kennedy International Airport in New York City), terminals fulfilling different functions (*e.g.,* Heathrow, Arlanda Airport near Stockholm, Barajas Airport near Madrid), or terminals serving different airlines (*e.g.,* Charles de Gaulle, John F. Kennedy, Dallas–Fort Worth). The successful operation of unit terminal airports has often required the design of rapid and efficient automatic people movers such as those at Changi Airport in Singapore, at Dallas–Fort Worth, and at Houston Intercontinental Airport in Texas.

The apron area. One of the important requirements in the design of a terminal complex is minimizing the time needed to service an aircraft when it is transiting an airport. This is especially important in the handling of short-haul aircraft, where unproductive ground time can consume an unacceptably large percentage of flight time. The turnaround time for a large passenger transport between short-haul flights can be as little as 25 minutes. During this period, a large number of service vehicles circulate on the apron. Therefore, an important aspect in the efficient operation of an airport facility is the marshaling of ground service vehicles and aircraft in the terminal apron area. Such an operation can become extremely complex at some of the world's busiest airports, where an aircraft enters or leaves the terminal apron approximately every 20 seconds.

Minimizing aircraft service time

Immediately after World War II, new airports were designed with their terminals entirely surrounded by runways, with access to the terminal areas provided by tunnels. Such layouts include John F. Kennedy at New York and London's Heathrow. Because many large airports now discourage general aviation use by small aircraft, making crosswind runways unnecessary, the newest designs (including airports at Dallas–Forth Worth, Munich, and Singapore) have linear layouts of terminals set between parallel runways. Figure 32 shows the linear development of Atlanta's Hartsfield terminal within a complex of four parallel runways. Such designs have been found to provide the most efficient relation between terminal apron space and runways, minimizing marshaling problems and delays at the apron entrances.

The design process. The design of airport passenger terminals is a complex procedure normally undertaken only by experts in that area. Buildings are planned to accommodate comfortably the passenger flows that are forecast to occur at some future date. In addition, it is anticipated that terminals will have to operate at levels slightly over capacity for short periods. The measurements used are various, including such terms as standard busy rate and typical peak-hour passenger flow, but by using these design procedures it is possible to plan a facility that will have insufficient capacity for no more than 30 operating hours per year. Under such conditions, only about 5 percent of passengers would be subjected to inconveniences caused by lack of capacity.

CARGO FACILITIES

The value of air freight

Less than 1 percent of all freight tonnage is carried by air. Nonetheless, this statistic significantly underestimates the importance of air freight because, in value of cargo moved, air transport dominates all other modes. For example, although Heathrow Airport handles less than three-quarters of a million tons of freight per year, in value of throughput it ranks as Britain's premier port. By the early 1990s Tokyo's Narita Airport, New York's John F. Kennedy, and Frankfurt International Airport handled in excess of one million tons of cargo per year.

As is the case with passenger facilities, freight terminals vary greatly in the volumes of material handled. Consequently, the scale of the building facilities and the nature of the handling methods also vary. Because only 10 percent of air cargo is carried loose or in bulk, all modern air-cargo facilities are designed to handle containers. In countries where labour is cheap and where freight throughputs at the terminal are not high, freight-handling systems can still be economically designed around the manhandling concept. This is not feasible in developed countries, where labour costs are high. Even at facilities with small throughputs, freight is moved by mobile mechanical equipment such as stackers, tugs, and forklift trucks. At high-volume facilities, a mixture of mobile equipment and complex fixed stacking and movement systems must be used. The fixed systems, which require complex engineering design and maintenance, are known as transfer vehicles (TVs) and elevating transfer vehicles (ETVs).

In the design of air-cargo facilities, special attention must be given to the handling of very heavy and oversized freight, perishables, urgent materials such as serums and human donor organs, high-value goods such as diamonds and gold, hazardous goods, and livestock.

An area of very fast growth in the air-cargo business is specialized movement by integrated carriers such as the U.S.-based Federal Express, which offer door-to-door delivery of small packages at premium rates. In its early years, this type of freight grew by more than 17 percent per annum. Cargo terminals for the small-package business are designed and constructed separately from conventional air-cargo terminals. They operate in a different manner, with all packages being cleared on an overnight basis.

LINKS TO LOCAL GROUND TRANSPORTATION

An airport should always be considered an interchange where different modes of transportation connect. Since the airport itself is not a primary destination, consideration must be given to access by surface vehicles. This is as critical a factor in airport layout and design as it is in the process of site selection. A large airport can quite easily generate in excess of 100,000 daily access trips by passengers and the same number of trips by workers, visitors, and suppliers. Such a scale of surface movement requires careful consideration of the design of internal circulation roads and access highways to the city centre and to the economic hinterland served by the airport. Additionally, road-based access requires the careful design of drop-down and pick-up areas and of both long-term and short-term parking. Larger airports are able to sustain economically viable links to taxi, limousine, and bus services. In addition, many of the world's largest airports are linked to intercity, suburban, and metro-style rail systems.

Journey time between airport and city centre

Peak traffic on airport approach roads tends to occur in the morning and evening, coinciding with other peaks of suburban traffic, so that it is difficult to forecast journey times between an airport and the downtown areas during rush periods. Even so, as major airports have gradually moved farther away from city centres, journey times to airports have increased. Road-traffic congestion has accentuated the problem, and the cost of delay has become critical in the economic comparison of air and surface transportation. In fact, large airports located on sites remote from the cities they serve have proved to be very unpopular with both passengers and airlines (*e.g.*, Mirabel at Montreal, Narita at Tokyo, and Dulles at Washington, D.C.).

A major argument for short takeoff and landing (STOL) aircraft lies in their lesser requirements for airport space, promising sites close to the central city. However, STOL airports have not proved economically viable, mainly because the aircraft used have been small, expensive to operate, and (at least in popular perception) uncomfortable. Furthermore, STOL aircraft have relatively low cruising speeds, which have limited their useful range. Helicopters, also known as vertical takeoff and landing (VTOL) aircraft, are in general too expensive for civil aviation transport, but the introduction of large tilt-rotor aircraft might revolutionize the use of centre-city airports.

AIRPORT CAPACITY

The various facilities at an airport are designed to cope adequately with the anticipated flow of passengers and cargo. The flow that any particular facility can accommodate without serious inconvenience to the users is considered to be its capacity. Limits on the traffic that can reasonably be accommodated at an airport are reached in a number of ways. These include air traffic delays to landing and takeoff movements; congestion on runways, taxiways, and aprons; crowding and delays in terminal buildings; or severe congestion in such access facilities as parking areas, internal roads, and public transport.

At smaller one-runway airports, limits to capacity usually occur in the terminal areas, since the operational capacity of a single runway with adequate taxiways is quite large. When passenger volumes reach approximately 25 million per year, a single runway is no longer adequate to handle the number of aircraft movements that take place during peak periods. At this point at least one additional runway, permitting simultaneous operation, is required. Airports with two simultaneous runways should be able to handle approximately 55 to 65 million passengers per year, and here, too, the main capacity problems are related to the provision of adequate terminal space. Layouts with four parallel runways, as shown in Figure 30, are estimated to have operational capacities of well over one million aircraft movements per year and annual passenger movements in excess of 100 million. The main capacity constraints of such facilities are in the provision of sufficient airspace for controlled aircraft movements and in the provision of adequate access facilities. It is likely that many of the world's largest airports will face access problems before they reach the operational capacity of their runways.

DRAINAGE

Large airports are actually urban complexes in which high-population activity centres are closely associated with very extensive paved areas. Typically a large airport can,

on a daily basis, handle more than 100,000 passengers and support a working population of over 50,000 employees. The sewage system of such an airport must cope with large daily flows of sanitary sewage effluent and, in addition, must accommodate runoff from rain and snow accumulating over several hundred acres of impervious pavement. The scale of the sewage problem at many large airports is such that some facilities have their own sewage treatment plants, especially for sanitary sewage. Because many airports are situated on low-lying ground, which is more likely to provide the flat land necessary for airstrips, the sewage system must often include extensive pumping facilities.

Contaminated runoff water

Growing concern about the environment and the increasing scale of activity at many airports has meant that runoff water can no longer be drained directly into bodies of surface water such as rivers and lakes. In particular, deicing chemicals used on aircraft and airfield pavements and cleaning chemicals used in aircraft maintenance are serious contaminants of groundwater and surface water. Consequently, some airports are required to provide at least primary treatment of all runoff discharges, and there are legal restrictions on the nature of the chemicals that can be used. In order to prevent groundwater pollution, the Munich Airport was designed to accommodate existing flows of surface water across the entire site and was also provided with extensive arrangements for the recycling of deicing chemicals.

NOISE

By the early 1960s, aircraft noise in the vicinity of urban airports had become a major problem. The cause of the problem was a rapidly increasing number of aircraft movements and the introduction of the first generation of turbojet aircraft with low climb performance, such as the early models of the Boeing 707 and the Douglas DC-8. Subsequently, public objections arose to the planned expansion of most urban airports. These objections often held up expansion for many years and, in cities such as London and Munich, ultimately severely modified the development of new airports. In addition, noise curfews were introduced at many existing airports, such as John F. Kennedy in New York, London's Heathrow, Kingsford Smith Airport near Sydney, Australia, and Hong Kong International Airport.

In response to national and international regulations aimed at certifying only quieter aircraft, major efforts have been made by aircraft manufacturers to reduce noise at the source. Successive generations of aircraft have been banned as they failed to meet increasingly severe requirements introduced by the International Civil Aviation Organization and the U.S. Federal Aviation Administration. The introduction of high-bypass turbofan engines and aircraft with high climb performance have helped considerably in reducing noise.

Airports can diminish aircraft noise in a number of ways. Hours of operation can be limited through the use of night curfews, night noise having been found much more objectionable to the public than day noise. Noisy aircraft can be restricted or even banned. Runways can be selected to limit or spread noise more evenly over the community, and approach and departure routes can be designated over less-populated areas. Airlines can also be encouraged to modify approach and departure gradients and operating procedures in order to reduce engine thrust over highly populated areas. The performance of the operators is monitored, and offending operators have financial penalties imposed upon them. Using such methods, many major airports—even those that anticipate long-term growth in passenger numbers and aircraft movements—can significantly improve the exposure of urban populations to aircraft noise. (N.J.A.)

CANALS AND INLAND WATERWAYS

Despite modern technological advances in air and ground transportation, inland waterways continue to fill a vital role and, in many areas, to grow substantially. This article traces the history of canal building from the earliest times to the present day and describes both the constructional and operational engineering techniques used and the major inland waterways and networks throughout the world.

Transport by inland waterways may be by navigable rivers or those made navigable by canalization (dredging and bank protection) or on artificial waterways called canals. Many inland waterways are multipurpose, providing drainage, irrigation, water supply, and generation of hydroelectric power as well as navigation. The lay of the land (topography) and particularly changes in water levels require that many rivers be regulated to make them fully navigable, thus enabling vessels to proceed from one water level to another. The chief regulating method is the lock, the development of which contributed significantly to the Industrial Revolution and the development of modern industrial society.

For many types of commodities, particularly such bulk commodities as grains, coal, and ore, inland waterway transport is still more economical than any other kind of transport. Thus, it is hardly surprising that modernized inland waterways, using the latest navigational aids and traction methods and traversing the great landmasses of North America, Europe, and Asia, play an increasingly important economic role.

History

ANCIENT WORKS

Most of the improvement of rivers and construction of artificial waterways in antiquity was for irrigation purposes. In the 7th century BC the Assyrian king Sennacherib built a 50-mile (80-kilometre) stone-lined canal 66 feet (20 metres) wide to bring fresh water from Bavian to Nineveh. The work, which included a stone aqueduct 300 yards (330 metres) long, was constructed in one year and three months, according to a plaque that survives on the site. Surprisingly advanced techniques were used, including a dam with sluice gates allowing regulation of the flow of the water stored. The Phoenicians, Assyrians, Sumerians, and Egyptians all constructed elaborate canal systems. The most spectacular canal of this period was probably Nahrawān, 400 feet wide and 200 miles long, built to provide a year-round navigation channel from near Sāmarrā to Al-Kūt, using water provided by damming the unevenly flowing Tigris. Many elaborate canals are known to have been built in Babylonia. In Egypt the Nile was dammed to control its floodwaters, and an extensive system of basin irrigation was established. The Persian king Darius in the 5th century BC cut a canal from the Nile River to the Red Sea. The Romans were responsible for very extensive systems of river regulation and canals in France, Italy, The Netherlands, and Great Britain for military transport. The legions in Gaul canalized one of the mouths of the Rhône to protect their overseas supply route. In the 1st century AD the Roman consul Marcus Livius Drusus dug a canal between the Rhine and Yssel to relieve the Rhine of surplus water, and the Roman general Corbulo linked the Rhine and Meuse with a canal 23 miles (37 kilometres) long to avoid the stormy North Sea passage from Germany to the coast. Attempting to reclaim the Fens in England, the Romans connected the River Cam with the Ouse by an eight-mile canal, the Nene with the Witham by one 25 miles long, and the Witham with the Trent by the Fosse Dyke (ditch), still in use.

Use of sluice gates

Outside Europe and the Middle East, between the 3rd century BC and the 1st century AD, the Chinese built impressive canals. Outstanding were the Ling Ch'ü in Kuangsi, 90 miles long from the Han capital; Ch'ang-an (Sian) to the Huang Ho (Yellow River); and the Pien Canal in Honan. Of later canals the most spectacular was the Grand Canal, the first 600-mile section of which was opened to navigation in 610. This waterway enabled grain

to be transported from the lower Yangtze and the Huai to K'ai-feng and Lo-yang. These canals had easy gradients (changes in water levels); and at about three-mile intervals there were single gates of stone or timber abutments with vertical grooves up or down along which the log closure was manually hauled by ropes to hold or release the water, thus controlling the water level. A few more elaborate gates had to be raised by windlasses. Where water level changes were too great for such simple devices, double slipways were built and vessels were hauled up the inclines.

MEDIEVAL REVIVAL

In Europe, canal building, which appears to have lapsed after the fall of the Roman Empire, was revived by commercial expansion in the 12th century. River navigation was considerably improved and artificial waterways were developed with the construction of stanches, or flash locks, in the weirs (dams) of water mills and at intervals along the waterways. Such a lock could be opened suddenly, releasing a torrent that carried a vessel over a shallow place. The commercially advanced and level Low Countries developed a system of canals using the drainage of the marshland at the mouths of the Schelde, Meuse, and Rhine; about 85 percent of medieval transport in the region went by inland waterway.

Use of stanches

Because shipping was handicapped where barges had to be towed over the weirs with windlasses or manually, the lock and lock basin were evolved to raise boats from one level to another. Although a primitive form of lock had been in operation as early as 1180 at Damme, on the canal from Brugge to the sea, the first example of the modern pound lock, which impounded water, was probably that built at Vreeswijk, Neth., in 1373, at the junction of the canal from Utrecht with the Lek River. Outer and inner gates contained a basin, the water level of which was controlled by alternatively winding up and lowering the gates. In the 15th century the lock-gate system was much improved with the addition of paddles to control the flow of water in and out of the lock chamber through sluices in the gates or sides of the lock.

Commercial needs soon encouraged canal construction in less ideal locations. The Stecknitz Canal, built in Germany (1391–98), ran 21 miles from Lake Möllner down to Lübeck, with a fall of 40 feet controlled with four stanches; the canal was later extended south to Lauenburg on the Elbe to establish a link between the Baltic and the North Sea. To deal with a fall from the summit to Lauenburg of 42 feet in 15 miles, two large locks were built, each capable of holding 10 small barges.

Italy, the other principal commercial region of medieval Europe, also made important contributions to waterway technology. The Naviglio Grande Canal was constructed (1179–1209) with an intake on the Ticino River, a fall of 110 feet in 31 miles to Abbiategrasso and Milan, the water level being controlled by sluices. To facilitate transport of marble from the quarries for the building of the Milan cathedral, the canal was linked with an old moat, and in Italy the first pound lock with mitre instead of the earlier portcullis gates was constructed to overcome differences in water level.

China may have been ahead of Europe in canal building. Between 1280 and 1293 the 700-mile northern branch of the Grand Canal was built from Huai-an to Peking. One section, crossing the Shantung foothills, was in effect the first summit-level canal, one that rises then falls, as opposed to a lateral canal, which has a continuous fall only. The Huang Ho (Yellow River) was linked with a group of lakes about 100 miles south, where the land rose 50 feet higher; and, to overcome water lost through operation of the lock gates, two small rivers were partially diverted to flow into the summit level.

THE 16TH TO 18TH CENTURY

The mitre lock

The development of the mitre lock, a double-leaf gate the closure of which formed an angle pointing upstream, heralded a period of extensive canal construction during the 16th and 17th centuries. The canals and canalized rivers of that period foreshadowed the European network to be developed over many years.

France. In France the Briare and Languedoc canals were built, the former linking the Loire and Seine and the latter, also known as the Canal du Midi, linking Toulouse with the Mediterranean. Both were remarkable feats of engineering. The Briare Canal (completed 1642) rose 128 feet to a plateau with a summit level 3.75 miles long and then dropped 266 feet to the Loing at Montargis. It included 40 locks, of which a unique feature was a staircase of six locks to cope with the fall of 65 feet on the descent from the Loing to Rogny. Construction of the 150-mile Canal du Midi joining the Bay of Biscay and the Mediterranean via the Garonne and the Aude ran through very rugged terrain. Begun in 1666 and finished in 1692, it rose 206 feet in 32 miles from the Garonne at Toulouse to the summit through 26 locks, and, after a three-mile stretch along the summit, then descended 620 feet through 74 locks for 115 miles. Near Béziers a staircase of eight locks was built, and six miles farther upstream a tunnel 180 yards long was constructed; three major aqueducts carried it over rivers, and numerous streams were diverted beneath it in culverts. The most notable technical achievement was a complex summit water supply that included unique diversion of flows and storage provision.

Flanders. The canal system in Flanders included one from Brussels to Willebroeck on the Rupel to shorten navigation by half, an 18½-mile canal with four locks; another of 44 miles was constructed from Brugge to Passchendaele, Nieuport, and Dunkirk and was later extended to Ostend, while Dunkirk was linked with the Aa River, at the mouth of which a large tide lock was constructed at Gravelines. The outstanding achievement in Flanders was a lock at Boesinghe on the canal from Ypres to Boesinghe beside the Yser River. The fall of 20 feet on this four-mile stretch was contained by a single large lock. Side ponds with ground sluices were provided for the first time to reduce the loss of water during the lock's operation. The ponds took one-third of the water when the lock was emptied and returned it for the filling.

Germany. In Germany the 15-mile Friedrich Wilhelm Summit Canal, completed in 1669, rose from Neuhaus on the Spree for 10 feet in two locks and from west of the summit fell 65 feet to Brieskow on the Oder. An extensive system of waterways in this part of Germany was finally established with the opening of the Plauer Canal in 1746, which ran from the Elbe to the Havel. The 25-mile Finow Canal along the Havel to the Liepe, a tributary of the Oder, had been built earlier but fell into decay because of flooding and neglect and was not rebuilt until 1751. In the late 17th and early 18th centuries, under the great elector of Brandenburg and Frederick I of Prussia, the three great rivers, the Elbe, Oder, and Weser, were linked by canal for commercial and political reasons, including the bypassing of tolls charged by the numerous states and petty principalities of the Holy Roman Empire. In the Low Countries, wars, political considerations, and the rivalry between the Dutch and Belgian ports handicapped canal building. The Dutch, for example, strongly opposed a Rhine-Meuse-Schelde canal, fearing diversion of trade to Antwerp.

The linking of the Elbe, Oder, and Weser rivers

Great Britain. The first lock was not built on an English canal until the 16th century, and the canal era proper dates from the construction of the Bridgewater Canal to carry coal from Worsley to Manchester in the 18th century by the engineer James Brindley. Opened for navigation in 1761, it was extended to the Mersey in 1776. Its success promoted a period of intense canal construction that established a network of inland waterways serving the Industrial Revolution and contributing to Britain's prosperity in the half-century preceding the railway era, which began in the mid-19th century. The Grand Trunk Canal established a cross-England route by linking the Mersey to the Trent, opened up the Midlands, and provided water transport for exports to European markets. There followed the link between the Thames and the Bristol Channel provided by the Severn Canal and the Gloucester and Berkeley Ship Canal from Sharpness on the Severn to Gloucester. Birmingham's growth and industrial prosperity were stimulated because the city became the centre of a canal system that connected London, the Bristol Channel, the Mersey, and

the Humber. The Caledonian Ship Canal across Scotland, joining the chain of freshwater lakes along the line of the Great Glen, was built between 1803 and 1822.

One of the few canals to be built after the canal era was the 36-mile-long Manchester Ship Canal, which was opened in 1894 to give oceangoing vessels access from the Mersey estuary to Manchester.

Technological development. This spate of canal construction was accompanied by technological development in both construction methods and operation. Locks, inclined planes, and lifts were developed to cope with changes in water level. At Bingley, for example, on the Leeds and Liverpool Canal, a lock staircase was built; and on the hilly areas at Ketley in Shropshire, inclined planes were constructed in 1788 to haul tugboats from one level to another. The longest plane, about 225 feet, was on the Hobbacott Down plane of the Bude Canal in Cornwall. Vertical lifts counterweighted by water were also used; a set of seven was built on the Grand Western Canal; while at Anderton in Cheshire a lift was later converted to electrical power and was still operating in the 20th century. The most spectacular inclined plane was built in the United States on the Morris Canal, which linked the Hudson and Delaware rivers. For a rise of 900 feet to the Alleghenies watershed, 22 locks were installed at the head of an inclined plane and, descending on a gradient of 1 in 10 to 1 in 12, ran down to the pound below. Barges 79 feet long with loads up to 30 tons were hauled up by trolleys running on rails, on which they settled as the lock emptied; the barges descended under gravity into the lower pound to float on an even keel when the water leveled off. In the reverse direction, they were hauled up by a drum-and-cable mechanism.

THE 19TH CENTURY

Europe. In Europe, where the canal era had also started toward the end of the 17th century and continued well into the 18th, France took the lead, integrating its national waterway system further by forging the missing links. In the north the Saint-Quentin Canal, with a 3½-mile tunnel, opened in 1810, linking the North Sea and the Schelde and Lys systems with the English Channel via the Somme and with Paris and Le Havre via the Oise and Seine. In the interior the Canal du Centre connected the Loire at Digoin with the Sâone at Chalon and completed the first inland route from the English Channel to the Mediterranean; the Sâone and Seine were linked farther north to give a more direct route from Paris to Lyon; the Rhine-Rhône Canal, opened in 1834, provided a direct north-to-south route; while the Sambre-Oise Canal linked the French canal system with the Belgian network via the Meuse. Toward the end of the 19th century, France embarked on the standardization of its canal system to facilitate through communication without transshipment. The ultimate result was a doubling of traffic between the opening of the century and World War II.

Industrial development in the early 19th century prompted Belgium to extend its inland waterways, especially to carry coal from Mons and Charleroi to Paris and northern France. Among the new canals and extensions built were the Mons-Condé and the Pommeroeul-Antoing canals, which connected the Haine and the Schelde; the Sambre was canalized; the Willebroek Canal was extended southward with the building of the Charleroi-Brussels Canal in 1827; and somewhat later the Campine routes were opened to serve Antwerp and connect the Meuse and Schelde. When the growth of the textile trade in Ghent created a need for better water transport, the Gent Ship Canal, cut through to Terneuzen, was opened in 1827, giving a shorter route to the sea. The Dutch extended their canals to serve the continental European industrial north. The Maastricht-Liège Canal was opened in 1850, enabling raw materials and steel to be transported from the Meuse and Sambre industrial areas by waterway throughout The Netherlands. In 1824 a long ship canal was built to bypass silting that obstructed navigation on the IJsselmeer (Zuiderzee) and to enter the North Sea in the Texel Roads. Later an even shorter ship canal was built to IJmuiden.

Dutch canals

In Scandinavia new canals were built to facilitate transport of timber and mineral products. In 1832 the new Göta Canal was opened, crossing the country from the Baltic to the Skagerrak and incorporating 63 locks. The political climate was less favourable for canal building in central Europe, but the Ludwig Canal, forming part of the Rhine-Main-Danube route, was opened in 1840. At the same time, steps were taken to improve river navigation generally, to provide speedier transport, and to enable a greater volume of freight to be carried. The Danube was regulated for 144 miles from Ennsmundung to Theuben, and the Franz Canal was dug in Hungary to join the Danube and Tisza. A nationwide Russian canal system connecting the Baltic and Caspian seas via the Neva and Volga rivers became navigable in 1718. A more direct route was established in 1804 with a canal between the Beresina and Dvina rivers. In the 19th century Russia made connections between the heads of navigation of its great rivers, the Volga, Dnepr, Don, Dvina, and Ob.

The Corinth ship canal

An outstanding engineering achievement in Greece was the cutting of a deep ship canal at sea level through the Isthmus of Corinth to connect the Aegean and Ionian seas. The Roman emperor Nero had first attempted this linking in the 1st century AD; the shafts sunk by him were reopened and sunk to their full depth. The canal, about 3.9 miles long, has a minimum depth of 26.2 feet and a minimum width of 68.9 feet at the bottom increasing to 80.7 feet at surface level. Dug in 1881–93, it is bounded by almost vertical rock cliffs that rise to more than 259 feet above water level in the canal's midsection.

United States. In the United States, canal building began slowly; only 100 miles of canals had been built at the beginning of the 19th century; but before the end of the century more than 4,000 miles were open to navigation. With wagon haulage difficult, slow, and costly for bulk commodities, water transport was the key to the opening up of the interior, but the way was barred by the Allegheny Mountains. To overcome this obstacle, it was necessary to go north by sea via the St. Lawrence River and the Great Lakes or south to the Gulf of Mexico and the Mississippi. A third possibility was the linking of the Great Lakes with the Hudson via the Mohawk Valley. The Erie Canal, 363 miles long with 82 locks from Albany on the Hudson to Buffalo on Lake Erie, was built by the state of New York from 1817 to 1825. Highly successful from the start, it opened up the Midwestern prairies, the produce of which could flow eastward to New York, with manufactured goods making the return journey westward, giving New York predominance over other Atlantic seaboard ports. The Champlain Canal was opened in 1823; but not until 1843, with the completion of the Chambly Canal, was access to the St. Lawrence made possible via the Richelieu River. Meanwhile, Canada had constructed the Welland Canal linking Lakes Ontario and Erie. Opened in 1829, it overcame the 327-foot difference in elevation with 40 locks, making navigation possible to Lake Michigan and Chicago. Later the St. Mary's Falls Canal connected Lake Huron and Lake Superior. To provide a southern route around the Allegheny Mountains, the Susquehanna and Ohio rivers were linked in 1834 by a 394-mile canal between Philadelphia and Pittsburgh. A unique feature of this route was the combination of water and rail transport with a 37-mile portage by rail by five inclined planes rising 1,399 feet to the summit station 2,334 feet above sea level and then falling 1,150 feet to Johnstown on the far side of the mountains, where a 105-mile canal with 68 locks ran to Pittsburgh. By 1856 a series of canals linked this canal system to the Erie Canal.

The Susquehanna-Ohio Canal

Meanwhile, the Louisiana Purchase of 1803 had given the United States control of the Mississippi River, and it became the main waterway for the movement of Midwestern produce via New Orleans and the Gulf of Mexico. Developments included the Illinois-Michigan Canal, connecting the two great water systems of the continent, the Great Lakes and the Mississippi. Entering Lake Michigan at Chicago, then a mere village, the canal triggered the city's explosive growth. Several canals were constructed subsequently to link up with the Erie and Welland canals and the St. Lawrence, and a comprehensive network of inland waterways was established.

Impact of the railways. With the development of rail transport in the 19th century, canals declined as the dominant carriers of freight, particularly in the United States and Britain. In continental Europe the impact was less marked, because the great natural rivers already linked by artificial waterways constituted an international network providing transport economically without transshipment; the terrain was more favourable and the canals larger and less obstructed by locks. Elsewhere canals could not compete with rail. They were limited both in the volume carried per unit and in speed; they were too small, too slow, and fragmented; and the railways, as they became integrated into national systems, provided a far more extensive service with greater flexibility. The canals were further handicapped because they were not, for the most part, common carriers themselves but were largely dependent on intermediate carrying companies. Although transport on the canals was for some time cheaper than rail, the railways gradually overcame this advantage. To modernize and extend the waterways to enable larger boats to ply them, to reduce the number of locks that slowed down movement, and to provide a more comprehensive service all required capital investment on a scale that made the return problematic. The railways exploited the difficulties of the canals by drastic rate cutting that forced many canal companies to sell out to them. In Britain a third of the canals had become railway-owned in the 1840s and '50s, and many were subsequently closed down. In the United States, half the canals were abandoned. The railways thus succeeded in eliminating their competition and obtained a near monopoly of transport, which they held until the arrival of the motor age.

Takeovers by railways

The Kiel Canal. The 19th century saw the construction of the Kiel and Suez canals. The former carries tonnage many times that of most other canals. Frequent attempts had been made to make a route from the Baltic to the North Sea and thus to bypass the Kattegat and the dangerous Skagerrak. The Vikings had portaged ships on rollers across the 10-mile Kiel watershed, but not until 1784 was the Eider Canal constructed between the Gulf of Kiel and the Eider Lakes. A little more than 100 years later, to accommodate the largest ships, including those of the new German navy, the Kiel Canal was widened, deepened, and straightened, cutting the distance from the English Channel to the Baltic by several hundred miles. Running 59 miles from locks at Brunsbüttel on the North Sea to the Haltenau locks on the Gulf of Kiel, the canal crosses easy country but has one unique engineering feature. At Rendsburg, to give clearance to the largest ships, the railway was made to spiral over the city on an ascending viaduct that crosses over itself before running on to the main span above the water.

The Suez Canal. The Isthmus of Suez so obviously provided a short sea route from the Mediterranean to the Indian Ocean and beyond as against the sea voyage around Africa that a canal was dug in antiquity; it fell into disuse, was frequently restored, and finally was blocked about the 8th century. Later there were many projects and surveys, but nothing happened until 1854, when Ferdinand de Lesseps, who had served as a diplomat for France in Egypt, persuaded Sa'īd Pasha, the viceroy of Egypt, to grant a preliminary concession for construction of a new canal across the isthmus. A later report recommended a sea-level lockless canal between Suez and the Gulf of Pelusium; and the original concession was superseded by one granted in 1856 to the Suez Canal Company, an international consortium. The concession was to last for 99 years from the canal's opening to navigation, after which it was to revert to the Egyptian government; the canal was to be an international waterway, open at all times to all ships without discrimination. In addition to the ship canal, the company undertook to excavate a freshwater canal from the Nile at Būlāq to Ismailia, with a branch extending to the Suez, to be available for smaller ships. Work on the ship canal lasted 10 years, during which time political, financial, contractual, and physical difficulties were overcome, and the canal was opened on Nov. 17, 1869. As ultimately constructed, it was a 105-mile lockless waterway connecting the Mediterranean and the Red Sea. From its northern terminal at Port Said, the canal passes through the salt marsh area of Lake Manzala, with the freshwater canal running parallel. About 30 miles from Port Said, a seven-mile bypass built between 1949 and 1951 enables convoys to pass. At about the halfway point the canal enters Lake Timsah and passes Ismailia. Thence the waterway proceeds through the Bitter Lakes and on to Port Tawfiq, the southern terminal on the Red Sea, a few miles from the town of Suez. Since its construction, the canal has been constantly improved; originally 200 feet wide with a maximum depth of 24 feet, it was widened in 1954 to 500 feet at water level and 196 feet at a depth of 33 feet, with the main channel 45 feet deep, enabling ships of a maximum draft of 37 feet to navigate the canal.

Improvements in the Suez Canal

The canal remained open despite much political controversy. Nationalized by Egypt in 1956, it was blocked in 1967 after the Arab-Israeli War and remained so until 1975.

The Panama Canal. After his success with the Suez Canal, de Lesseps was attracted to the Isthmus of Panama, where many projects had been suggested for cutting a canal to join the Atlantic and Pacific oceans and thus make unnecessary the passage around South America. De Lesseps proposed a sea-level route via Lake Nicaragua, but construction difficulties forced him to abandon this project in favour of a high-level lock canal via Panama. Further problems, especially yellow fever among the work force, halted construction after about 78,000,000 cubic yards (60,000,000 cubic metres) of material had been excavated. Meanwhile, U.S. interest had been actively maintained, but the situation was complicated by political difficulties and questions of sovereignty. A treaty between Britain and the United States recognized the exclusive U.S. right to construct, regulate, and manage a canal across the isthmus; but Panama was Colombian territory, and the Colombia Senate refused ratification of a treaty with the United States. After a revolt, a treaty was signed with independent Panama that granted the United States exclusive use, occupation, and control of the Canal Zone in perpetuity. (See below *Waterway systems: Administration* for later history.)

Although preliminary work started in 1904, little real progress was made because of disputes over the type of canal that should be built; not until 1906 was the high-level lock plan finally adopted, as opposed to the previously favoured sea-level plan. Largely responsible for this decision was John F. Stevens, who became chief engineer and architect of the canal. Completed in 1914, the canal is 51.2 miles long. At its start from the large harbour area in Limon Bay on the Caribbean Sea, it rises more than 80 feet above sea level to the Gatun Lake through the Gatun Locks and is retained at the north by these locks and dam and at the south by the Pedro Miguel Locks and Dam. The waterway then runs through the Gaillard Cut, which channels it through the Continental Divide, then between the Pedro Miguel Locks and the Miraflores Lake at an elevation of 54 feet, ships to the Pacific Ocean being lowered by them to the Balboa Harbor entrance. The Gatun Lake, with its area of 166 square miles, is an integral part of the waterway and the principal source of its water. The minimum channel depth throughout the length of the canal is 37 feet and its width 300 feet. There are 23 angles or changes of direction between the entrances. Ships normally travel through the canal under their own power except in the locks, through which they are towed by electric locomotives. (E.A.J.D.)

Modern waterway engineering

Waterways are subject to definite geographic and physical restrictions that influence the engineering problems of construction, maintenance, and operation.

The geographic restriction is that, unlike roads, railways, or pipelines, which are adaptable to irregular natural features, waterways are confined to moderate gradients; and where these change direction, the summit pounds (ponds) require an adequate supply of water, while valley pounds need facilities for disposal of surplus.

The primary physical restriction is that vessels cannot

The transport unit

travel through water at speeds possible for road vehicles or railway wagons. Because transport economics are based on the transport unit (x tons moved y miles in 1 man-hour), waterways must provide larger tonnage units than those possible on road or rail in order to be competitive.

Modern waterway engineering, therefore, is directed toward providing channels suitable for larger vessels to travel faster by reducing delays at locks or from darkness and other natural hazards. While such channels and associated works are designed to minimize annual maintenance costs, the costs of operating vessels, locks, wharves, and other waterway works can be minimized by increased mechanization.

CHARACTERISTICS OF BASIC TYPES

Fundamentally, waterways fall into three categories, each with its particular problems: natural rivers, canalized rivers, and artificial canals.

On natural rivers navigation is subjected to seasonal stoppages from frost, drought, or floods, all of which lead to channel movements and to the formation of shoals. While minimizing natural hazards, attention is directed primarily to retaining the channel in a predetermined course by stabilization of banks and bed, by elimination of side channels, and by easing major bends to obtain a channel of uniform cross section that follows the natural valley.

On canalized rivers navigation is facilitated by constructing locks that create a series of steps, the length of which depends on the natural gradient of the valley and on the rise at each lock. Associated with the locks for passing vessels, weirs and sluices are required for passing surplus water; and in modern canalizations, such as the Rhône and the Rhine, hydroelectric generation has introduced deep locks with longer artificial approach channels, which require bank protection against erosion and, in some strata, bed protection against seepage losses.

On artificial canals navigation can depart from natural river valleys and pass through hills and watersheds, crossing over valleys and streams along an artificial channel, the banks and sometimes the bed of which need protection against erosion and seepage. The route of an artificial canal can be selected to provide faster travel on long level pounds (stretches between locks), with necessary locks grouped either as a staircase with one chamber leading directly to another or as a flight with short intervening pounds. Where substantial differences of level arise or can be introduced, vertical lifts or inclined planes can be constructed. Storage reservoirs must be provided to feed the summit pound with enough water to meet lockage and evaporation losses; other reservoirs can be introduced at lower levels to meet heavier traffic movements entailing more frequent lockage operation. If supplies are insufficient to offset the losses, pumps may be needed to return water from lower to upper levels.

CHANNELS

Channel design. Natural rivers and canalized rivers away from artificial cuts need no protection against seepage and only light protection of banks against erosion. The widening or cutting off of major bends assists navigation, but wholesale straightening is undesirable because the natural sinuosity of the river, though modified, should be retained. Local widening is effected by dragline excavators cutting into the channel and dumping the material ashore, where it is either used to form levees or removed elsewhere. Deepening or widening beyond the reach of shore-based excavators requires a floating plant that discharges to hopper barges for transport to a disposal point or to pipelines for pumping ashore.

Cross section of artificial canals

Artificial canals should provide a waterway with a cross-sectional area at least five, and preferably seven, times the cross-sectional area of the loaded vessel. In rock cuttings, such as those of the Corinth Canal, the waterway cross section could be rectangular, but the normal cross section is trapezoidal, with bed width three to four times, and surface width six to eight times, the width of the vessel, while the depth must be enough to allow the water displaced by the moving vessel to flow back under the hull.

Channel construction. The physical construction of a canal has been facilitated by the development of very large mechanical excavators. Walking draglines with 20-ton buckets, such as were used on the St. Lawrence Seaway, are more suitable for quarry or opencut coal workings; for general channel construction the more versatile tracked machines are preferred. Scrapers and dumper trucks with oversize pneumatic tires for fast travel over rough ground readily dispose of excavated materials to form embankments or other fill.

Water losses by percolation through bed or banks must be prevented on embankments and wherever permeable strata are encountered. While the watertight skin was originally obtained by a layer of puddled clay with protective gravel covering, other materials later became available, such as fly ash from power stations, sometimes with a cement admixture; bentonite; bituminous materials; sheet polythene; or concrete.

Bridges, aqueducts, and tunnels for waterways. Canals must frequently cross over or under roads and railways, rivers, and other canals. These crossings are made by a variety of bridges, sometimes carrying the road or railroad, sometimes carrying the canal. Most are fixed, though movable bridges are also used. On the Weaver River in England, four movable bridges, carrying main roads across the waterway, swing on pontoons.

Canals originally crossed valleys on heavy masonry structures supporting the full formation, including puddled clay lining. Cast-iron flanged and bolted troughs later provided a lighter and watertight channel; current practice uses concrete with bituminous sealing.

Canals were originally carried through hills and watersheds in small bricked tunnels through which vessels were propelled by manual haulage, by poling, or by legging—that is, by crewmen lying on their backs on the cabin and pushing with their feet against the tunnel roof. Later, tunnels were provided with towpaths.

Bank protection. On natural or canalized rivers of relatively large cross section, bank erosion can be checked by rubble roughly tipped or by natural growth such as reeds or willows.

On artificial canals of smaller dimensions, where passing vessels create a serious wash, some revetment (bank protection) is essential. Sloping banks are readily protected by close-laid stone pitching, by bundles formed of interwoven willow branches, or by bituminous carpet; more permanent protection is provided by steel or concrete piles, which are close-driven, overlapping or interlocked, and protected against impact damage by horizontal fendering above the waterline and by roughly tipped rubble below the waterline. In cuttings the slopes are stabilized by berms (level strips) 6 to 10 feet wide at intervals determined by the nature of the soil. On long embankments safety stop gates can minimize water losses in the event of a breach.

Towpaths. Originally provided for animal haulage, towpaths were adapted on many French canals for mechanical and electrical haulage until the general use of powered craft terminated this service in 1969. But the towpaths are still useful; in addition to providing ways for some local haulage by mechanical tractor, they provide valuable access to the canals for inspection and maintenance.

LOCKS

On canalized rivers and artificial canals, the waterway consists of a series of level steps formed by impounding barriers through which vessels pass by a navigation lock. Basically, this device consists of a rectangular chamber with fixed sides, movable ends, and facilities for filling and emptying: when a lock is filled to the level of the upper pound, the upstream gates are opened for vessels to pass; after closing the upstream gates, water is drawn out until the lock level is again even with the lower pound, and the downstream gates are opened. Filling or emptying of the chamber is effected by manually or mechanically operated sluices. In small canals these may be on the gates, but on larger canals they are on culverts incorporated in the lock structure, with openings into the chamber through the sidewalls or floor. While the sizes of the culverts and openings govern the speed of filling or emptying the chamber, the number and location of the openings determine the

extent of the water disturbance in the chamber: the design must be directed toward obtaining a maximum speed of operation with minimum turbulence. The dimensions of the chamber are determined by the size of vessels using, or likely to use, the waterway. Where the traffic is dense, duplicate or multiple chambers may be required; in long chambers intermediate gates allow individual vessels to be passed.

Lock dimensions

Lock dimensions vary from the small, narrow canal locks of England, with chambers 72 feet long and 7 feet wide, to the 1,500-ton capacity waterways of Europe, with chambers 650 by 40 feet. On the St. Lawrence Seaway the dimensions are approximately 800 by 80 feet; on the Mississippi and Ohio rivers, where push-towing units are operating, the dimensions rise to 1,200 by 110 feet.

On canalized rivers the present trend is for locks to be deeper, particularly where they form an integral part of a hydroelectric dam. On the Rhône the lock at Donzère-Mondragon has a depth of 80 feet; in Portugal, where the Douro was being developed in the early 1970s for power and navigation, the Carrapatelo Lock has a depth of 114 feet.

On artificial canals, where conservation of water is essential, depths do not normally exceed 20 feet: water consumption can be reduced by the provision of side pounds either adjacent to the lock, as at Bamberg on the Rhine-Main-Danube waterway, or incorporated in the lock walls, as in the (1899) Henrichenburg Lock on the Dortmund-Ems Canal.

Locks are located to provide good approach channels free from restrictions on sight or movement. Where traffic is heavy or push tows operate, adequate approach walls are needed both to accommodate vessels awaiting entry and to provide shelter from river currents while vessels move slowly into or out of the lock.

Lock gates. Movable gates must be strong enough to withstand the water pressure arising from the level difference between adjacent pounds. The most generally used are mitre gates consisting of two leaves, the combined lengths of which exceed the lock width by about 10 percent. When opened, the leaves are housed in lock wall recesses; when closed, after turning through about 60°, they meet on the lock axis in a V shape with its point upstream. Mitre gates can be operated only after water levels on each side have been equalized.

On small canals gates may be manually operated by a lever arm extending over the lock side; on large canals hydraulic, mechanical, or electrical power is used. On the Weaver Navigations Canal in England the hydraulic power for operating the lock gates has been derived for 100 years from the 10-foot head difference between the pounds.

Vertical gates, counterweighted and lifted by winch or other gearing mounted on an overhead gantry, can operate against water pressure; as the gate leaves the sill, water enters the chamber, supplementing or replacing the culvert supply. The turbulence is more difficult to control, and the overhead gantries impose restrictions on masts and other superstructures of a vessel.

The use of sector gates, which turn into recesses in the wall, depends on the physical characteristics of the site and on the traffic using the waterway; falling gates lower into recesses in the forebay, and rolling gates run on rails into deep recesses in the lock walls.

Lock equipment. Ladders recessed into the walls provide access between vessels and the lockside and are vital in case of accidents.

Bollards (mooring posts) on the lockside are used for holding vessels steady by ropes against the turbulence during lock operation; mooring hooks set in recesses in the walls provide an alternative anchorage against surging. Floating bollards are provided in deep locks; retained in wall recesses, they rise or fall with the vessel, obviating the need for continuous adjustment of the ropes. Signals, physical or visual, erected at each end of the lock indicate to approaching craft whether the lock is free for them to enter and, in the multiple-chamber locks, which chamber they should use. Control cabins, centrally situated, enable all operations of the lock gates, sluices, and signals to be carried out by one person from a push-button control panel. Telephone or radio communication between adjacent locks gives advance information enabling the operator to have a lock prepared in anticipation of a vessel's arrival. Experiments in France in the early 1970s were directed toward the automatic passage of a vessel through a flight of locks, the various operations at each lock, once initiated, continuing automatically until the vessel left.

Communications between locks

Lock bypasses. The passage of a small pleasure boat through a deep lock is an expensive operation if it is passed alone and can be hazardous if it is passed with large barges that might surge against it. Canoes are normally brought ashore and manually moved around a lock on a portable trolley; larger pleasure craft can be transported on a cradle that is mechanically towed on a lockside rail track.

Water chutes have been introduced in Germany for canoes and rowboats where there are rises of 30 to 80 feet; although more costly to install than a lockside rail track, they are more popular. The canoeist, entering the approach channel, pushes a button actuating the head gates, which rise to allow the water to carry the canoe into and down the chute, where it is kept in the centre of the chute by guide vanes. For upstream passage, canoes are kept afloat by descending water but require manual towage.

Boat lifts. Vessels can be transported floating in a steel tank or caisson between adjacent pounds by a vertical lift, replacing several locks. Vertical lifts can be operated by high-pressure hydraulic rams, by submersible floats, or by geared counterweights. Hydraulic lifts with twin caissons were constructed in 1875 at Anderton, Eng., with a 50-foot lift for 60-ton vessels; in 1888 lifts were constructed at Les Fontinettes, Fr., for 300-ton vessels and at La Louvière, Belg., for 400-ton vessels. Similar hydraulic lift locks were constructed at Kirkfield and Peterborough in Ontario, Can.; the latter, completed in 1904, has a lift of nearly 65 feet. Float lifts were constructed in 1899 at Henrichenburg, Ger., with a 46-foot lift for 600-ton vessels; in 1938 at Magdeburg, Ger., with a 60-foot lift for 1,000-ton vessels; and in 1962 a lift at Henrichenburg for 1,350-ton vessels.

Counterweighted lifts were introduced in 1908 when the Anderton lift was reconstructed. Each caisson was separately counterbalanced by a series of weights and ropes with electrically driven gearing. This method was used in 1932 at Niederfinow, Ger., with a 117-foot lift for 1,000-ton vessels.

Inclined planes. In the 18th century, inclined planes were constructed to transport small boats on trucks between adjacent pounds, using animal power and gravity and, later, steam. A series of planes was built in the United States on the canal between the Delaware and Hudson rivers to transport 80-ton vessels in caissons; a similar plane for 60-ton vessels was built at Foxton, Eng., to bypass 10 locks.

Three planes have been constructed in Europe, at Ronquières, Belg., for 1,350-ton vessels; at Saint-Louis-Arzviller, Fr., for 300-ton vessels; and at Krasnoyarsk, Russia, for 1,500-ton vessels. At Ronquières and Krasnoyarsk, vessels are carried longitudinally up relatively gentle inclines with gradients of 1 in 21 and 1 in 12, respectively, while at Arzviller the site permitted only a steep gradient of 1 in $2^1/_2$, necessitating vessels being moved transversely. At Ronquières the plane rises 220 feet and replaces 17 locks; at Arzviller the rise is 150 feet, and here, too, 17 locks have been replaced. At Krasnoyarsk the plane rises 330 feet from the downstream water level of the Yenisey River to surmount the hydroelectric dam; on top of the dam the caisson moves on to a turntable, where it is rotated through 38° before passing to a second plane running down to the water level impounded upstream of the dam.

INLAND WATERWAY CRAFT

While early navigation of natural rivers was dependent on the use of sail for upstream operation, towpaths and animal haulage were provided when rivers were canalized and artificial canals constructed. Later, mechanical haulage was developed and is still used for local movement of unpowered craft.

Steam, and later diesel, tugs improved speed of travel, particularly where lakes or estuarial lengths were encoun-

Tugs and barges

tered. Powered barges, towing one or more unpowered (dumb) barges, were introduced on rivers with adequate lock chambers; but on artificial canals double (or treble) lockage operations made this method uneconomical; and, except for local lighterage (loading, transporting, and unloading) or maintenance duties, dumb barges are little used on artificial canals.

To meet competition from road haulage, with its greater flexibility and higher speeds, water transport must find its solution in its capacity for larger units, thus necessitating the enlargement of channels and locks. Consequently, the 300-ton barges operating economically early in this century have been replaced by craft as large as 1,350 tons and more.

In North America, transport operators grouped dumb barges into assemblies, lashing them on either side or ahead of a power unit with similar barges secured in rows ahead. These assemblies of unpowered and individually unmanned barges are known, somewhat illogically, as push tows, and the power unit as a push tug. While these assemblies operate most advantageously on natural rivers, their development has justified heavy capital expenditure for enlarging lock chambers on some canalized rivers to avoid delays and increased operational costs arising from multiple lockage. In Europe, push tows normally operate with fewer than six barges, but on the Mississippi, with its deep channel and 700 miles without a lock, a push tow may aggregate 40,000 tons, an assembly of 40 barges being controlled by one 9,000-horsepower push tug, with cabins and facilities for 24-hour operation. On the Ohio River the original 600-foot lock chambers were lengthened to 1,200 feet to obviate double lockage.

Push tows and push tugs

Movement of push tows around bends, as on the Moselle River, is facilitated by portable power units attached to the bows and operated as required. Similar units can be attached to individual barges for transfer from push tow to wharf or vice versa; they can also be used for handling dumb barges in docks and for moving hopper barges short distances from dredger to disposal site.

WATERWAY MAINTENANCE

Inspection vessels, self-propelled and equipped with echo-sounding appliances, are necessary for regular survey of the waterway. On natural and canalized rivers, which are subject to droughts and floods, attention is particularly directed to the location of the navigable channel: transverse soundings reveal channel movements and enable marker buoys or perches to be relocated and shoals removed by dredging; longitudinal echo-sounding readings normally suffice to locate shallow lengths on artificial canals.

The dredging plant

The dredging plant is an expensive item of waterway maintenance. Bucket dredgers for major operations are supplemented by suction, or grab, dredgers for localized work; hopper barges are required for transporting dredged materials to disposal sites, which should be numerous enough to minimize the transporting period, so that the dredger remains fully operational with a minimum of hopper barges and towage units.

Bank revetment requires special vessels for carrying piling frames and light lifting tackle; other service craft are needed for concrete mixing and general duty.

Lock gate renewal is normally planned to ensure that a predetermined number of gates are replaced annually; special vessels equipped with heavy lifting tackle are needed for transport and site handling.

Divers carry out underwater inspections and repairs; although scuba-type diving has been developed for some underwater operations, helmet diving is still needed for prolonged work. Both types of diving require special craft with specialized crew and equipment for servicing the divers. Salvage craft equipped with pumps and heavy-lifting tackle are used for removing obstructions from the channel or for raising sunken vessels. Tugs handle the service vessels because many are used only intermittently, and thus power units are not economical. Dry docks or slipways, workshops, fitting shops, welding bays, and other special facilities, usually grouped in the vicinity of the administration offices, are part of every modern canal-maintenance system. (C.M.)

Waterway systems

ADMINISTRATION

Modern inland waterway development has been largely carried out by governments, in contrast to early canal construction, which was mainly undertaken by private enterprise. Most of the older canals were subsequently acquired by the state and are administered by them or their agencies and are subject to comprehensive regulation, frequently by independent commissions. International commissions representing the states concerned regulate navigation on the international waterways. In the United States the waterways are basically a federal responsibility, with their development undertaken by the U.S. Army Corps of Engineers, but state governments and local authorities also participate in the administration of many local waterways. The Interstate Commerce Commission has responsibility for the regulation of the common carriers and requires them to publish their rates. For some major multipurpose projects, public corporations were established to undertake and administer them.

In Europe and the former Soviet Union the national networks, mainly based on navigable and canalized rivers linked by canal, were developed by the governments, which retained responsibility for finance and administration. In Britain most canals were brought under government ownership beginning Jan. 1, 1948, and are administered by the British Waterways Board.

European waterway regulation

Europe's main waterways have long been accepted as international waterways with navigation free to all vessels and equality of treatment of all flags guaranteed. The chief regulatory commissions are the Central Commission for the Navigation of the Rhine, the Danube Commission, and the commission for the canalized Moselle. There are also a number of bilateral agreements between states. Wars and political considerations following them have from time to time interrupted the freedom of navigation. A provisional Rhine Commission was operating in the early 1970s; a new Danube Commission was established in 1953 after the signing of the Austrian state treaty, when freedom of navigation throughout the river's length was fully restored. With the creation of a number of international organizations in Europe, a high degree of cooperation between states for the development of the inland waterways and the regulation of navigation was achieved, particularly through the United Nations Economic Commission for Europe, the European Economic Community, the Organization for Economic Cooperation and Development, and the Council of Europe.

In North America a U.S.-Canadian International Joint Commission has functioned since 1909 with general authority over the boundary waters. The St. Lawrence Seaway is a joint project, administered by the St. Lawrence Seaway Authority in Canada and the St. Lawrence Seaway Development Corporation in the United States.

The Panama Canal was originally administered under the Panama Canal Convention of 1903 by the United States, under the supervision of the army. Panama-U.S. relations were frequently strained, and in 1964 the United States agreed to negotiate new treaties concerning the existing canal and construction of a new canal at sea level. Later both countries agreed to a new treaty recognizing Panama's sovereignty over the Canal Zone.

The international status of the Suez Canal, constructed and administered by the Suez Canal Company, has frequently been a matter for dispute, peaceful and otherwise. Only in 1904, under an Anglo-French agreement, was the Constantinople Convention of 1888, establishing the Suez Canal as an international waterway open to all in war and peace, finally implemented. In 1956 British presence in the area ended, and troops were withdrawn from the canal zone; the Egyptian government nationalized the assets of the canal company and the administration was assumed by Egypt, but the 1967 war closed the canal until 1975.

MAJOR INLAND WATERWAYS AND NETWORKS

Europe. After the end of World War II, the growth of transport by inland waterway in Europe, coordinated by the various international authorities, resulted in an en-



Figure 33: Inland waterway systems of western Europe.

larged and integrated network brought up to a minimum common standard for craft of 1,350 tons. With the Rhine, the Moselle, and their tributaries dominating the German system and providing outlets for the Dutch and Belgian systems and connecting with the French network, main improvements were concentrated on the international Main-Danube Canal and on improving the north-south route of the Nord-Sud Canal (or Elbe-Seitenkanal). The latter canal (completed in 1976) leaves the Elbe about 20 miles above Hamburg and, running south, joins the Mittelland Canal near Wolfsburg, Ger., reaching a total of 71½ miles and shortening the route between Hamburg and the Ruhr by 134 miles.

The Main-Danube waterway

The Main-Danube waterway connecting the Rhine with the Black Sea was completed in 1992 and provides a route for traffic between eastern and western Europe through Germany, accommodating craft of 1,350 tons throughout its length. Following the Main River to Bamberg in Germany, the route proceeds by artificial waterway, including a section of the Regnitz Canal to Dietfurt, thence by the Altmühl River to a point below Kelheim, where it joins with the Danube, crossing the Austrian border at Jochenstein. The 44-mile Bamberg-to-Nürnberg canal section, completed in 1972, includes seven locks with a combined lift of 268 feet. All locks are 623 feet long and able to accommodate vessels of 1,500 tons. Improvements of the channel of the German Danube, begun in 1965, include a pair of locks at Kachlet, just above Passau. In Austria four pairs of locks to take 1,350-ton craft have been built.

The damming of the Danube at Đerdap (1970–72), the Iron Gate rapids, on the border between Serbia and Romania, was undertaken in conjunction with the improvement of navigation through these dangerous waters; it incorporates vast hydroelectric power plants. Two locks, 1,017 feet long and 112 feet wide, with two chambers each, are being built to facilitate passage through the Iron Gate. Journey time for ships traveling from Black Sea ports upstream to Belgrade, Vienna, and central Europe will be reduced from approximately 100 to 15 hours by this project, and traffic is expected to rise from the present 12 million tons annually to 50 million tons.

France's waterway network of nearly 5,000 miles is based primarily on its rivers, but many of the low-capacity canals are being raised to the 1,350-ton standard. A major development planned in the 1970s in cooperation with West Germany was the construction to this standard of the North Sea–Mediterranean waterway via the canalized Rhône and Rhine rivers. With four existing locks built for the Grand Canal d'Alsace, a projected lateral canal between Huningue and Strasbourg, the project was modified in 1956, and the four remaining dams were to be built on the Rhine itself and bypassed with short canals including four locks, three with two chambers each. Canalization of the Rhône started with the building of the Port of Edouard-Herriot downstream from Lyon, and work proceeded on 12 locks and dams. Two new ports, serving Valence and Montélimar, were being constructed. Improvements were also made on the Marne-Rhine wa-



Figure 34: Canals and inland waterways in the Soviet Union.

terway, which provides an important internal trade route connecting the Paris Basin with the industrial regions of Alsace-Lorraine. The improvements included major works on either side of the Vosges summit level, replacing 23 old locks. At Réchicourt a new lock with a lift of 32½ feet bypasses six locks and a winding section of the old canal; on the other side of the summit a new canal section bypasses 17 locks, which formerly required 8 to 12 hours to navigate. On this section the inclined plane of Saint-Louis-Arzviller deals with a difference in level of 146 feet with a horizontal length of 422 feet. Two tanks each carry a 350-ton barge. Their 32 wheels run on four rails, and two sets of 14 cables connect the tanks to the two concrete counterweights. Improvements have been made to routes connecting the Seine with the north and east. The Canal du Nord was completed in 1965, and a bottleneck was removed on the Oise Lateral Canal with the building of two locks to accommodate through convoys to Paris.

In The Netherlands the extensive canal system based on large natural rivers and serving the ports of Rotterdam and Amsterdam has required comparatively little modernization; but to avoid the Maas (Meuse) River, between Roermond and Maastricht, the Juliana Canal was built in 1935 and improved after World War II. The Twente Canal, opened in 1936, improved communication with the industrial east. Most important of the postwar projects was the building of the Amsterdam-Rhine Canal to enhance the capital's value as a transshipment port. The Noord-Hollandsch Canal from Amsterdam to Den Helder was constructed, and the IJsselmeer was linked with the Ems estuary across the north of Holland. To shorten the distance between Rotterdam and Antwerp by 25 miles, the Schelde-Rhine Canal has been built.

Italy's waterway system, based on the Po Valley, is cut off from the European network by the Alps, but it is also being brought up to higher standards.

In Scandinavia there are two major commercial artificial waterways: the first, the Trollhätte Canal, connects the Götaälv (river) upward from Göteborg with Lake Vänern and with the Finnish lakes and connecting canals; the second, the Saimaa Canal, in southeast Finland, connecting the vast Saimaa Lake system to the sea, was being reconstructed at the time of World War II. After the Soviet-Finnish War, part was ceded to the Soviet Union; but in 1963 it was leased back to Finland, modernization continued, and the canal, with eight large locks replacing the previous 28, was reopened in 1968.

Principal Soviet links

In the Soviet Union, water navigation played a major role in the country's economy; and after World War I its great rivers—the Dnepr, Dvina, Don, Vistula, and Volga—were linked to form an extensive network, making through navigation possible from the Baltic to both the Black Sea and the Caspian. The Black Sea and the Baltic are connected by three different systems, of which the most important is the link between the Dnepr and the Bug, a tributary of the Vistula, by way of the Pripyat and Pina rivers, a 127-mile canal connecting with the Mukhavets River, a tributary of the western Bug. This system is the sole wholly inland waterway connection between western Europe and the Soviet systems, giving through access to the Caspian and Black seas. When the Rhine-Danube and Oder-Danube canals are completed, a second route will be provided, via the Berezina River, a tributary of the Dnepr, the Viliya, a tributary of the Niemen, and a 13-mile canal through Latvia to Riga. The last link reaches the Baltic through Lithuania and Poland from the Dnepr by way of the Szara, a tributary of the Niemen; the Jasiolda, a tributary of the Pripyat; and a 34-mile canal. Other important links are the Volga-Don Canal, 63 miles long and completed in 1952, and the Moscow-Volga Canal, built between 1932 and 1937, which flows 80 miles from the Volga to the Moskva River at Moscow. The White Sea–Baltic Canal, built in 1931–33, runs from Belomorsk on the White Sea through the canalized Vyg River across Lake Vyg and through a short canal to Povenets at the northern end of Lake Onega, through which it passes to the canalized Svir River, Lake Ladoga, and the Neva River to the southern terminal at Leningrad. The total length of the system is 140 miles, reducing sea passage between Leningrad and Arkhangelsk by 2,400 miles; through its 19 locks it rises to 335 feet above sea level. In the Soviet Union the Ob and Yenisey in Siberia are connected by canal, and the

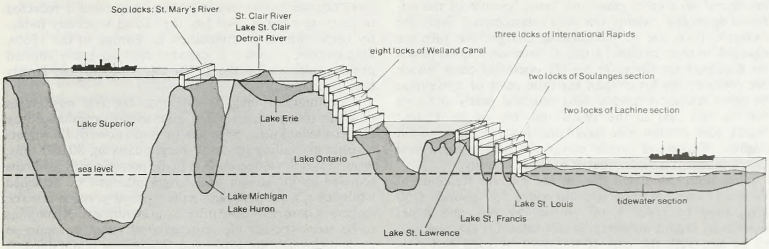


Figure 35: Cross section of the St. Lawrence Seaway showing how the succession of locks lifts the channel from sea level to its highest interior point at Lake Superior.
By courtesy of St. Lawrence Seaway Development Corp.

Karakumsky Kanal has been built from Kerki on the Amu Darya and is being continued westward to the Caspian.

North America. The U.S. and Canadian networks of inland waterways are based on the great navigable rivers of the continent linked by several major canals. Additionally, to reduce the hazards of navigating the Atlantic seaboard and to shorten distances, intracoastal waterways (protected routes paralleling the coast) have been developed. The total inland U.S. system, including protected coastal routes, approximates 25,000 miles, of which well over half has a minimum depth of nine feet. The largest system is based on the Mississippi, which is navigable for about 1,800 miles from New Orleans to Minneapolis, and its vast system of tributaries. This system connects with the St. Lawrence Seaway via Lake Michigan, the Chicago Sanitary and Ship Canal, and the Illinois River and with the Atlantic coast via the New York State Barge Canal (Erie Canal) and the Hudson River. The two intracoastal waterways are the Atlantic and the Gulf, the former extending from Boston, Mass., to Key West, Fla., with many sections in tidal water or in open sea. The Gulf Intracoastal Waterway comprises large sheltered channels running along the coast and intersected by many rivers giving access to ports a short distance inland. New Orleans is reached by the Tidewater Ship Canal, a more direct and safer waterway than the Mississippi delta. The Pacific coast canals are not linked with the national network, but two major projects of importance are the Sacramento Deepwater Ship Canal and the Columbia River development, which will provide more than 500 miles of navigable river from the Pacific to Lewiston, Idaho.

Mississippi system

The opening of the St. Lawrence Seaway in 1959 saw the fulfillment of a project that had been envisaged from the times of the earliest settlements in Canada. A continuous, navigable, deep waterway from the Atlantic to the Great Lakes was the obvious route for opening up the interior of North America; but natural obstacles, such as the Lachine Rapids north of Montreal, had prevented its realization. The completion of such a waterway required agreement between the United States and Canada, which was difficult to achieve. In 1912 the Canadian government decided to improve the Welland Canal to provide a 27-foot depth with locks 800 feet long and 80 feet wide; but because of World War I it was not completed until 1932. Although a joint project to include hydroelectric power development on the International Rapids section had been provisionally agreed upon, final agreement between Canada and the United States was not reached until the early 1950s. The Canadian government undertook to raise the standard of the waterway to a 27-foot navigation depth between Montreal and Lake Erie, and the United States agreed to carry out other works, including the bypassing by canal and locks of the Barnhart Island–Cornwall generating dam at the foot of the Long Sault Rapids. This agreement enabled work on the seaway to begin in 1954. The resultant deep waterway, navigable by oceangoing ships, extends about 2,300 miles from the Atlantic Ocean to the head of the Great Lakes in the heart of North America.

After Montreal Harbour the first lock is the St. Lambert, which rises 15 feet to the Laprairie Basin and proceeds 8.5 miles to the second Côte Ste. Catherine Lock, which rises 30 feet to Lake St. Louis and bypasses the Lachine Rapids. Thereafter, the channel runs to the lower Beauharnois Lock, which rises 41 feet to the level of Lake St. Francis via a 13-mile canal. Thirty miles farther, the seaway crosses the international boundary to the Bertrand H. Snell Lock, with its lift of 45 feet to the Wiley-Dondero Canal; it then lifts another 38 feet by the Dwight D. Eisenhower Lock into Lake St. Lawrence. Leaving the western end of the lake, the seaway bypasses the Iroquois Control Dam and proceeds through the Thousand Islands to Lake Ontario.

Locks

Eight locks raise the water 326 feet over 28 miles from Lake Ontario to Lake Erie. The St. Marys Falls Canal, with a lift of about 20 feet, carries the waterway to Lake Superior, where the seaway terminates.

Indian subcontinent. In Asia the full potential of the rivers for navigation had not been developed by the second half of the 20th century; though the rivers constitute one of the main means of transport, there are few navigable canals and no integrated system of inland waterways. The Indus River basin was being developed in the 1970s, however, and there were plans for international cooperation in developing inland waterway transport on the Mekong and Ganges-Brahmaputra systems.

ECONOMIC SIGNIFICANCE

Despite the large capital investment required to modernize existing inland waterway systems and for new construction, water transport has demonstrated competitive strength as a carrier for commodities in the movement of which the time factor is not of prime importance, such as minerals, timber, and many agricultural products. In the same way as the canals of the 19th century contributed to the development of the Midwest in the United States, the St. Lawrence Seaway has led to an expansion of industrial activity on the regions bordering the Great Lakes. Economic expansion along North America's rivers has followed capital investment in improvement of navigation along them. In the Soviet Union, similar development of vast areas was made possible by linking the major rivers to provide through routes.

Growth of water-transport tonnage

In continental Europe the eight member countries of the Conference of European Ministers of Transport (ECMT) experienced a growth in total tons carried by inland waterways from 385 million tons to 472 million tons in the years 1964–68. Whereas in 1938 Germany carried 90 million tons of freight on its inland waterways, by the end of the 1960s the Federal Republic of Germany alone was carrying over 230 million tons a year; East Germany was carrying an additional 12 million tons. Nor was this increase limited to the earlier years of the decade, as is shown by the volume of goods passing along the Rhine, which rose from 187 million tons in 1963 to 265 million tons in 1969. Most European countries had the same experience: the Soviet Union, which carried over its 233,000 miles of navigable waterways 239.5 million tons in 1963, transported 322.7 million tons in 1969.

It is difficult to judge the economics of water transport

compared with other transport forms because of the different operating systems. On most international rivers, for example, there are no navigational charges; but tolls are charged on most national artificial waterways. Costs of water transport are therefore mainly operating costs, which are considerably lower than the total costs of movement by other transport modes. This situation partly accounts for the fact that in the 1950s and '60s in the United States, costs per ton-mile stayed practically the same or fell slightly. Mergers of carrier companies and technological developments also contributed to price stability.

It has been calculated that in the Federal Republic of Germany one horsepower could move 330 pounds (150 kilograms) by road, 1,100 pounds by rail, and 8,800 pounds by inland waterway. Water-transport cost was said to be one-sixth the cost of transport by road and two-thirds the cost of transport by rail. Other transport carriers contend that such comparisons are not valid, because public investment in permanent structures (*i.e.,* canals and locks) is not always taken into account, whereas for railways private investment in right-of-way costs is reflected in carrying charges. Nor has the inland waterway industry been without its difficulties. In Europe in the 1960s, for example, a surplus of carrying craft adversely affected profits, although by the 1970s this problem had largely been overcome.

In summary it may be said that the real advantages of water transport are being maintained or enhanced by modern techniques, especially by more powerful towboats capable of hauling up to 50 barges carrying 80,000 tons; around-the-clock operation is made possible with towboats refueled in midstream and barges attached or detached while the tow proceeds along the river; at ports, automatic loaders cut turnaround time to a minimum. It remains to be seen whether the resurgence of water transport so evident through the 1960s and '70s will be maintained. A major question mark is the barge-carrying ship, analogous to railway piggybacking of truckloads, which promises to provide through transport by barge from inland ports across oceans to foreign inland destinations. (E.A.J.D.)

DAMS

A dam is a structure built across a stream, river, or estuary to retain water. Its purposes are to meet demands for water for human consumption, irrigation, or industry; to reduce peak discharge of floodwater; to increase available water stored for generating hydroelectric power; or to increase the depth of water in a river so as to improve navigation. An incidental purpose can be to provide a lake for recreation.

Auxiliary works at a dam may include spillways, gates, or valves to control the discharge of surplus water downstream from the dam; an intake structure conducting water to a power station or to canals, tunnels, or pipelines for more distant use; provision for evacuating silt carried into the reservoir; and means for permitting ships or fish to pass the dam. A dam therefore is the central structure in a multipurpose scheme aiming at the conservation of water resources. The multipurpose dam holds special importance in less developed countries, where a small nation may reap enormous benefits in agriculture and industry from a single dam.

Dams fall into several distinct classes, by profile and by building material. The decision as to which type of dam to build depends largely on the foundation conditions in the valley and the construction materials available. Broadly, the choice of materials now lies between concrete, soils, and rock fill. Although in the past a number of dams were built of jointed masonry, this practice is now largely obsolete. The monolithic form of concrete dams permits greater variations in profile, according to the extent to which water pressure is resisted by the deadweight of the structure, is transferred laterally to buttresses, or is carried by horizontal arching across the valley to abutments formed by the sides of the valley.

History

ANCIENT DAMS

The Middle East. The earliest recorded dam is believed to have been on the Nile River at Kosheish, where a 49-foot- (15-metre-) high masonry structure was built about 2900 BC to supply water to King Menes' capital at Memphis. Evidence exists of a masonry-faced earthen dam built about 2700 BC at Sadd-el-Kafara, about 19 miles (30 kilometres) south of Cairo; this dam failed shortly after completion when, in the absence of a spillway, it was overtopped by a flood. The oldest dam in use is a rock-fill structure about 20 feet high on the Orontes in Syria, built about 1300 BC.

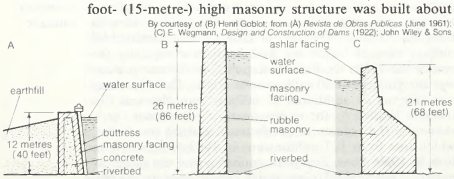


Figure 36: *Cross sections of early dams.*
(A) Proserpina, built by the Romans in southwestern Spain about AD 100. (B) Kebar, in Iran, constructed about AD 1300. (C) Almanza, in Spain, completed about AD 1586.

The Assyrians, Babylonians, and Persians built dams between 700 and 250 BC for water supply and irrigation. Contemporary with these was the earthen Ma'rib Dam in South Arabia, 50 feet high and nearly 1,970 feet long. Flanked by spillways, this dam delivered water to a system of irrigation canals for more than 1,000 years. Other dams were built in this period in Ceylon (modern Sri Lanka), India, and China.

The Romans. Despite their undoubted skill as civil engineers, especially in the field of water supply, the Romans' role in the evolution of dams is not remarkable for quantity or for advances in height. Their skill lay in the comprehensive collection and storage of water and in its transport and distribution by aqueducts. Remarkably, at least two Roman dams in southwestern Spain, Proserpina and Cornalbo, are still in use, although a third, the Alcantarilla Dam, has overturned, and the reservoirs of some others have filled with silt. The Proserpina Dam, 40 feet high, has a masonry-faced core wall of concrete backed by earth; it may be regarded therefore as a forerunner of the modern earthen dam. The Proserpina is strengthened on the upstream face by buttresses (Figure 36A). Of similar construction, 46 feet high and 1,804 feet in length, Alcantarilla Dam was supported by a great weight of earth, which eventually caused failure of the wall. Cornalbo represented a further advance in design; the masonry wall was constructed of cells, which were filled with stones or clay, and faced with mortar. It differs from Proserpina and Alcantarilla in having a sloping upstream face and in being straight in plan. Proserpina and Alcantarilla were polygonal in plan. The merit of curving a dam upstream was not apparently fully appreciated by the Romans until such a curved structure, the forerunner of the modern "arch-gravity" dam, was built in AD 550 at Dâra on the present Turkish-Syrian border by Byzantine engineers.

Proserpina Dam

Early dams in the Orient. Dam construction evolved quite independently in East Asia. In 240 BC a stone crib was built across the Gukow River in China; this structure was 98 feet high and about 985 feet long. Many earthen dams of moderate height (in some cases, of great length) were built by the Sinhalese in Ceylon after the 5th century BC to form reservoirs or tanks for extensive irrigation works. The Kalabalala Tank (formed by an earthen dam 79 feet high and nearly 3.75 miles in length) had a perimeter of 37 miles and helped store monsoon rainfall for irrigating the country around the ancient capital of

The Kalabalala Tank

Anuradhapura. Many of these tanks in Ceylon are still in use today.

In Japan the Diamonike Dam reached a height of 105 feet in AD 1128. Numerous dams were also constructed in India and Pakistan. In India a design employing hewn stone to face the steeply sloping sides of earthen dams evolved, reaching a climax in the 10-mile-long Veeranam Dam, Tamil Nadu, built from AD 1011 to 1037.

In Iran the Kebar, a pioneer arch dam, was built early in the 14th century (Figure 36B). Spanning a narrow limestone gorge, it reached 26 metres high with a thickness of less than 5 metres. The central curved portion, 38 metres in length and radius, was supported on two straight abutments.

FORERUNNERS OF THE MODERN DAM

The 15th to 18th century. In the 15th and 16th centuries, dam construction resumed in Italy and, on a larger scale, in Spain, where Roman and Moorish influence was still felt. Of these dams, the Tibi (1579–89) was an arch-gravity structure 138 feet high; this height was not surpassed in western Europe until the building of the Gouffre d'Enfer Dam in France almost three centuries later. Almanza Dam in Spain is illustrated in Figure 36C. An attempt to build a dam 170 feet high near Lorca, Spain, at the end of the 18th century failed disastrously in 1802, when earth and gravel below the piled structure washed out. In Europe, where rainfall is ample and well distributed throughout the year, dam construction before the Industrial Revolution was on only a modest scale and was restricted to forming water reservoirs for towns, driving water mills, and making up water losses in navigation canals. An exception was the 115-foot-high earthen dam built in 1675 at Saint-Ferréol, near Toulouse, Fr., to supply the Canal du Midi; for more than 150 years it was the highest earthen dam in existence.

19th century. Up to the middle of the 19th century, dam design was entirely empirical. Knowledge of the properties of materials and structural theory had been accumulating for 250 years; Galileo, Sir Isaac Newton, Gottfried Wilhelm Leibniz, Robert Hooke, Daniel Bernoulli, Leonhard Euler, and Charles-Augustin de Coulomb had made outstanding contributions. In the 1850s William John Macquorn Rankine, professor of civil engineering at Glasgow University, successfully demonstrated how applied science could help the practical engineer. Rankine's work on the stability of loose earth, for example, provided a better understanding of the principles of dam design and performance of structures. Furthermore, in certain countries Rankine's work encouraged acceptance of civil engineering as a subject for university study and added to the status of civil engineers. Much remained to be learned of soils and natural rocks in the 100 years after Rankine. Many scientists and engineers made, and continue to make, noteworthy contributions.

Rankine's contributions

Development of modern structural theory. Concrete dam design is based on conventional structural theory. In this relationship two phases may be recognized. The first, extending from 1853 until about 1910 and represented by the contributions of a number of French and British engineers, was actively concerned with the precise profile of gravity dams in which the horizontal thrust of water in a reservoir was resisted by the weight of the dam itself and the inclined reaction of the dam's foundation. Starting about 1910, however, engineers began to recognize that concrete dams were monolithic three-dimensional structures in which the distribution of stress and the deflections of individual points depended on stresses and deflections of many other points in the structure. Movements at one point had to be compatible with movements at all others. Owing to the complexity of the stress pattern, model techniques were gradually employed. Models were built in plasticine, rubber, plaster, and finely graded concrete. In recent years the digital computer has facilitated use of the analytic method of finite elements, by which a monolithic structure is divided into an assembly of separate blocks. Study of both physical and digital models permits deflections of a dam's foundations and structure to be taken into account.

During the 100 years up to the end of World War II, experience in design and construction of dams advanced in many directions. In the first decade of this century many large dams were built in the United States and western Europe. In succeeding decades, particularly during the war years, many impressive structures were built in the United States by federal government agencies and private power companies. Hoover Dam, built in 1936, is an outstanding example of an arch-gravity dam built in a narrow gorge across a major river and employing advanced design principles. It has a height of 726 feet from its foundations, a crest length of 1,244 feet, and reservoir capacity of 48×10^9 cubic yards (37×10^9 cubic metres).

Among earthen dams, Fort Peck Dam, completed in 1940, contained the greatest volume of fill, 126 million cubic yards (96 million cubic metres). This volume was not exceeded until the completion in 1975 of Tarbela Dam in Pakistan (190 million cubic yards).

The modern dam

BASIC PROBLEMS IN DAM DESIGN

Most modern dams continue to be of two basic types: masonry (concrete) and embankment (earth or rock fill). Masonry dams are typically used to block streams running through narrow gorges, as in mountainous terrain; though such dams may be very high, the total amount of material required is limited. Embankment dams are preferred to control broad streams, where only a very large barrier, requiring a great volume of material, will suffice. The choice of masonry or embankment and the precise design depend on the geology and configuration of the site, the functions of the dam, and cost factors.

Site investigation and testing. Investigation of a site for a dam includes sinking trial borings to determine the strata. The borings are supplemented by shafts and tunnels which, because of their cost, must be used as sparingly as possible. In the shafts and tunnels, tests can be made to measure strength, elasticity, permeability, and prevailing stresses in strata, with particular attention given to the properties of thin partings, or walls, between the more massive beds. The presence in groundwater of chemical solutions harmful to the materials to be used in the construction of the dam must be assessed. Sources of construction materials need exploration. As dams continue to increase in height, the study of foundation conditions becomes increasingly critical.

Model tests play a major part in the structural, seismic, and hydraulic design of dams. Structural models are particularly useful in analysis of arch dams and in verifying analytical stress calculations. Various materials have been used for model tests; on some early tests for Hoover Dam, rubber was employed. The need for accurate reproduction of stress patterns in complex models is met by using material of low elasticity. In a sense, dams themselves are models for future design. The instruments built into them to record movements under load, strains within materials after construction, temperature and pressure changes, and other factors are installed primarily to study the performance of the structure and to warn of possible emergencies, but their value in confirming design assumptions is important.

The digital computer has permitted considerable advance in analytic methods of design. Its ability to handle a great volume of data and to solve large sets of simultaneous equations containing many variables has made practicable the finite-element method. In this method a complicated structure is divided into a number of separate equilibrium conditions, and strains are rendered compatible, thus leading to a complete analysis of stress and strain distribution throughout the structure.

Finite-element method

Problems of materials. Each of the two basic dam materials, concrete and earth or rock fill, has a weakness that must be overcome by the proper design of the dam.

Weaknesses of concrete. Concrete is weak in tensile strength; that is, it can be pulled apart easily. Concrete dams must therefore be designed to place minimum tensile strain on the dam and to make use of concrete's great compressive strength, or ability to support vertical loads.

The chief constituent of concrete, cement, shrinks as it sets and hardens, because of water absorption in the crystalline structure, evaporation of water to the atmosphere, and cooling from the higher temperatures reached when the chemical reactions in the cement are in progress during hydration. Because of the large volume of concrete in a dam, shrinkage presents a serious cracking hazard.

Various expedients are used to overcome the problem. Concrete is usually cast in separate blocks of limited height. Gaps may be left to permit heat losses and filled in later. Low-heat cements may be used; these are specially blended so that rates of heat evolution are retarded. Cement content can be safely reduced in the interior concrete in the dam, in which strength and resistance to climatic and chemical deterioration are less important. The cement content, and therefore the heat caused by hydrating, can also be reduced by using aggregate (the other major constituent of concrete) of larger stones. Another expedient is to use other fine-grained materials, such as fly ash (pulverized fuel), as filler, reducing the total cement volume in the concrete. Another is to use certain additives, surface-active agents, and air-entraining agents that permit using a lower water-to-cement ratio in mixing the concrete. Techniques used to speed the cooling process include replacing some of the water in the mix by ice, circulating water through pipes laid in the concrete, and extracting excess water from surfaces by vacuum.

Weaknesses of earth and rock fill. Soils and rock fragments lack the strength of concrete, are much more permeable, and possess less resistance to deterioration and disturbance by flowing water. These disadvantages are compensated for by a much lower cost and by the ability of earth fill to adapt to deformation caused by movements in the dam foundation. This assumes, of course, sufficient usable soil available close to the dam site. In bare mountain country it may be necessary to quarry rock and construct a rock-fill rather than an earth-fill dam. Earth fill is of course more economical, and often a suitable borrow area can be found close to the site.

Seepage in earthen dams

Soil consists of solid particles with water and air in between. When the soil is compressed by loading, as occurs in dam construction, some drainage of air and water takes place, causing an increase in pressures between the solid particles. When there is a high rate of seepage, the soil tends to develop differential pressures and reach a condition called quick, in which it behaves as a fluid. Even if it does not reach this condition, there is often some weakening of its structure, and steps must be taken to counter this.

The earthquake problem. Many large dams have been built in the seismically active regions of the world, including Japan, the western United States, New Zealand, the Himalayas, and the Middle East. In 1968 the Tokachi earthquake damaged 93 dams in Honshu, the main Japanese island; all were embankment dams of relatively small height.

Despite a great deal of work on the distribution of seismic activity, the measurement of strong ground motions, and the response of dams to such motions, earthquake design of dams remains imprecise. The characteristics of strong ground motions at a given site cannot be predicted, and all types of dams possess some degree of freedom, imperfect elasticity, and imprecise damping. Nevertheless, the digital computer and model testing have given promise of considerable progress. It is now possible to calculate the response of a concrete dam to any specified ground motion; this has been done for the Tang-e Soleyman Dam in Iran and the Gariep Dam in South Africa.

There has also been considerable advance in the theoretical estimation of the effects of ground motion on embankment dams.

TYPES OF DAMS

The modern concrete dam. *Concrete gravity dams.* Concrete gravity dams share certain features with all types of concrete dams. Running in virtually a straight line across a broad valley, they resist the horizontal thrust of the retained water entirely by their own weight; at each level in their height the water's thrust is deflected down toward the foundation by the weight of the concrete. In this action their purpose resembles that of the abutment of an arched bridge or the buttresses and pinnacles of a church. A gravity dam is a right-angled triangle; its hypotenuse forms the sloping downstream face. The base width is approximately three-quarters the height of the dam.

The three main forces acting on a gravity dam are the thrust of the water, the weight of the dam, and the pressure, or reaction exerted by the foundation, which is necessarily inclined in respect to the superstructure. It is also essential to consider the thrust exerted on the upstream face by silt deposited in the reservoir or by ice on the water surface, the inertial forces that can be caused by seismic action, and, in particular, the buoyant uplift force of water seeping under the dam or into the horizontal joints.

Uplift from seepage has caused sustained discussion among engineers. It calls for the greatest of care in design and construction. Where a dam is founded on solid rock, a simple downward projection of concrete into the rock will generally suffice to cut off seepage and eliminate uplift pressures. Usually, however, the rock foundation is permeable, sometimes to considerable depths, so construction of an absolutely reliable cutoff is either difficult or impossible. Reliance must then be placed on an extensive system of grouting the fissured rock and on relieving uplift pressures by means of drainage. Many dams possess both cutoffs and underdrainage.

Post-tension construction

A relatively new development in the construction of gravity dams is incorporation of posttensioned steel into the structure. This helped reduce the cross section of Allt Na Lairige Dam in Scotland to only 60 percent of that of a conventional gravity dam of the same height. A series of vertical steel rods near the upstream water face, stressed by jacks and securely anchored into the rock foundation, resists the overturning tendency of this more slender section. This system has also been used to raise existing gravity dams to a higher crest level, economically increasing the storage capacity of a reservoir.

Of special interest are three concrete gravity dams, all of which feature a straight sloping downstream face. Bratsk, built across the Angara River at Irkutsk in Russia and completed in 1964, stands 410 feet above foundation level and, excluding the earthen side dams, is nearly 5,000 feet in length; it contains 5,900,000 cubic yards of concrete. Grand Coulee Dam, completed in 1941, was built across the Columbia River in Washington state, U.S.; its main structure is 550 feet high and 5,223 feet (1,592 metres) long and contains almost 12,000,000 cubic yards (9,000,000 cubic metres) of concrete. Grande Dixence Dam in Switzerland, completed in 1962 across the narrower valley of the Dixence, has a crest length of 2,296 feet and contains approximately 7,790,000 cubic yards of concrete; at 935 feet it was the highest dam in the world until the Nurek Dam was completed in 1980. By comparison, the Great Pyramid of Giza contains 3,400,000 cubic yards of masonry.

Concrete buttress and multiple-arch dams. Unlike gravity dams, buttress dams do not rely entirely upon their own weight to resist the thrust of the water. Their upstream face, therefore, is not vertical but inclines about 25° to 45°, so the thrust of the water on the upstream face inclines toward the foundation. Embryonic buttresses existed in some Roman dams built in Spain, among them the Proserpina. As technology advanced, dams with thin buttresses of reinforced concrete supporting inclined panels of similar construction were built. In today's buttress dams, less account is taken of effecting maximum economy in the use of concrete. The trend is to reduce the area of costly formwork necessary and to avoid use of steel reinforcement. With greater heights, modern buttress dams are inevitably less slender.

Design of a multiple-arch dam

Several variations are possible in the design of the junction between the buttresses at the water face. Where no relative movement in the buttress foundations is anticipated, the design can link individual buttress heads rigidly, by means of arches, to form a multiple-arch dam. A recent Canadian example of this type is the 703-foot-high multiple-arch Daniel Johnson Dam on the Manicouagan River in Quebec. The dam uses a total of 14 buttresses

in its crest length of 4,297 feet; two very much larger buttresses support the structure over the original riverbed.

Where buttress foundations might yield, the design must allow some freedom of movement between the heads of the buttresses. This is normally achieved by enlarging the heads until they are almost in contact and then joining them with flexible seals. Thus joined, the heads present a solid face to the water. Such a design was used in the construction in the Farahnaz Pahlavi Dam in Iran. Built for the Tehrān Regional Water Board, this dam has a maximum height of 351 feet and a crest length of nearly 1,181 feet.

A comparison between the Daniel Johnson multiple-arch dam and the Farahnaz Pahlavi buttress dam shows that the buttresses have to be placed much closer together than is necessary with a multiple-arch dam. This allows each buttress to be more slender, however, and spreads the load more easily over the foundation. The detailed design at the bottom of the Farahnaz Pahlavi buttresses was necessitated by weak foundation conditions at the site and by the need to limit the length of each buttress to reduce its response to seismic action. By contrast, the Daniel Johnson buttresses could be founded individually, exploiting fully an important advantage of buttress dams over gravity dams—that of smaller uplift forces.

Arch dams. The advantages of building a dam curved in plan, utilizing the water pressure to keep the joints in the masonry closed, was appreciated as early as Roman times. An arch dam is a structure curving upstream, where the water thrust is transferred either directly to the valley sides or indirectly through concrete abutments. Theoretically, the ideal constant angle arch in a V-shaped valley has a central angle of 133° of curvature. This leads to the development of the cupola (or variable radius) arch dam with the crest portion overhanging downstream (Figure 37). The constant radius arch dam generally has a vertical upstream face. There are many other factors to take into account, however, including fixity at the abutments at the upper levels and the vertical cantilever effect of the arch at the riverbed level.

From W. Creager, J. Justin, and J. Hinds, *Engineering for Dams* (1945); John Wiley & Sons, Inc.

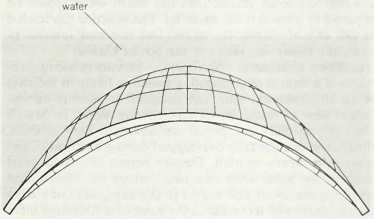


Figure 37: Plan view of an arch dam showing double curvature.

An arch dam is therefore a shell structure, admittedly sometimes of significant thickness, that owes its strength essentially to its curved profile but that is supported at the riverbed and up the valley sides by constraints that cause both flexure and shear on the membrane. Dependent for its strength upon effective support at its abutments, its very strength and rigidity make it sensitive to movements at the abutments. Only favourable sites providing sound rock are suitable for arch dams.

Vaiont Dam disaster

The great reserves of strength inherent in an arch dam were dramatically displayed in 1963 when the reservoir behind Vaiont Dam in Italy was virtually destroyed by a landslide. Vaiont, at that time the second highest dam in the world, was built across a narrow gorge on limestone foundations so that the crest 858 feet above the valley bottom was only 623 feet in length. Some large-scale instability in the mountainside above the reservoir had been observed earlier by the engineers during filling; they were allowed to proceed very slowly, and three years later, on Oct. 9, 1963, with filling still incomplete, about 314 million cubic yards of soil and rock slid down into the reservoir, sending a tremendous volume of water to a height of 853 feet on the opposite side of the valley. The flood overtopped the dam to a depth of 328 feet and surged down the valley, causing a major tragedy, the destruction of several villages with a large loss of life. Yet, only superficial damage was caused to the dam, which, at its crest, is about 11.2 feet thick.

Embankment dams. *General characteristics.* Earlier embankments were undoubtedly built as simple homogeneous structures, with the same material used throughout. No effort was made at first to subdivide the dam into separate zones with the best-suited material in each zone. The homogeneous dam nicely illustrates the general behaviour of an embankment dam and demonstrates the reasons for the rather baffling pattern of heterogeneous dam profiles employed.

Like a concrete gravity dam, the weight of an embankment dam deflects the horizontal thrust of the water pressure down to the foundation. The resultant pressures on the foundation must not cause excessive deformation or collapse.

Water penetration

Unlike concrete, embankment dam materials possess only limited resistance to water penetration. The rate of penetration depends on the pressures exerted by the water in the reservoir, the length of seepage paths through the dam, and the permeability of the material of construction. Soils and rock range from substantially impermeable clays through silts and sands to coarse-graded gravels and rock fragments that possess little resistance to the movement of water. The range is extremely wide; the seepage rate through clean gravel is 10,000 times that through sand, 10,000,000 times that through silt, and 100,000,000 times that through dense clay.

An embankment dam must be stable in itself. Its side slopes must not slip or slide; liquefaction of the soils must not occur; erosion of the soils—as a result of water overtopping the crest, wave action on the upstream face, or seepage washing out the fine material through the coarser—must be avoided. As with a concrete dam, seepage of water from the reservoir through the foundation and under the actual embankment also must be controlled.

Potential weakness. There are three parts of a dam where weakening of the soil structure and liquefactions can occur. In Figure 38A the pattern of seepage through a homogeneously filled dam is shown. Near the downstream toe, the gradient of the pore-water pressures is steep, and constraints holding the soil structure together are low; this is one area of weakness in an embankment dam. One solution is to introduce drainage as indicated in Figure 38B, where the area of steep seepage gradients has been moved to where the soil is constrained near the centre of the dam. Seepage gradients at the vulnerable downstream toe are eliminated.

A second area of potential weakness is the upstream face, when the water in the reservoir is rapidly drawn down. If the pore-water pressures cannot adjust themselves fast enough to this change in the free-water surface in the reservoir, severe seepage gradients begin; these can cause failure. A zone of freely draining fill of coarser grading can be placed on the upstream face to counter this.

Water seepage from the reservoir through the foundations under the dam is another potential weakness. Owing to their great widths, embankment dams can be constructed on unfavourable sites, such as open-jointed rock

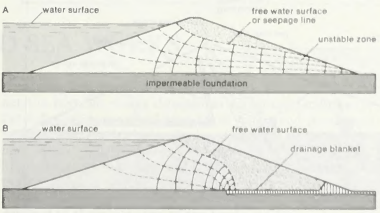


Figure 38: *Paths of seepage through embankment dams.*
(A) Homogeneous dam. (B) Dam with a drainage blanket.

or weaker and possibly locally permeable clay. It is necessary, however, either to check or to drain away harmlessly the seepage water that would otherwise weaken the downstream parts of the dam and, in extreme cases, cause it to fail. Several countermeasures, possibly in combination, can be employed: the foundation can be grouted or a cut-off trench excavated and backfilled with an impermeable material; a drainage blanket can be constructed at the base of the downstream part of the dam, or individual drainage wells or galleries can be excavated; the length of the seepage paths under the dam can be extended by means of an impermeable blanket laid on the upstream side of the dam; or additional free-draining fill can be placed at the downstream toe of the dam.

Construction techniques. Today all large embankment dams have a core of lower permeability built near their centre. Suitable materials, such as a plastic clay, are weaker than more permeable soil. The width of the core is restricted to that necessary to lower sufficiently the pore pressures in the downstream part of the dam. Although the top of the core must be at the crest of the dam, the core itself need not be vertical. On some rock-fill dams the core can slope forward to an extreme position where it lies on the upstream face. Usually a sloping core occupies an intermediate position so it can be constructed on a sloping face of a partially built dam.

Where seepage is inevitable, the use of finely graded core material in proximity to coarser material is avoided. Bands of intermediately graded material must be inserted to prevent the finely graded material from leaching through the coarse zones. Filter zones are graded so each band is four to five times coarser than the preceding band.

Aswān High Dam

Figure 39A shows a typical section of the Aswān High Dam, an embankment 365 feet high built of dune sand and rock fill on a very permeable foundation of deep alluvium. Here the central clay core is vertical; this barrier to seepage is extended to the original riverbed as grouted sand and below the riverbed to a depth of 740 feet as a grout curtain. A corrugated blanket of clay extends upstream within the dam from the base of the core. Within the upstream and downstream cofferdams, partly of rock fill, much of the filling is of compacted sand. Filter layers separate the cofferdam filler from the outer layers of freely draining rock fill. Drainage wells will be observed below the downstream toe. The early stages of construction were carried out under deep water; hence the grouted coarse sand between the clay core and the grout curtain.

The rather simpler section of Oroville Dam, located in California, U.S., is shown in Figure 39B. Until the 1,040-foot-high Nurek Dam in the Soviet Union was completed, Oroville Dam (774 feet) was the highest embankment dam in the world. Unlike the Aswān High Dam, Oroville was not built on deep permeable alluvium, nor was it necessary to place part of the fill under water. One unusual feature is the concrete block at the base of the sloping core designed to fill in the incised gorge of the Feather River Canyon. The grout curtain, compared with that of the Aswān High Dam, is of nominal depth. On each side of the sloping core, transition zones separate the core from the main mass of more pervious filling. The downstream transition zone is backed by a curtain drain of selected pervious material connected to a drainage blanket on the downstream side. The upstream face of the dam is protected against wave action by a three-foot layer of broken stone (riprap).

From (A) 9th International Congress on Large Dams; (B) *Civil Engineering*, A.S.C.E. (June 1963)

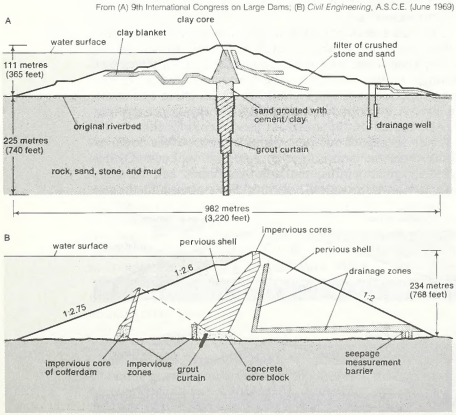


Figure 39: Sections of (A) the Aswān High Dam, Egypt, and (B) Oroville Dam, California, U.S.

Efficient compacting of soils requires maximum density of dry particles consistent with an economic number of passes of the compacting plant. The process of compacting a soil by kneading it involves expelling as much of the air as practicable; water content is not normally much reduced. The optimum water content for maximum dry density—which results in maximum strength—can be achieved for a given amount of work done on the soil in compaction. In arid climates, water must often be added to excavated soils. In temperate climates, however, water content is usually too high, except in deeply excavated and well-drained soils.

Normally, soils are placed in embankment dams in thin layers individually compacted by rolling. Finer soils, such as those used in cores, may be harrowed before rolling. Coarser soils, including rock fragments, are compacted by vibration and rolled. Coarse rock fragments (rock fill) are compacted to a limited extent by impact on being dumped from the construction plant; compaction of smaller fragments is assisted by sluicing with water.

Hydraulic filling

In the process of hydraulic filling, sands are dredged from borrow pits, transported in water by pipelines to the filling area, and deposited there by draining off the surplus water. Hydraulic filling is widely practiced in maritime works, if sand is the only construction material available. It has also been used in the construction of embankment dams, although on some inland sites too much water would be required to transport the material. The practice has tended to fall out of favour for dams, but renewed interest in hydraulic filling was taken in the Soviet Union.

Auxiliary structures. *Spillways.* Serious consequences follow if a dam is overtopped. Disaster is likely in the case of an embankment dam not designed to permit uncontrolled flow of water on its downstream slope. In March 1960 the partially completed embankment dam at Orós, Brazil, was accidentally overtopped during a period of unexpectedly heavy rainfall. Despite heroic efforts to avert disaster, the water level rose nearly three feet above crest level, eroded about half the fill in the dam, and cut a deep breach about 660 feet wide in the structure. Although there was time to evacuate 100,000 people living downstream, half were subsequently rendered homeless and about 50 perished. Spilling over a concrete gravity dam is also serious, because the floodwater erodes the foundations at the downstream toe. Arch dams possess greater resistance to failure after overtopping.

Flood hydrology is a difficult subject. Much effort is being made to establish relationships between rainfall and river discharge, both in quantity and in time lag. Such statistical methods cannot estimate the maximum possible flood. At best they indicate only the probability of a specified flow being exceeded in a particular period. In constructing the Kariba Dam on the Zambezi River, analyses of the available records of river discharge yielded the estimate that a flood of 9,950 cubic yards per second should be expected once in four years. During the first year of construction on the riverbed, a flood of 11,100 cubic yards per second was experienced, and in the second year the Zambezi discharged 21,200 cubic yards per second.

In these circumstances civil engineers attach much importance to the design of spillways on dams. Inadequate spillway capacity caused failure of many older earthen dams built before modern flood data became available.

Four general aspects of spillways should be emphasized.

First, the uncontrolled discharge of surplus water past the dam should be automatic and, like a safety valve on a steam boiler, not under human control. Second, the spillway intake should be wide enough so that the largest floods can pass without increasing the water level in the reservoir enough to cause a nuisance to riparian inhabitants and property upstream. Third, the rate of floodwater discharge should not increase much above that experienced before the construction of the dam. An increase creates a flood nuisance downstream. A dam usually reduces the peak discharge rate, owing to the lag effect caused by a flood passing through the reservoir. Fourth, floodwater discharged over the height of a dam can be destructive to the dam structure itself and to the riverbed unless its energy is controlled and dissipated in harmless turbulence.

With embankment dams, a separate spillway structure is normally constructed to one side of the dam itself. With concrete gravity dams the sloping downstream face of the structure serves as the spillway. The water travels at very high speeds (about 100 miles per hour in the case of a dam 330 feet high) and forms a standing wave where it enters the riverbed; it proceeds downstream at lower mean velocity but in a highly turbulent state. Grand Coulee Dam utilizes a spillway of this type. An obstruction known as a kicker, placed at the toe of the dam to project the water slightly upward, can move further downstream the area in which erosion of the riverbed is most intense. With higher dams, it is possible to deflect the jet of spilling water from a level above the base of the dam; this is known as a ski-jump spillway.

Ski-jump spillway

Spillways need not be open to the atmosphere. Shaft and tunnel spillways can destroy the energy of the water at a predetermined point downstream of the dam. At the upstream end, the intake can be self-priming siphons or bell-mouthed drop shafts; the latter are also known as morning-glory spillways.

With arch dams it is convenient to construct gated openings in the shell structure at some distance below the crest of the dam, ensuring that the discharging jets fall well clear downstream. A line of six such gates is used in the design of Kariba Dam.

Spillways constructed to one side of earthen dams are featured in the design of Oroville Dam and of Mangla in Pakistan. The spillway at Mangla discharges 36,600 cubic yards of water per second; the upper stilling basin has the dimensions of an Olympic Games stadium with its grandstands.

Rock-fill dams, specially designed to be overtopped in times of flooding, are a new development. The first such permanent dams have been constructed in Australia. For temporary works, the technique has been used on the Blue Nile.

Gates, sluices, and outlets. In addition to spillways, openings through dams are also required for drawing off water for irrigation and water supply, for ensuring a minimum flow in the river for riparian interests downstream, for generating power, and for evacuating water and silt from the reservoir. These gated openings normally are fitted with coarse screens at the upstream ends to prevent entry of floating and submerged debris. Provision for cleaning these screens is essential.

Several forms of gates have been developed. The simplest and oldest form is a vertical-lift gate that, sliding or rolling against guides, can be raised to allow water to flow underneath. Radial or tainter gates are similar in principle but are curved in vertical section better to resist water pressure. Tilting gates consist of flaps held by hinges along their lower edges that permit water to flow over the top when they are lowered.

Drum gates can control the reservoir level upstream to precise levels automatically and without the assistance of mechanical power. One drum gate design consists of a shaped-steel caisson held in position by hinges mounted on the crest of the dam and supported in a flotation chamber constructed immediately downstream of the crest. Water pressure in the reservoir and buoyancy of the caisson in the flotation chamber hold the caisson in rotational equilibrium. Raising or lowering the water level in the flotation chamber causes the caisson to rotate in the same direction, thus reducing or increasing flow from the reservoir over the gate. This action can be linked to and operated automatically by a float control device in the reservoir. Two drum gates are installed at Pitlochry Dam in Scotland.

Reservoirs. Modern engineers have learned the value of giving attention early to potential problems in reservoir maintenance. Sediment in rivers seriously influences the effective life of a reservoir and therefore the financing of a dam. Some modern dams have been rendered useless for storing water because the reservoir has filled with silt. In many others, effective storage capacity has been seriously reduced. At the Nile barrages, the heavy silt-laden floodwater is allowed to pass through the sluices and only the cleaner water at the end of the flood season is stored.

Fish passes. For many years hydroelectric dam design has taken into account the need to conserve certain species of migratory fish. Success has been achieved with salmon in Scotland and on certain rivers in the United States and Canada. Notable examples of conservation measures are to be found at Bonneville, Priest Rapids, and Wanapum dams and at many dams in Scotland.

Adult salmon striving to reach their spawning grounds upstream must be prevented by screens from entering the turbine tailraces at power stations and induced instead to enter a fish pass that allows them to surmount the dam. Similarly, young salmon must be allowed to pass a dam safely on their journey downstream to feeding grounds in the ocean. Young salmon are remarkably insensitive to sudden changes of pressure and have been known to pass safely through turbines operating at heads of up to 160 feet. Nevertheless, it is preferable to induce them to use the fish passes.

Fish passes usually take the form of fish ladders and fish locks. A fish ladder is utilized at Pitlochry Dam in Scotland; it consists of a series of stepped pools through which water is continuously discharged during the migratory seasons. The individual pools may be separated by a series of low weirs or linked by short inclined underwater pipes to provide the necessary steps of one to two feet in water levels. Sometimes both weirs and pipes are provided.

Fish ladder

The Borland fish lock was developed in Scotland as an alternative to fish ladders. It operates on the same intermittent principle as a ship lock but is constructed as a closed conduit. Intermittent closure of the gates at the bottom causes the continuous flow through the lock to fill the conduit at intervals and thus allows fish waiting in the bottom chamber to be raised through the height of the dam. The lock also serves at other seasons to flush young salmon down past the dam. (J.G.B.)

HARBOURS AND SEA WORKS

The construction of harbours and sea works offers some of the most unusual problems and challenges in civil engineering. The continuous and immediate presence of the sea provides the engineer with an adversary certain to discover any weakness in the structure built to resist it.

Principles of maritime engineering

Objectives. The principal objectives of such works fall broadly into two classifications: improvement of transportation, and reclamation and conservancy of land. Under the first fall works directed at providing facilities for the safe and economical transfer of cargo and passengers between land vehicles and ships; fishing ports for the landing and distribution of the harvest of the sea; harbours of refuge for ships and small craft; and marinas for the mooring or laying up of small private craft. Under the heading of reclamation and conservancy come works directed to the protection of the land area from encroachment by the sea, to the recovery and conversion to land

use of areas occupied by the sea, and to the maintenance of river estuaries as efficient means for the discharge of inland runoff. In many places, without continuous attention to such maintenance, the coincidence of high tides with heavy rainfall would lead to frequent disastrous flooding of inhabited areas.

The civil engineering techniques used for either of these objectives are broadly similar, and indeed the realization of both objectives at the same time will frequently be a feature of the same project. An operation of maintaining a river estuary at a depth sufficient for navigation, for example, may at the same time greatly improve its capacity for the drainage of upland floodwaters.

Hydraulic models. The planning of maritime civil engineering works, whether for transportation, reclamation, or conservancy, has been facilitated by the development of the technique of model studies. Once regarded as scientific toys, such studies are now considered an essential preliminary step to any large-scale redevelopment of a port or coastal area and are useful even for minor modifications or additions.

Scale models of the area, harbour, or estuary are made so that water can be caused to flow in such a way as to reproduce the various tidal and other streams in the same direction and with velocities equivalent to those occurring on the site. A variety of devices, usually electronically controlled, have been developed to produce both wave and tidal effects.

By courtesy of the British Transport Docks Board

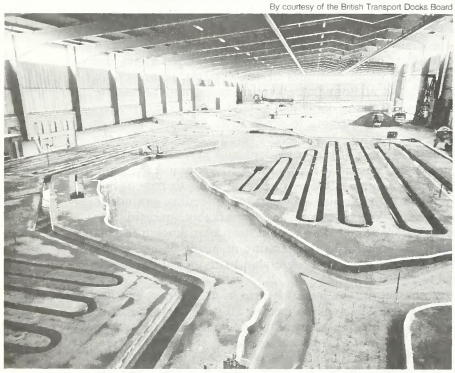


Figure 40: Tidal model of the Humber estuary in England.

Predicting long-range tide effects

The value of these experiments derives from the reduction in the time scale, which has been found to correspond to the reduction in the dimensional scales of the model. Thus, the large model of the Clyde estuary of Scotland works on a tidal cycle of about 14 minutes, or about 50 times the actual frequency. The effect of three years of tides following any modification of the profile of the harbour can thus be studied on the model in a matter of three weeks, and any tendency to otherwise unanticipated scour (clearing by powerful current) or siltation can probably be detected. The relative values of alternative positions of breakwaters in affording shelter can be similarly studied using the wave-generating devices available; and the development of secondary, or reflected, waves with undesirable disturbances within the sheltered area may be anticipated and, if possible, forestalled.

Natural and artificial harbours. In certain favoured points on the world's coastlines, nature has provided harbours waiting only to be used, such as New York Bay, which the explorer Giovanni da Verrazano described as "a very agreeable location" for sheltering a ship. Such inlets, bays, and estuaries may require improvement by dredging and must be supplied with port structures, but basically they remain as nature made them, and their existence accounts for many of the world's great cities. Because such natural harbours are not always at hand where port facilities are needed, engineers must create artificial harbours. The basic structure involved in the creation of an artificial harbour is a breakwater, sometimes called a jetty, or mole, the function of which is to provide calm water inshore. Locations for artificial harbours are of course chosen with an eye to the existing potential of the coast; an indentation, however slight, is favoured. Yet it has often been found justifiable on economic or strategic grounds to construct a complete harbour on a relatively unsheltered coastline by enclosing an area with breakwaters built from the shore, with openings of minimum width for entry and exit of ships.

Sea works for transportation

CLASSICAL HARBOUR WORKS

Improvements to natural harbours and construction of artificial harbours were undertaken in very ancient times. There is no conclusive evidence for the date or locality of the first artificial harbour construction, but it is known that the Phoenicians built harbours at Sidon and Tyre in the 13th century BC.

The engineers of those days either knew or thought little about conservancy even as applied to the ports they constructed. Evidence is to be seen in the once thriving ports around the shores of the Mediterranean that now are not merely silent ruins but seem so far from even sight of the sea that it is difficult to imagine the presence of seagoing ships at the wharves, the alignment of which can occasionally be traced in the fertile alluvial land now occupying the site. Ephesus, Priene, and Miletus, on the Aegean shores of Asia Minor, are examples of this type of harbour disappearance, the destructive agent in each of these cases being the picturesque Meander (now the Menderes) River, whose creation of new land from the sea is readily perceivable from high ground adjacent to the river mouth. The formation of further bars is proceeding visibly—and, as there is currently no port in the vicinity whose livelihood can be threatened, it is interesting to speculate how far out to sea this process will ultimately continue in the course of the next millennium or so.

Disappearance of ancient harbours

At Side, facing the island of Cyprus, the remains of an ancient breakwater, built to protect the anchorage, can still be seen, but the area enclosed between it and the advancing shoreline is now not a stone's throw wide. In this case, not only the river in the vicinity but also littoral drift, (the movement of sediments by a current parallel to the coast), which produces and maintains extensive beaches to the east and the west, must be held partly responsible for the scale of siltation.

Of many of the ancient port structures, no physical trace remains, but knowledge of the fact that they existed and even a measure of technical description has come down through the written word. With these descriptions and the monuments that still remain, some picture may be formed of the work undertaken by the maritime civil engineers of ancient times.

Given the frailty of the craft for which they were providing, shelter from the weather was the prime consideration; and much effort was devoted to the construction of breakwaters, moles, and similar enclosing structures. Cheap labour was abundant, and the principal material used was natural stone. Surviving structures built in this way are likely to give an appearance of indestructibility, which occasionally attracts favourable comparison with the lighter, more rapidly depreciating modern structures. It is not, however, necessary to credit the engineers of antiquity with a conscious intention to build forever. Given the materials they had to use and the purposes they were implementing, they could do little else; moreover, because there was no rapid pace of advance in the development of ships or land transport, they were undisturbed by the shadow of obsolescence. In the 20th century, far from wanting to build forever, the port engineer has to be careful to avoid saddling posterity with structures that may long outlast their usefulness and turn into liabilities. The modern balance between excessive durability and dangerous frailty is one that the ancients never had to strike.

The durability of ancient sea works

Aided by the characteristics of the material they employed, the ancients constructed maritime works on a scale that is certainly remarkable to this day. Interesting technical practices included the use by the Romans of the semicircular arch in constructing moles or breakwaters, an arrangement that allowed a measure of ingress and egress by the sea to produce a beneficial scouring action in the harbour. The Romans underpinned their structures with timber piling and frequently resorted to the construction of cofferdams (watertight enclosures) that they could dewater by the employment of Archimedean screws and waterwheels. This practice enabled them to carry out much of their foundation work in the dry; and the use of their famous hydraulic cement, pozzolana, gave their structures a durability far exceeding that afforded by the lime cement available to their predecessors.

Among the more interesting harbours of the ancient world are Alexandria, which had on the island of Pharos the first lighthouse in the world; Piraeus, the port of Athens; Ostia, the port of Rome; Syracuse; Carthage, destroyed and rebuilt by the Romans; Rhodes; and Tyre and Sidon, ports of the earliest important navigators, the Phoenicians.

BREAKWATERS

Because the function of breakwaters is to absorb or throw back as completely as possible the energy content of the maximum sea waves assailing the coast, they must be structures of considerable substance. The skill of the designer of a breakwater lies in achieving the minimum initial capital cost without incurring excessive future commitments for maintenance. Some degree of maintenance is of course unavoidable.

Breakwater design. A common breakwater design is based on an inner mound of small rocks or rubble, to provide the basic stability, with an outer covering of larger boulders, or armouring, to protect it from removal by the sea. The design of this outer armouring has fostered considerable ingenuity. The larger the blocks, the less likely they are to be disturbed, but the greater the cost of placing them in position and of restoring them after displacement by sea action. Probably the least satisfactory type of armour block, frequently used because of its relative ease of construction, is the simple concrete cubic, or rectangular, block. Even the densest concrete seldom weighs more than 60 percent of its weight in air when fully immersed in seawater; consequently, such blocks may have to be as much as 30 tons (27,000 kilograms) in weight to resist excessive movement.

Boulders of suitably dense natural rock are generally much more satisfactory and, in a project completed in the United Kingdom in the 1960s, it was found by experiment, and subsequently confirmed in experience, that armouring of this type could be composed of blocks of as little as six to eight tons to resist the action of waves up to 18 feet (5 metres) in height. The same experiments showed that, to afford the same protection in the same circumstances, concrete blocks of 22 tons would have been necessary.

In such cases, an intermediate layer of smaller blocks or boulders is inserted between the armouring and the inner core to prevent the finer material in the core from being dragged out by sea action between the interstices of the armour—a process that leads to ultimate settlement and possible breaching by overtopping of the breakwater.

The increasing cost and frequent unavailability within economic distance of suitable natural rock has provoked considerable thought to the design of concrete armour units that can, by reason of their shape, overcome the disadvantages of the simple cubic, or rectangular, block. One of the most successful has been the tetrapod, a four-legged design, each leg projecting from the centre at an angle of $109\frac{1}{2}^{\circ}$ from each of the other three. Legs are bulbous, or pear-shaped, with the slightly larger diameters at the outer end. These units have the property, when placed, of knitting into each other in such a way that the removal of a single unit without the displacement of several others is almost impossible, while the interstices between them act as an absorbent of wave energy. Weights substantially less than those needed for cubic blocks are adequate in the case of tetrapods in similar storm conditions. The tetrapods can be mass-produced adjacent to the site through the employment of reusable steel forms.

Tetrapods

It is usual to construct some form of roadway along the crest of a breakwater, even when this is not required for any other dockside purposes, to facilitate inspection and access for labour, materials, and equipment for damage repairs.

Solid breakwaters. In certain circumstances, particularly in parts of the world where clear water facilitates operations by divers, vertical breakwaters of solid concrete or masonry construction are sometimes employed. Some preparation of the seabed by the depositing and leveling of a rubble mound to receive the structure is necessary, but it is usual to keep the crest of such a mound sufficiently below the surface of the water to ensure its not becoming exposed to destructive action by breaking waves. Repulsion of the waves by vertical reflection rather than their absorption is the philosophy of protection in all such cases, but it is not possible to state categorically which arrangement produces the most economical structure.

This type of breakwater can be conveniently constructed through the use of prefabricated concrete caissons, built on shore and floated out, sunk into position on the prepared bed, and filled with either concrete or, less frequently, simple rubble or rock filling. A historical example of this arrangement was the Mulberry Harbour, built by the Allies and floated into position for the invasion of Normandy in 1944. No previous preparation of the seabed was possible, and only partial filling of the caissons had been carried out when the progress of the war rendered further operations unnecessary. Nevertheless, the fact that several of the caissons remained in position basically undamaged for nearly a decade after the invasion on this notoriously stormy coast demonstrated the possibilities of the method.

Floating breakwaters. Because of the large quantities of material required and the consequent high cost of breakwaters of normal construction, the possibility of floating breakwaters has received considerable study. The lee of calm water to be found behind a large ship at anchor in the open sea illustrates the principle. The difficulty is that, to resist being torn away in extremes of weather, the moorings for a floating breakwater must be very massive. They are therefore difficult to install and subject to such constant chafing and movement as to require substantial maintenance. Another problem arises, especially in areas of large tidal range. The unavoidable—indeed, essential—slack in the moorings may allow the breakwater to ride large waves, so that they pass underneath it carrying a considerable proportion of their energy into the area to be sheltered.

One approach to the problem is based on the concept of causing the waves to expend their energy at the line of defense by breaking on a large, floating horizontal platform.

Pneumatic breakwaters. Finally, the pneumatic, or diffusion, breakwater has been widely discussed. Experiment and limited experience have shown that a curtain of air bubbles blown up from the seabed through a row of perforated nozzles acts as a barrier to the movement of waves over the surface. The mechanics of the arrangement appear to be that the rising bubbles generate streams flowing on the surface, outward in both directions, and the flow meeting the oncoming waves can be made sufficient to hold them up. There is reason to believe that jets of water would be almost as effective as air. Although the volume of air or water necessary to restrain completely the waves generated in severe weather over a wide front would require installation of a plant of uneconomical size, the device can be useful for the temporary protection of a short length of shore to allow the execution of specific works. The air or water pipes can be laid on the seabed at the perimeter of the area to be protected and fed from a mobile plant on shore, and the whole body of equipment can be removed after the operations have been completed.

DOCKS AND QUAYS

Because the principal operation to which harbour works are dedicated is transfer of goods from one transportation

form to another (*e.g.*, from ships to trucks), it follows that docks, wharves, and quays are the most important assets of a port.

Facilities required in a dock

Ships must lie afloat in complete shelter within reach of mechanical devices for discharging their cargoes. Although in emergencies ships have been beached for unloading purposes, modern vessels, particularly the larger ones, can rarely afford contact with the seabed without risking serious structural strain. The implications of cargo handling, as far as civil engineering works are concerned, do not differ much whether the loading and discharge are effected by shore-based cranes or by the ship's own equipment. In either case, large areas of firm, dry land immediately alongside the ship are required; the engineer must find a way to support this land, plus any superimposed loading it may be required to carry, immediately adjacent to water deep enough to float the largest ship.

The capital cost of such works probably increases roughly in proportion to the cube of the deepest draft of ship capable of being accommodated; thus the economic challenge posed by the increase in the size of modern ships is considerable. The advent of containerization—the packaging of small units of cargo into a single larger one—has not fundamentally altered this problem, except perhaps to reduce the number of separate individual berths required and to increase greatly the area of land associated with each berth. A figure of 20 acres (8 hectares) per berth is freely mentioned as a reasonable requirement. The problem of land support at the waterline remains the same.

Gravity walls. The solution initially favoured, and indeed predominant for many years, was that of the simple gravity retaining wall, capable of holding land and water apart, so to speak, through a combination of its own mass with the passive resistance of the ground forming the seabed immediately in front of it. To ensure adequate support without detrimental settlement of the wall, to ensure its lateral stability, and to prevent problems of scour, it is necessary to carry the foundations of the wall below the seabed level—in some cases a considerable distance below. In earlier constructions, the only guide to this depth in the planning stage was previous knowledge of the ground and the acumen of the engineer in recognizing the characteristics of the ground upon seeing it. Many projects were carried out in open excavation, using temporary cofferdams to keep out the sea. In particularly unfavourable or unstable soils, accidents caused by collapse of the excavation were not unknown.

In modern practice, no such project is initiated without exhaustive exploration of the soil conditions by means of borings and laboratory tests on the samples. Continuous monitoring of the soil conditions during construction is also considered essential. Even so, accidents caused by soil instability still occasionally occur.

Materials for gravity walls

The material composing the walls is today almost universally concrete, plain or reinforced, according to the requirements of the design. This material has entirely superseded the heavy ashlar (natural rock) masonry at one time used for such construction, when the techniques for the large-scale production of concrete were not so well developed as they are today.

In some circumstances, particularly those in which the water is reasonably clear or the design and soil conditions do not require very deep excavation into the seabed, the construction of quay walls is adopted by means of large blocks, sometimes of stone but generally of concrete, placed underwater by divers. The economics of this method of construction are influenced by the high cost of skilled divers and by the cumbersome nature of diving equipment. The development of lightweight, self-contained equipment, which leaves the diver considerably more mobile, may relieve this problem.

Concrete monoliths. The risks and difficulties attendant on the construction of gravity walls have been avoided, in suitable conditions, through the use of concrete monoliths sunk to the required foundation depth, either from the existing ground surface or, where the natural surface slopes, from fill added and dredged from the front of the quay wall on completion. This technique amounts to the construction above the ground of quite large sections of the intended wall, usually about 50 feet square in plan, which are then caused to sink by the removal, through vertical shafts, of the underlying soil. Another lift of wall is then constructed on top of the section that has sunk, more soil is removed, and the process is repeated until the bottom has reached a foundation level appropriate to the required stability. Considerable skill is sometimes necessary in the sinking process to prevent the monoliths (usually provided with a tapered-steel cutting edge to the lowest lift) from listing, an eventuality that can occur if any part of the periphery encounters material that is particularly difficult to penetrate. Differential loading of the high side and special measures to undercut the material composing the obstruction may be necessary.

The shafts through which the excavated material is removed are generally flooded throughout the operation simply from the intrusion of the groundwater; if necessary, this water can be expelled by the use of compressed air. The excavation of difficult material in detail and in the dry can then be undertaken. It is an operation of some delicacy, because the flotation effect of the compressed air adds a further element of instability to the monolith, and a blow (sudden leakage of air) under the cutting edge may result in flooding of the working chamber. When the bottom edge of the monolith has reached the designed level, the excavation shafts are sealed by concrete plugs. The shafts themselves can then be filled, either with concrete or with dry filling to give the final wall the required mass for stability.

The problem of Southampton quay

Success in this form of construction cannot be guaranteed. In the case of the Western Docks at Southampton, Eng., constructed between World War I and World War II, it was found impossible, except at inordinate cost, to get the monoliths to sink through the opposing strata to the depth required for stability as a retaining wall. It was therefore necessary to reduce the thrust involved in this function by cutting the retained material back to a natural slope and spanning the gap between the back of the monoliths and the top of this slope by means of a reinforced-concrete relieving platform, supported along its other edge on reinforced-concrete piles. This arrangement has served well enough as far as the quay wall itself is concerned, but the maintenance of the natural slope, stone-pitched as a protection against erosion, has been a continuing liability. In addition, the presence behind the quay of the relieving platform constitutes a formidable obstacle to further construction work—*e.g.*, warehouses or multistory transit sheds.

Concrete caisson walls. In situations in which the depth from ground level to the final dredged bottom is not excessive and the material available for retention as reclamation is of good self-supporting qualities, quay walls can be constructed of precast concrete caissons floated into position and sunk onto a prepared bed in the same manner as that described for breakwaters. Care is taken to design caissons able to withstand the thrust of the retained material, which is carefully selected for the areas immediately behind the quay wall. The conditions suitable for this form of construction are generally typical of the Mediterranean, where the slightness of the tidal variation keeps the depth required to a minimum. An outstanding example of this kind of construction is the extension to the area of the Principality of Monaco, which is being increased by as much as 22 percent by reclamation retained by this technique. Similarly constructed installations for transportation and ship-repair purposes exist elsewhere in the Mediterranean, in parts of which the earthquake factor is an additional influence on the retaining-wall design.

In all cases of dock wall construction by concrete monolith or caisson, it is the basic structure of the wall that is provided by these means; the final superstructure, above highest tide level, will depend for its detail on the requirements for dockside services, crane tracks, and other elements.

The piled jetty. The high cost, difficulties, and possible dangers of providing dock and quay walls of the kind just described have always encouraged a search for alternative solutions that would eliminate the need for operations on or below the seabed. Of these, the earliest and most obvi-

ous is the piled jetty—its piles can be driven from floating craft and the deck and superstructure added thereto, working wholly above water. In regions in which there is a large tidal range, it may sometimes be both advantageous and necessary to take the opportunity provided by extremely low tides to make attachments to the piles for bracing and stiffening purposes. With a reasonable programming of the work, this operation can usually be done without particular difficulty, assuming that the seabed is of a composition reasonably amenable to penetration by piles to a sufficient depth to secure the lateral stability of the structure. Hard rock is not suitable, although some of the more friable rocks can be pierced by steel piles.

Timber piles

Piles may be of timber, reinforced concrete, or steel. Timber is a popular choice if there is a large natural supply. Lateral stiffness and stability can be achieved by using a sufficiently close spacing of the piles in both directions and adequate rigid bracing between the tops, timber being a material readily amenable to the workmanship required. Its chief drawback is lack of durability, particularly in the area between wind and water, although a timber jetty with reasonable maintenance can often resist normal operational obsolescence. There are examples of construction in which the piles are connected together by casting a reinforced-concrete slab around the heads, its soffit (underside) just below lowest water level. By this means, the timber is kept continually submerged, a condition under which its durability is prolonged. On the other hand, in tropical or semitropical waters or waters kept warm by industrial effluents, the use of timber may be inhibited by the presence of marine borers. Timber jetties have a considerable advantage in the comparative ease with which repairs to accident damage or deterioration can be effected.

Reinforced-concrete piled piers and jetties, soundly constructed, exhibit great durability. Attachment to the piles for bracing and similar purposes tends, however, to be more complicated than in the case of timber. This is a disadvantage that applies also to subsequent maintenance and repairs.

Sheet-piled quay. An extension of the piled jetty concept is a quay design based on steel sheetpiling, the design becoming increasingly popular with improvements in the detail and manufacture of the material. Steel sheetpiling consists in essence of a series of rolled trough sections with interlocking grooves or guides, known as clutches, along each edge of the section. Each pile is engaged, clutch to clutch, with a pile previously driven and then driven itself as nearly as possible to the same depth. In this way a continuous, impervious membrane is inserted into the ground. In most designs the convexity of the trough sections is arranged to face alternately to one side and the other of the line along which the membrane is driven, so that a structure of considerable lateral stiffness is built up. At the same time, a measure of flexibility in the clutches allows some angular deviation so that a membrane curved in overall plan is obtainable, a feature of considerable convenience in developing the layout of a series of wharves or quays.

The development of steel sheetpiling over the years has largely been characterized by the increasing weight and stiffness of the sections available from the rolling mills. In one design, the clutch is a separate unit from the main structural element, generally of broad-flanged or universal beam section. In this case, the clutch unit appears in a profile of two grooves, or channels, back to back, each capable of embracing the flanges of adjacent beams, which are thus locked together in a continuous sheet, or membrane, of considerable strength. Each universal section is entered, when pitched for driving, into the clutch on the previously driven section and usually carries the clutch for the next section with it. In another design, made economically possible by the advances in the technique of automatic continuous welding, rolled universal beam sections are welded by one flange into the troughs, or pans, of conventional sheet piles, the composite construction producing a unit of unique strength and stiffness.

The development of steel sheetpiling has kept ahead of the development of hammers capable of driving it, probably because the stiffer the section is, the greater the length of pile that can be incorporated in a design. The combination of heavier section and greater length demands a greater proportion of the energy delivered by the hammer being unproductively absorbed in the temporary elastic compression of the pile, leaving less energy to drive the pile further into the ground. Simply increasing the amount of energy delivered, by using a heavier hammer or a higher drop, does not necessarily provide the solution; it may only result in damage to the head of the pile without achieving greater penetration. This difficulty has been in part overcome by the use of high-strength steel piles. Nevertheless, it is not unknown for a pile to appear to be going down with little or no head damage when it is, in fact, sustaining extensive damage below seabed level that gravely compromises its efficiency as a retaining quay wall. This situation, usually the buckling of a pile, can occur particularly where the material of the seabed contains boulders or similar obstacles to penetration.

The problem has obvious major implications for the construction of quay walls and has provoked much debate among engineers. The skill of the quay designer and the advice of the soil mechanics specialist both contribute to the satisfactory reconciliation of the various conflicting factors outlined in order to achieve the most effective and economical solution.

Protecting the quay

In the normal design of sheet-piled quay or wharf wall, the sheetpiling itself forms the quay face, although it is generally found advisable to protect the piles from the impact of ships berthing by timber fenders. Vertical timbers at intervals are generally used. Horizontal walings (wooden ridges) between these timbers can also be employed, but they have a disadvantage, particularly at small wharves and with ships having their own protective belting: on a rising tide the beltings become entangled with the walings, occasioning damage or even minor disaster.

The upper part of the sheetpiling, being laterally unsupported on the sea side, is generally anchored back to resist the thrust of the retained soil. This resistance is commonly effected by using tie rods secured to anchors buried in the retained soil itself to a depth that, for reasons of overall stability, is beyond the natural slope line of the soil. As often as not, these anchors are themselves composed of lengths of sheetpiling driven, if possible, below the retained soil into the strata beneath. The mild, or alloy, steel tie rods, coated and wrapped against corrosion, can be carried through the exposed sheetpiling of the quay wall with large retaining nuts on the outside or can be secured to welded attachments at the back of the piling. The latter practice is the more commonly favoured arrangement, largely on account of its more finished appearance. The sheet-piled quay just described is completed by casting a reinforced-concrete cope beam to cover as well as contain the exposed heads of the sheetpiling.

The advantage of this type of quay wall is that the space behind the wall is not occupied—as in the case of the suspended pile-supported deck—by a monolithic, fully structural element, the arrangements of which can be disturbed for subsequent modification of the services layout only at some cost and usually by a potentially complicated design operation. As in the case of a gravity wall, the space can be filled with suitable material that can subsequently be treated, for all intents and purposes, as natural ground in which service ducts can be buried if required. This arrangement is often an advantage in the case of freshwater mains for fire fighting or watering ships because they can thus be protected from frost. Alternatively, it is possible to place concrete-lined service and cable trenches in this material, sometimes conveniently by the use of precast sections, because the ground loads imposed are seldom sufficient to give rise to serious settlement problems.

Structural reinforcement. Identifiable structural loading—arising, for example, from crane tracks—can be supported on reinforced-concrete beams on piles driven through the filling to the strata beneath. Dockside railways, a decreasing requirement because of the transfer of much shore-to-ship delivery to road vehicles, need not necessarily have piled support, because the loading from these can be spread to remain within the bearing capacity of the filling. Some settlement is bound to take place, and

the need for compensating by packing up and releveling of the track has the incidental disadvantage of breaking up the surfacing of the quay, which is almost always provided to facilitate quayside access by road vehicles.

Sheet-pile quay walls are readily applicable to sites at which only relatively shallow or medium-depth water alongside is needed. As the required depth increases, a sheet-pile section of sufficient strength and stiffness to hold the retained material without further assistance becomes impractical from the point of view of handling and driving. A solution increasingly favoured is the so-called Dutch quay. In this design, after the line of sheetpiling has been driven using one of the heavier and stiffer sections, the ground behind is excavated for a distance determined by the natural slope of the material to be used as filling and taken down as far as possible to lowest water level. At this level, a reinforced-concrete relieving platform is constructed up against the sheetpiling but with independent vertical support from bearing piles driven through the bottom of the excavation to an appropriate depth. Piles for crane tracks are driven at the same time as these—that is, before the construction of the relieving platform.

The Dutch quay

Filling material is returned above the relieving platform, and, although the latter now prevents further pile driving in the area, the probability of this being required is remote, whereas the retained load against the sheetpiling is much reduced. The advantages of having filled material behind the sheetpiling for installing services remain. In addition, the relieving platform affords the sheetpiling considerable help in resisting horizontal blows from the impact of berthing ships, and in order to increase this resistance some of the piles supporting the platform are often driven toward the quay face. Reinforced-concrete counterforts between the platform and the sheeting can be an additional help.

Durability. A question that hung over the use of steel sheetpiling in salt water in its early years concerned its durability in potentially hostile conditions. The rate of corrosion, particularly at the waterline or within the tidal range, varied from one locality to another according to the state of the water and the effect of such factors as salinity and industrial effluents. Precoating of the pile with a protective film such as tar or a bituminous paint is of only transient value, requiring regular renewal, and is effective only down to the lowest water level.

The inclusion in the composition of the sheet-pile steel of a very small percentage of copper was tried as a means of increasing its durability, but the effectiveness is doubtful.

Cathodic protection

The confirmation of the electrochemical basis of much of the corrosion affecting steel sheetpiling led to the development of cathodic protection, a process that has wide application in many other fields, especially shipbuilding. Electrolytic corrosion arises from the passage through the piling of electric currents, causing the pile, or part of it, to become the anode, or positive pole, in what amounts to a galvanic cell, or battery. In such a cell, metal is normally removed from the anode and may reappear on the cathode, or negative pole, which remains unaffected. These currents in sheetpiling may arise from stray leakages from adjacent electrical installations or be generated within the pile itself by differences in the electrolytic conditions at differing levels on the pile.

Cathodic protection is a means whereby cathodic polarity is imposed upon the whole pile, and its operation as an anode (with consequent deterioration) is prevented. This can be done either by supplying from a suitable source—*e.g.*, a battery—an electric current that will overcome and reverse the direction of the naturally generated current or by connecting the piling at intervals to sacrificial anodes of an element—generally aluminum or magnesium—whose atomic relationship to the steel in the piling is such as to generate a current without external assistance. These anodes are buried in the surrounding ground, and care must be taken to ensure full electrolytic continuity between them and the piling to complete the circuit. It is sometimes necessary, in order to ensure electrical continuity between the anode connections in the piling itself, to weld adjacent piles together after driving.

By whatever means cathodic protection is applied, a small liability for operational maintenance arises, either for the continuous supply of the imposed current or for the periodic renewal of the sacrificial anodes. The considerably increased durability of the structure usually justifies this.

Enclosed docks. Whenever possible, commercial quays are built open to the tide range to provide maximum freedom for shipping. There are, however, some parts of the world in which the range between low water and high water is so great that the resulting variations in the level of the ship's decks and hatches impose unacceptable disabilities on the handling of cargo. In such circumstances the quay walls may become of such dimensions as to be uneconomical. (The net clear height of the quay walls, disregarding depth of foundations, must span the distance from the lowest seabed level acceptable for navigation at low tide to an adequate freeboard for the coping of the quay wall above the level of the highest high tide. This condition is equally applicable in cases in which only the berths themselves are made to be usable no matter what the stage of the tide.)

The problem can be met by constructing the quays as enclosed docks in which the water level is kept constant and access to the tidal areas is by means of a lock or locks. An obvious condition for the success of such an arrangement is that the strata of the bed under the enclosed dock area be sufficiently impervious to preclude any significant loss of water through the bottom during low-tide conditions. In this way the tidal range, as a limit on the height of the quay walls, can be eliminated.

Locks

Apart from the fact that they have gates at each end, the structure of maritime navigation locks and the problems involved in their design are very similar to those of dry docks. Although, in normal usage, a lock is never completely dry, it is essential that it should be designed to be capable of withstanding the stresses imposed by this condition so that it may be possible to dewater the lock completely for inspection and maintenance.

It is common practice to design enclosed docks so that the normal water level maintained is not below the highest likely high tide because the invasion of an enclosed dock by a high tide significantly above the normal water level can be disastrous.

Although enclosed docks are frequently of such an area that they can supply the lockage water lost when a ship passes through the lock without any drop in level that cannot be made up on the next high tide, it is normal to provide a measure of impounding capacity in the form of pumps for lifting additional water from outside into the dock. Such a provision is essential for situations in which it is required to keep the enclosed dock water level above the highest tide.

It has sometimes been possible to accommodate ships of larger draft than originally planned for in large but relatively old enclosed docks. This is done by installing impounding pumps for topping up the water level to give an increased depth.

Enclosed docks generally suffer the operational disadvantage of restricted times of entry and exit because they are subject to a fairly rigid tidal schedule. First of all, the lower the tide level outside, the greater the amount of water lost in the locking operation; and, second, it is seldom economically feasible to maintain full navigation depths in the approach channel to the lock entrance at all levels of the tide. This situation is particularly the case in which enclosed docks are sited adjacent to and operating from a tidal river estuary. The tidal lock at Dunkirk, Fr., opening to allow the passage of the night channel ferry, which runs on a timetable, is an example of a tidal lock operated whatever the state of the tide.

If possible, the access locks are usually duplicated, lest an accident involving the gates or the structure of the lock put the whole dock area out of operation. Stability calculations of the quay walls within an enclosed dock are important; in older installations such calculations may have been based on the continuing presence of water at the designed normal level, and in the event of a serious failure at the lock—resulting in a considerable drop in the water level—the stability of the quay walls could come into question.

ROLL-ON, ROLL-OFF FACILITIES

An enormous increase in the use of the roll-on, roll-off technique of loading and unloading developed in the late 1960s. The principle of embarking whole vehicles under their own power was not new. The report of Hannibal ferrying his elephants over the Rhône in the 3rd century BC might be regarded as the earliest example from which the vast amphibious operations of the invasion of Normandy in 1944 were descended. Since the 1960s, however, the spectacular increase in the use of road transport for heavy freight and the increase in handling charges at ports for the loading and discharge of cargo by conventional means have combined to provide the impetus for the rapid commercial development of the roll-on, roll-off technique. In addition, the tendency to assemble machinery at its place of manufacture in larger and larger units has encouraged the development of special transport vehicles, and the economies of moving load and vehicle together from origin to destination can be considerable.

Berthing and access problems

The principal problem for the port engineer is to provide special berthing for the ferry vessels and means of access for vehicles from the shore to the ship's decks. Railcar ferries, involving somewhat similar problems, have been known for some time, but, because of the severer limits on gradients for such vehicles, there has been a tendency to limit the operation of these services to terminals at places where the tidal range is inconsiderable. For the Dover-Dunkirk ferry, opened shortly before World War II, a special enclosed dock was constructed at Dover in which the water level could be kept constant for loading and unloading, while at Dunkirk the entire dock system is totally enclosed, accessible through sea locks.

Many of the new roll-on, roll-off terminals for road services are, by contrast, in tidal water; and, where the tide range is large, access bridges of considerable length are often needed to keep the change of gradient between low and high tide within acceptable limits. The change in the ship's trim between conditions of light loading and full loading creates yet another problem.

Disadvantages of pontoons

At first sight, the solution might appear to be to support the outer end of the link span on a float, or pontoon, so that it would automatically follow the rise and fall of the tide. Several disadvantages of structural detail arise, however, and the system is vulnerable to damage caused by the movement of the pontoon under adverse weather conditions. A means to adjust the height between the span and the supporting pontoon to accommodate changes in a ship's trim is still required; and, therefore, the overall economies of a pontoon are less than might at first be imagined.

Thus it is almost universal practice to support the outer end of the link span from an overhead structure, either through conventional wire-rope hoisting gear or by means of hydraulic rams. The level of the end of the span can thus be continually adjusted, either automatically or by manual control, to match changes in the level of the ship's deck, whether caused by the tide, by the trim of the ship, or by differences in deck levels between one ship and another. Maximum flexibility of access has become increasingly important with the appearance, on some services, of ships with two independent car decks, both of which must be equally accessible to the link span. This situation has sometimes been achieved by the use of double-decker link spans, a technique that has the effect of keeping the length and—unless the span is intended to carry loads on both decks simultaneously—the weight of the span to a minimum.

The sudden proliferation of roll-on, roll-off services simultaneously has led to a rather unfortunate development: a number of the terminals have tended to be tailor-made to suit particular ships and to be unable to accept different ships without, in some cases, quite major structural alteration. This feature clearly reduces the otherwise great flexibility of this technique, and an international commission to examine the question was appointed in 1970 by the Permanent International Association of Navigation Congresses.

Maximum advantage of the roll-on, roll-off technique is gained in relatively short sea passages. On longer voyages, the idle road vehicles make the economies questionable. This problem can be overcome to some extent by embarking only semitrailers and leaving the tractive units ashore; the practice has no effect on the terminal details.

BULK TERMINALS

The enormous increase in the marine transit of materials in bulk, with petroleum leading the way, has given rise to the development of special terminals for the loading and discharge of such materials. The principal factor influencing the design of these installations is the still-increasing size of the ships. A single example of the effect of this change on design limits will be sufficient. The "Queen" liners, long the world's largest ships, never drew more than 42 feet of water. Supertankers, on the other hand, when fully loaded, draw up to 72 feet. If these ships required berthing structures of the type provided for conventional cargo and passenger liners and if the formula relating the capital costs of such structures to the deepest draft were applied, the cost of building an appropriate berth for such a tanker would reach a figure more than six times the cost of the *Queen Mary*'s old berth. Fortunately, the high mobility of the cargo renders such drastic and expensive measures unnecessary. Heavy capacity access for individual shore-based vehicles to carry away the cargo is not required, nor does the provision of services for the relatively small crews who man these great ships present any problem. The berthing positions can therefore be sited well out from the shore in deep water, and the structure itself can be limited to that required to provide a small island with mooring devices.

Unloading oil from tankers

In the case of oil terminals, the link to shore can be a relatively light pier or jetty structure carrying the pipelines through which the cargo is pumped ashore, with a roadway for access by no more than average-size road vehicles, which will probably be used in small numbers or even only one at a time (see Figure 41). Because the ship itself carries the pumping machinery for delivering the cargo ashore, heavy mechanical gear for cargo handling is not required.

In the case of bulk carriers bringing solid commodities, such as iron ore, the problem is more complicated. Hoisting grabs for lifting the ore out of the holds are necessary, even though transit between ship and shore can still be effected by continuous conveyors, corresponding to pipelines. Heavier foundation work is probably necessary at the berthing point to carry this machinery, and, for this reason, ore terminals have not been built as far out in deep water as oil terminals. It seems unlikely that the size of ore carriers will reach anything like the dimensions already attained by supertankers.

The employment of piled structures to meet these requirements is almost universal, and a variety of techniques have evolved for handling and sinking into the seabed the

By courtesy of British Petroleum



Figure 41: The oil jetty at Rotterdam, Neth.

long heavy piles required. At the sites likely to be chosen, penetration by piles may not be easy, particularly in places where most of the reasonably accessible deepwater sites tend to be located on the rockier shores.

One problem that arises is that of shelter in adverse weather conditions. While the ships themselves are reasonably robust, the relatively fragile berthing structures might break up, setting the ship loose, possibly without power immediately available, threatening disaster. As the cost of building breakwaters to protect sites in the depth of water required is likely to be prohibitive, the search has been for natural shelter. In the British Isles the sheltered creeks of the western shores, such as Milford Haven, Wales, have become valuable. Milford Haven had known little shipping other than fishing fleets since the early 19th century, but in the early 1970s it boasted four bulk oil terminals. Two supply refineries were built on the spot; the third pumps to a refinery 60 miles away.

Another aspect of the terminals is the need for protection against the effects of unavoidable collision impacts. A slight impact from a vessel of these dimensions, by reason of the large kinetic energy of such a mass, can cause considerable damage to the light berthing structure. Much ingenuity and theoretical analysis have gone into devising fendering systems that will absorb this energy. Some systems use the displacement against gravity of large masses of material disposed pendulumwise in the berthing structure as the energy absorbent; others use the distortion by direct compression, shear, or torsion of heavy rubber shapes or sections; still others rely on the displacement of metal pistons against hydraulic or pneumatic pressure. The common feature of all the devices is that at least part of the energy absorbed is not dissipated but is used immediately to return the ship to its correct berthing position. This feature is not exhibited by the older forms of fenders, which relied on the compression and, in extreme cases, on the ultimate destruction of coiled rope or timber to absorb the impact. A major question is the exact ship velocity to be allowed for, the determination of which is primarily an exercise in probability, balancing the economics of designing to a specified velocity against the cost of repairs after impacts at greater velocities. The key factor is the frequency of such impacts, which can be determined only by experience.

DRY DOCKS

The largest single-purpose structure to be built by the maritime civil engineer is not directly connected with loading, unloading, or berthing but is indispensable to prolonging the life of ships. This is the dry dock, which permits giving necessary maintenance to the underwater parts of ships. The problem of dry-docking is aggravated by the tendency of ships to grow in size by increases in beam (width) and draft (depth below waterline) rather than in length, a process that rapidly renders many of the world's largest dry docks useless for servicing an increasing proportion of the traffic.

Danger of obsolescence

A classic example is the King George V Drydock at Southampton, Eng. Opened in 1933, it was 1,200 feet long and 135 feet wide and was capable of accommodating the largest vessels afloat—namely, the two Cunard liners *Queen Mary* and *Queen Elizabeth,* each more than 80,000 tons deadweight. The later supertankers have deadweight tonnages of 135,000 tons and more, within a length of about 1,150 feet but with a beam of about 175 feet, which precludes them from entering the King George V dock. The lengthening of a dry dock would be a comparatively simple and economical operation; widening, on the other hand, involves at least the complete demolition of one sidewall and its rebuilding to give the increased clear width to the other wall, assuming space can be made available. Increasing the depth would mean a new dock altogether, but, because tankers generally dry-dock in the unloaded condition in which their draft can be considerably less than that of a conventional ship, this problem has not so far been a practical one.

Structural requirements. Moreover, in a great many cases, the maximum state of stress in a dry dock occurs not when it is carrying the weight of the ship (always considerably less than the weight of the water occupying the dock when flooded) but when it is completely empty and subject to the pressures generated by water in the surrounding ground, particularly under the floor, the support of which may lie at a considerable depth below the level of the adjacent water table. To ensure against any tendency to lift under this pressure, the floor must either have sufficient weight in itself (1 foot, or 300 millimetres, depth of concrete will resist a little less than 2½ feet head [depth] of water) or be designed as a structural element capable of transmitting this pressure laterally to the walls of the dry dock, which can then be designed to contribute the additional extra weight required. Obviously an operation involving both the complete rebuilding of one wall of a dry dock and the strengthening of the floor to cover an increase in its span as an inverted arch or beam is almost tantamount to the construction of a complete new dock.

This problem received somewhat tardy recognition, so that, although several large new dry docks were built around the world in the 1960s, only a minority were capable of allowing the entry of tankers of more than 200,000 tons.

Design. The design of a dry dock probably depends more on ground conditions than does any other engineering structure, with the possible exception of large dams. Mention has been made of the need in many cases to resist upward pressures under the floor. Apart from the simple solution of using the weight of the dock structure itself for this purpose, which is not economical, devices that have been tried include "pegging" the floor to the underlying strata by means of piles or prestressed anchors and extending the floor slab itself beyond the sidewalls, thereby gaining assistance from the weight of the material filling behind the walls, which are designed to act as retaining walls to this filling. Venting of the floor to relieve water pressure can sometimes be of help provided the volume of water so released is not excessive. If it is, continuous pumping to keep the dock dry will be necessary. On sites in which water pressures do not have to be resisted, the design is generally simpler, and sufficient strength and stiffness to spread the loads from the ships' keels over the underlying ground so as not to exceed the bearing resistance of the latter is the controlling floor-design factor.

Dry docks in shipbuilding

The use of dry docks for the building rather than the maintenance of ships is a practice that has been increasingly adopted. Both the building and the launching of a ship in these circumstances can be considerably simplified. The designs of such dry docks are no different from those hitherto described; what is possibly the largest dry dock in the world was completed in Belfast, N.Ire., in 1970. This dock, built along the site of a former channel between two open basins, is capable of accommodating the three Cunard liners *Queen Mary, Queen Elizabeth,* and *Queen Elizabeth 2* simultaneously and is to be used for the building of large tankers. It is spanned by a crane of 400 tons lifting capacity to handle large prefabricated ship sections.

Entrances. Dry dock entrances are closed by gates of different designs, of which the sliding caisson and the flap gate, or box gate, are perhaps the most popular. The sliding caisson is usually housed in a recess, or camber, at the side of the entrance and can be drawn aside or hauled across with winch and wire rope gear to open and close the entrance. The flap gate is hinged horizontally across the entrance and lies on the bottom, when in the open position, to be hauled up into the vertical position to close the dock—a process occasionally facilitated by rendering the gate semibuoyant through the use of compressed air.

The ship type of caisson gate, a quite separate vessel floated and sunk into its final position across the entrance, is largely out of favour. Although it was comparatively easy to remove for maintenance and had the further advantage that a spare caisson could be kept in reserve in case of damage, the tie-up of capital is usually found unnecessarily expensive merely as an insurance premium.

The maximum degree of watertightness obtainable between the gate and its seating is essential if continuing and expensive operational commitments for pumping out leakage water are to be avoided. The pressure of the water outside the gate is available to provide a powerful sealing

force, but special treatment of the actual contact faces is necessary to make this force fully effective. For a long time it has been held that the only satisfactory arrangement was by the use of a timber lining (generally greenheart) around the contact face on the gate, bearing against stops in the dock structure composed of granite dressed and polished to a high degree of accuracy. The increased expense of such methods and the diminishing number of skilled labourers capable of dressing the granite have led to a search for alternatives. These include such devices as the use of stainless facing bars set in concrete, in place of the dressed granite, and rubber linings on the gates themselves. While these have generally proved effective when first installed, more experience is needed to determine their durability as compared with older methods.

Keel and bilge blocks. Keel and bilge blocks, on which the ship actually rests when dry-docked, are of a sufficient height above the floor of the dock to give reasonable access to the bottom plates. Such blocks are generally made of cast steel with renewable timber caps at the contact surfaces. Individual blocks can generally be dismantled under the ship to allow access to that part of the plates, if required, and can be reassembled to take their appropriate share of the weight after the operation required has been completed. Most modern ships, particularly tankers, are of nearly square section over a large part of their middle length and can be kept upright in dry dock by the support of the bilge blocks under their bilge keels. In the most up-to-date dry docks, the bilge blocks are provided with mechanical means for traversing them across the dock and altering their height by remote control while the dock is still flooded. This arrangement permits them to be adjusted in their correct position according to the shape of the ship while the latter is still just afloat but in contact with the centre-line keel blocks. The economic advantage of this arrangement is considerable because it allows one ship to be removed and another put into the dry dock on the same opening of the gate, whereas under previous practice it would have been necessary to close the dock and pump it out to reset the bilge blocks to the known profile of the next ship. Apart from the time needed, the power consumed in pumping out a large dry dock is a considerable factor.

Because of the increasing number of ships suitable for bilge docking, the use of side shores to keep hulls upright in dry dock is a rapidly dying process, and indeed the altars provided for this purpose in dry docks of more old-fashioned design are often an embarrassment to the accommodation of a modern square-sectioned ship. Frequently this situation is remedied by cutting away some altars, an operation that must be conducted with discrimination because the removal of any quantity of material from the sidewalls may have a damaging effect on their stability.

Construction. *Basic technique.* Dry docks are usually constructed in open excavation in the dry, shutting out the sea by means of a cofferdam. Sometimes it is found convenient to construct the sidewalls first, in trench, next to remove the loose material between them, and then to lay the floor in stages so as not to endanger the stability of the walls before the floor is in position to give them toe support. Extensive pumping, to keep the excavations from filling with water during construction, is generally necessary.

Keeping out the sea

In one rather unusual case, a dry dock for 240,000-ton tankers was constructed almost wholly under water because large fissures in the rock running through to the sea flooded the site beyond the capacity of any reasonable assembly of pumping equipment. The entire space required for the structure was therefore excavated to formation level by dredging, and the sidewalls were constructed first, using prefabricated concrete caissons sunk into place and filled with concrete. The spaces between adjacent caissons were sealed by filling with concrete in the same way. Stone aggregate, to a depth of 23 feet, was then deposited between these walls and consolidated into a concrete floor by a process of grouting in which colloidal cement grout was forced under pressure between the interstices of the aggregate, subsequently setting to form the whole into concrete. A similar process across the floor at the entrance incorporated a cofferdam of interlocking steel sheetpiling, which allowed the sill and gate hinge to be constructed in the dry. The gate, of the flap variety already mentioned, was floated and stepped into place by divers after the removal of the cofferdam. Only then was it possible to pump out the main body of the dock, which was completed by laying a reinforced concrete topping over the floor in order to provide a satisfactory working surface.

Floating dry docks. Floating dry docks have the initial advantage that they can be built and fully equipped in shipyard and factory conditions, in which their construction is not subject to unforeseen hazards arising from weather and variations in the ground conditions from those anticipated during design. The floating dock can be towed to the site, moored, and made ready for operation in a comparatively short time. Expenditure on temporary works, often a large fraction of the cost of a fixed dry dock, is also avoided.

Floating dry docks are usually fully self-contained (Figure 42). The sidewalls provide much of the residual buoyancy and stability required to keep the dock afloat when it has been submerged far enough to allow the entry of a ship into the docking space over the main deck. Most of the machine tools and workshop equipment required for all the normal operations of ship repair and maintenance are also housed in the walls as well as the generating plant (usually diesel driven) to supply power for the operation of the dock and its equipment. Traveling cranes, for handling material off and onto the ship, run on the tops of the sidewalls.

A floating dry dock can be moved at relatively short notice to another site, should a long-term change in shipping-traffic patterns dictate a change. This advantage may be more apparent than real, because the large work force required to man it may not be so readily transferable.

Moreover, floating dry docks tend to have large maintenance costs because the steel structure, being continually afloat, requires regular chipping and painting, as the hull of a ship does. The above-water structure presents no particular problem and can generally be given maintenance care without putting the dock out of use. The most vulnerable areas, those immediately adjacent to the waterline,

From A. Amirikian in "Transactions of the Society of Naval Architects and Marine Engineers," 1957

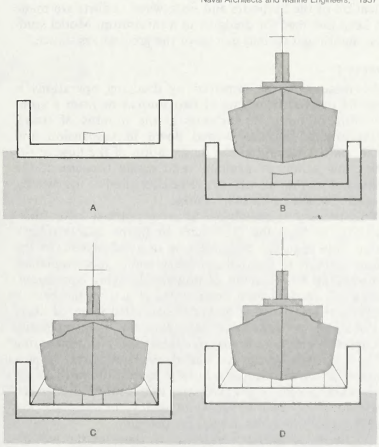


Figure 42: *Operation of a floating dry dock.*
(A) Dock afloat in its light-draft condition, with little or no water in its ballast tanks. (B) Dock submerged to its maximum draft by filling the ballast chambers; a vessel floats freely in the flooded channel. (C) Dock floor emerges from the water, and the vessel comes to rest on the support blocking. (D) All ballast is pumped out, raising the dock floor completely out of the water and providing a dry working area.

can be reached by careening, a process that involves filling the water ballast tanks along one side to induce a list that lifts those on the other side part of the way out of the water. On completion, the process can be reversed for the other side.

Underwater scaling and painting

Maintenance. Methods of underwater scaling and painting, or the use of limpet dams with which small areas can be covered with watertight enclosures inside of which people can work under compressed air, allow a limited measure of attention to be given to the bottom plating outside. Occasionally it is necessary to detach one of the sections of the dock, which is usually constructed in separate sections for this reason, and dry-docking it in the remainder, repeating the process until the whole dock has been renovated. This costly and tedious process is resorted to only for compelling reasons.

To give a floating dock sufficient depth of water for submerging the docking blocks below the keel of the ship to be docked, it may be necessary to dredge a berth for it. In areas subject to heavy siltation, this dredged area will almost certainly act as a silt trap. Periodic removal of the dock from the berth to allow the latter to be redredged is an additional source of expenditure in such cases. Finally, in places where the tide range is of consequence, special mooring arrangements are necessary to restrain excessive lateral drift of the dock as the mooring chains become slack on low water.

The arrangement of keel and bilge blocks is generally similar to those described for fixed dry docks.

Sea works for reclamation and conservancy

An indispensable item of equipment over a wide range of the maritime civil engineer's activities is the dredge with its ancillary units, such as hopper barges, tugs, reclamation units, and servicing craft. There are few navigable harbours or harbour approaches that do not require, at varying intervals of time, removal of deposits of unwanted material, the continuing accumulation of which can ultimately obstruct navigation. With the current trend toward larger ships, dredging is especially important.

Extensive research has been devoted to the development of dredging equipment. Through more sophisticated techniques—including, in some cases, permanent profile modification of the harbours and waterways—efforts are made to keep the need for dredging to a minimum. Model studies, mentioned earlier, can be of the greatest assistance.

DREDGING

The material to be removed by dredging operations is usually derived from one of two sources or from a combination of both. In harbours at the mouths of rivers, quantities of silt are carried down in suspension and tend, partly because of the deceleration of the flow in the increased waterway available and partly because of the effects of increasing salinity, to be deposited at the mouth, usually the site of harbour works.

Estuarial silt

This process has produced areas of marked agricultural fertility, such as the Nile delta in Egypt. While over a large time span the action is one of great benefit, in the short term it is generally a considerable inconvenience. The skillful employment of modern dredging equipment, however, has indicated possibilities of getting the best of both worlds. The other source of deposited material likely to obstruct navigation is littoral (coastal) drift, especially in areas where there is a sizable tidal range. The incoming tide frequently brings suspended material, some proportion of which settles to the bottom around the turn of the tide when the movement of water is at a minimum. In the absence of any countervailing tendency, an accumulation takes place, which again requires dredging.

For many years the workhorse of many of the world's dredging fleets has been the bucket-ladder dredge, operating a continually moving chain of open-ended shovels or scoops. At the bottom of the ladder the scoops are pushed into the face of the material and empty themselves as they turn over at the top, the material falling into chutes that divert it into hopper barges for removal. A four-point mooring system enables the craft, and with it the bucket ladder, to be held up to the working face and, at the same time, swung sideways across it in either direction. By this means, an often remarkably level bed to the sea bottom can be closely controlled by adjusting the position of the ladder under the dredge's bottom. The positive action in filling the buckets enables such a dredge to tackle material of considerable stiffness, thereby extending its use to works of dredging and harbour development in which soils other than recently deposited silt or sand have to be excavated. Even some of the softer rocks can be removed in this way if the buckets are provided with hardened and stiffened edges and ripping teeth.

The principal disadvantage of the bucket-ladder dredge is the need for an elaborate system of fixed moorings. The area that can be covered by one placing of the moorings is limited. Continuous lifting and replacing of the moorings are not only time-consuming but must be carried out in such a way as to offer minimum obstruction to navigation, a requirement that sometimes involves a great number of interruptions in dredging operations.

The suction dredge

In areas in which the deposited silt is highly mobile and accumulates in considerable quantities, it can be economically removed by a suction dredge, which pumps water mixed with silt into open hoppers. By adjustment of the capacity of the hopper to the rate of flow from the pump, the water can be made to remain in the hopper long enough to deposit most of the silt. Careful design of the pumping machinery is required to assume a continuous mixture of maximum silt with minimum water.

The first suction dredges generally operated from moored positions in the same way as bucket-ladder dredges, but a less elaborate system of moorings generally sufficed because the leveling of the seabed could be left to occur naturally through the mobility of the material. A marked advance was achieved by the elimination of much of the lifting and laying of moorings through the development of the trailer suction dredge. This craft has the capacity to dredge while on the move and cruises up and down the waterway or other area, sucking up silt as it goes. This operation does not eliminate all interference to navigation, because a working trailer suction dredge moves more slowly than a ship under normal steerage way, but the obstruction is markedly less. The dredge's turn at the end of each sweep is usually facilitated by the incorporation of a bow side thrust propeller.

The growing tendency to use dredged material for reclamation purposes and the suitable condition for such purposes of the spoil as delivered by a suction dredge have encouraged its development. The seabeds and river bottoms in their natural state are often largely composed of relatively soft material and can be deepened by the use of suction dredges operating normally. Where rock or other hard material must be handled, conditions are favourable to the use of the suction-cutter dredge, which incorporates at the suction head a powerful rotating screw cutter that fragments the hard material. The increased dredging stresses arising from the use of a cutter require that a craft so equipped should be operated as a stationary dredge with moorings. Because such operations seldom take place in areas already under use by traffic, the obstruction problem is not often critical. Additionally, in modern equipment, the incorporation of heavy spud legs in the craft to anchor in the seabed reduces the number of separately laid moorings required.

A useful ancillary piece of equipment to all the above is the grab dredge, either self-propelled or towed to the site. Grab dredges are especially suitable for dredging close up to existing quay walls or other structures with minimum risk of damage, and the grab equipment is often capable of lifting individual boulders. Not infrequently, grab dredges have value for maintenance dredging, particularly in restricted areas and with silt of sufficient mobility to level out the individual holes that are almost inevitably left behind. Although the return fall of the grab takes place with the bucket empty and is, to that extent, nonproductive, with skillful operators this element can be reduced to a minimum, and, with some large craft operating four grabs simultaneously, considerable outputs can be achieved.

Dredges are characteristically designed to deliver their

output either overside into attendant hopper barges or, in the case of self-propelled dredges, into hopper compartments incorporated in their own structure. These hopper compartments are essential in the case of trailing suction dredges, but their value in other cases depends on the circumstances and on the chosen method of disposal of the spoil. When a long journey to the depositing area is involved, it is obviously more economical to leave the dredge continuously at work and to remove the spoil in separate barges.

When the journey is short and the spoil is to be simply dumped, for which purpose the hoppers are provided with bottoms that fall open, an economical work cycle between dredging area and spoiling ground, using one craft only, can frequently be established.

The side-boom dredge

A special case is the side-boom dredge, which discharges straight back overside; by making the work coincide with an appropriate state of the tidal current, this arrangement secures the removal of the dredged silt by the tide's operation.

Dredged spoil is less and less often disposed of by dumping out at sea, a practice that was once almost universal; instead it is used for the reclamation of land from the sea and foreshore. This reclamation process has been stimulated by the rise in the value of the land so created and by the discovery that, in many instances, spoil taken out to sea frequently returns. This phenomenon has been investigated, both on hydraulic models and by mixing radioactive tracers with the dumped spoil in small quantities, permitting its subsequent movements to be followed with Geiger counters.

A variety of procedures have been developed for the combined operation of dredging and reclamation. Where the area to be dredged and the area to be reclaimed are in close proximity, as sometimes happens, the whole operation can be carried out by a single suction dredge pumping ashore through a floating pipeline. When, as is more often the case, there is a considerable distance between the two sites, transport in hopper barges is more economical. At the reclamation site, the barges either can be pumped out by a suction reclamation unit or occasionally can dump their loads on the bottom; from there the material can be pumped ashore by the unit acting as a stationary suction dredge.

The layout of reclamation areas is a matter to which adequate scientific investigation should be devoted, covering such aspects as the adequacy and subsequent maintenance of any navigable waterways it is intended to provide through them, the design of the banks required to contain the pump spoil while the solids settle, and the relative positions of delivery and runoff points to obtain the maximum recovery of solid matter. Such schemes for reclamation, carried out in this way, can simultaneously ensure more valuable new land and improve navigation facilities.

THE DELTA PLAN

It was noted at the beginning of this section that maritime engineering has two large objectives: improvement of transportation and reclamation and conservancy of land. Outstanding among examples of human ingenuity in the second category has been the long effort of the people of The Netherlands to keep their country, large areas of which are below sea level, habitable and productive.

The purpose of these efforts has generally been twofold: first to recover, reclaim, and retain more land for occupation; and second to prevent the percolation of seawater

By courtesy of The Netherlands Sluice and Tunnel Construction Co.

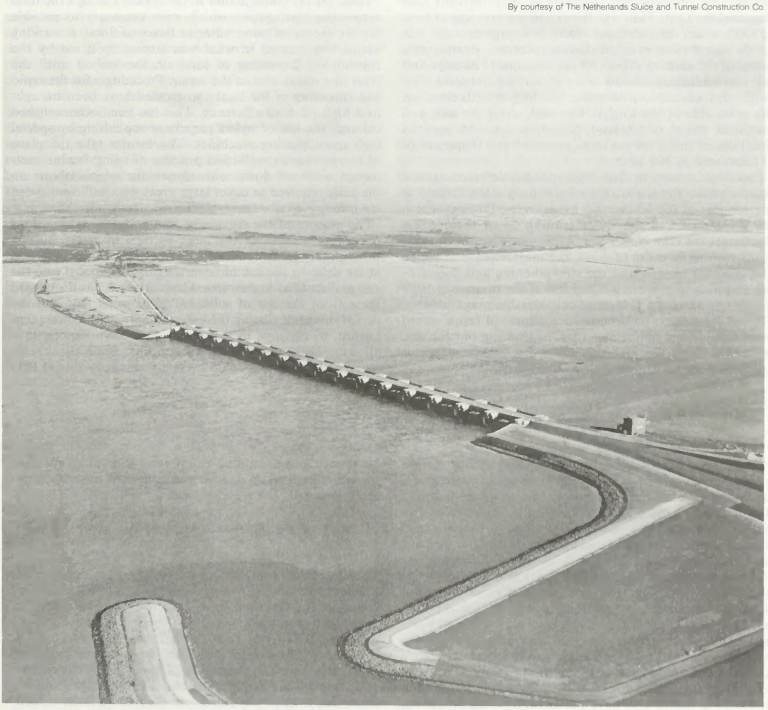


Figure 43: The giant sluices across the Haringvliet, part of the Dutch Delta Plan.

into the water table of both the recovered and the original ground—which, if not prevented, would seriously reduce or even altogether destroy the value of the land for agricultural purposes. This second purpose has sometimes been described as "pushing back the salt line."

Enclosure of the Zuiderzee

A prime example of the first purpose was the enclosure in 1926–32, by means of a dike some 17 miles in length, of a large inlet known as the Zuiderzee (renamed the IJsselmeer after its enclosure). Considerable areas of this body of water have since been reclaimed by the pumping ashore of dredged sand, and the reclamation of further areas is either in hand or planned for the future. A large proportion of the area will, nevertheless, be maintained as a freshwater lake by the flow of the river IJssel, which takes off from one of the outfalls of the Rhine, known as the Lek, or Neder Rhine, just south of Arnhem. In the 1960s it was found necessary to place a dam across the Lek just below the takeoff of the IJssel to divert an increased quantity of Rhine water down the IJssel to the IJsselmeer. The growth of shipping traffic on the canal, which connects Amsterdam with the North Sea, the locking operations of which necessarily discharge quantities of salt water into the IJsselmeer, would otherwise tend to nullify the effects of the freshwater flow of the IJssel.

To maintain navigation in the Lek, in spite of the reduction in water flow, two further dams are provided downstream toward Rotterdam, and all three dams are capable of being opened in the event of excessive floodwater coming down the Rhine.

DESALINIZATION

The second purpose, that of desalinization, has been at the heart of the Delta Plan, one of the most imaginative civil engineering projects ever undertaken. The incident that triggered the Delta Plan was the disastrous flooding of Feb. 1, 1953, when the notorious North Sea surge brought tide levels higher than ever previously recorded, overtopping many of the existing dikes and causing untold damage and salt contamination of vast areas of agricultural land. The surge also caused considerable flooding and damage on the other side of the English Channel, along the east and southeast coasts of England. Statistical research suggests that tides of this level are to be expected at a frequency of at least once in 300 years.

The weak points in The Netherlands' defenses against flooding from the sea are the several deep inlets formed at the mouths of the Rhine and Maas rivers, through which the greater part of the water coming down these rivers discharges into the North Sea. Around the shores of these inlets run many miles of dikes, the maintenance of which is a constant burden and the strengthening and heightening of which to prevent a repetition of the disastrous 1953 floods represented a project of considerable magnitude.

It was considered that the most economical result would be obtained by a major operation of shutting out the sea, more or less at the main coastline, by a series of dams across the mouths of the inlets. By this means some 435 miles of dikes would be cut off from direct sea attack and reduced to a secondary function, whereas the total of the new dams that might still require a measure of maintenance is only 19 miles. By conserving and controlling the vital flows from the Rhine and the Maas, the inlets themselves would be gradually transformed into freshwater lakes, thus greatly contributing to "pushing back the salt line."

A secondary effect in this direction will be an increase in the flow of fresh water toward Rotterdam as a result of the raising of the levels in the estuarial inlets, particularly in the northernmost inlet, the Haringvliet. This result should greatly assist desalinization in the Rotterdam area, where the penetration inland of the salt line had reached alarming proportions as a result of the improvement in the navigational approaches to the port, effected by the construction of the channel known as the New Waterway from the Hook of Holland.

A further benefit to be gained is the great improvement in communications between the mainland and the hitherto somewhat isolated communities on the islands lying between the inlets; the new dams across the inlets will provide foundations for motor roads.

Extent of the Delta Plan

The Delta Plan construction was scheduled to take nearly a quarter of a century, and the total cost represents a significant percentage of The Netherlands' national budget.

Although the authors of the plan stress that it is not properly a land-reclamation scheme (little or no extra land will be created by it), there is no doubt that many of the techniques developed for reclamation work are of the utmost value in carrying out the plan, and, conversely, lessons learned in the course of the project will no doubt find useful application in future reclamation work the world over.

Thus, for the construction of the sluices through the dam across the Haringvliet, which were necessary to provide for the escape of river water in times of flood, a working island was created in what was almost open sea by the continuous depositing of sand on the seabed until the level rose above that of the water. Procedures for the rapid waterproofing of the banks so created have been brought to a high pitch of efficiency. This has been accomplished through the use of nylon carpets or asphalting by special high-speed placing machines. The former take the place of the previously well-tried practice of using fascine mattresses weighted down with stones, for which labour on the scale required to cover large areas with sufficient speed is no longer available.

The closure of the final gaps in the dams, a hazardous operation because of the large volume of water rushing through the narrow remaining gap at this stage, is effected at the delta by the use of concrete caissons floated into the gap and scuttled in position. The technique has progressed there from the use of solid-walled caissons that had the disadvantage of closing the gap suddenly, with consequent hazard, to caissons incorporating their own sluices, thus allowing the flow of water to continue until all were in place and the sluices could be safely closed. (J.H.J.)

LIGHTHOUSES

For centuries lighthouses have had an important function in marine navigation, warning the mariner of hazards, establishing his position, and guiding him to his destination. Traditionally, lighthouses and buoys offer identifiable seamarks by visual cues such as illuminated beacons and painted symbols, but electronic aids such as radar beacons may also be provided.

The development of electronic navigation systems has had a great effect on the role of lighthouses. Powerful lights are becoming superfluous, especially for landfall, but there has been a significant increase in minor lights and lighted buoys, which are still necessary to guide the navigator through busy and often tortuous coastal waters and harbour approaches. Among mariners there is still a natural preference for the reassurance of visual navigation, and lighted marks also have the advantages of simplicity, reliability, and low cost. In addition, they can be used by vessels with no special equipment on board, providing the ultimate backup against the failure of more sophisticated systems.

History of lighthouses

LIGHTHOUSES OF ANTIQUITY

The forerunners of lighthouses proper were beacon fires kindled on hilltops, the earliest references to which are contained in the *Iliad* and the *Odyssey* (*c.* 8th century BC). The first authenticated man-made lighthouse was the renowned Pharos of Alexandria, which stood some 350 feet (about 110 metres) high. The Romans erected many lighthouse towers in the course of expanding their empire, and by AD 400 there were some 30 in service from the Black Sea to the Atlantic. These included a famous lighthouse at Ostia, the port of Rome, completed in AD 50, and

lighthouses at Boulogne, Fr., and Dover, Eng. A fragment of the original Roman lighthouse at Dover still survives.

The Phoenicians, trading from the Mediterranean to Great Britain, marked their route with lighthouses. These early lighthouses had wood fires or torches burning in the open, sometimes protected by a roof. After the 1st century AD, candles or oil lamps were used in lanterns with panes of glass or horn.

MEDIEVAL LIGHTHOUSES

The decline of commerce in the Dark Ages halted lighthouse construction until the revival of trade in Europe about AD 1100. The lead in establishing new lighthouses was taken by Italy and France. By 1500, references to lighthouses became a regular feature of books of travel and charts. By 1600, at least 30 major beacons existed.

These early lights were similar to those of antiquity, burning mainly wood, coal, or torches in the open, although oil lamps and candles were also used. A famous lighthouse of this period was the Lanterna of Genoa in Italy, probably established about 1139. It was rebuilt completely in 1544 as the impressive tower that remains a conspicuous seamark today. The keeper of the light in 1449 was Antonio Columbo, uncle of the Columbus who crossed the Atlantic. Another early lighthouse was built at Meloria, Italy, in 1157, which was replaced in 1304 by a lighthouse on an isolated rock at Livorno. In France the Roman tower at Boulogne was repaired by the emperor Charlemagne in 800. It lasted until 1644, when it collapsed owing to undermining of the cliff. The most famous French lighthouse of this period was one on the small island of Cordouan in the estuary of the Gironde River near Bordeaux. The original was built by Edward the Black Prince in the 14th century. In 1584 Louis de Foix, an engineer and architect, undertook the construction of a new light, which was one of the most ambitious and magnificent achievements of its day. It was 135 feet in diameter at the base and 100 feet high, with an elaborate interior of vaulted rooms, richly decorated throughout with a profusion of gilt, carved statuary, and arched doorways. It took 27 years to build, owing to subsidence of the apparently substantial island. By the time the tower was completed in 1611, the island was completely submerged at high water. Cordouan thus became the first lighthouse to be built in the open sea, the true forerunner of such rock structures as the Eddystone Lighthouse.

Lanterna of Genoa

The influence of the Hanseatic League helped increase the number of lighthouses along the Scandinavian and German coasts. At least 15 lights were established by 1600, making it one of the best-lighted areas of that time.

During this period, lights exhibited from chapels and churches on the coast frequently substituted for lighthouses proper, particularly in Great Britain.

THE BEGINNING OF THE MODERN ERA

The development of modern lighthouses can be said to have started about 1700, when improvements in structures and lighting equipment began to appear more rapidly. In particular, that century saw the first construction of towers fully exposed to the open sea. The first of these was Henry Winstanley's 120-foot-high wooden tower on the notorious Eddystone Rocks off Plymouth, Eng. Although anchored by 12 iron stanchions laboriously grouted into exceptionally hard red rock, it lasted only from 1699 to 1703, when it was swept away without a trace in a storm of exceptional severity; its designer and builder, on the lighthouse at the time, perished with it. It was followed in 1708 by a second wooden tower, constructed by John Rudyerd, which was destroyed by fire in 1755. Rudyerd's lighthouse was followed by John Smeaton's famous masonry tower in 1759. Smeaton, a professional engineer, embodied an important new principle in its construction whereby masonry blocks were dovetailed together in an interlocking pattern. Despite the dovetailing feature, the tower largely relied on its own weight for stability—a principle that required it to be larger at the base and tapered toward the top. Instead of a straight conical taper, though, Smeaton gave the structure a curved profile. Not only was the curve visually attractive, but it also served to dissipate some of the energy of wave impact by directing the waves to sweep up the walls.

Towers at Eddystone Rocks

Owing to the undermining of the foundation rock, Smeaton's tower had to be replaced in 1882 by the present lighthouse, constructed on an adjacent part of the rocks by Sir James Douglass, engineer-in-chief of Trinity House. In order to reduce the tendency of waves to break over the lantern during severe storms (a problem often encountered with Smeaton's tower), Douglass had the new tower built on a massive cylindrical base that absorbed some of the energy of incoming seas. The upper portion of Smeaton's lighthouse was dismantled and rebuilt on Plymouth Hoe, where it still stands as a monument; the lower portion or "stump" can still be seen on the Eddystone Rocks.

Following the Eddystone, masonry towers were erected in similar open-sea sites, which include the Smalls, off the Welsh coast; Bell Rock in Scotland; South Rock in Ireland; and Minots Ledge off Boston, Mass., U.S. The first lighthouse of the North American continent, built in 1716, was on the island of Little Brewster, also off Boston. By 1820 there were an estimated 250 major lighthouses in the world.

Modern lighthouses

CONSTRUCTION

While masonry and brick continue to be employed in lighthouse construction, concrete and steel are the most

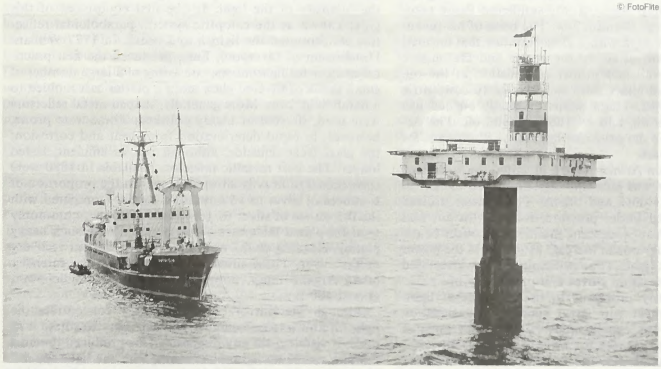


Figure 44: The Royal Sovereign Lighthouse, a concrete "float out" lighthouse in the English Channel.

widely used materials. Structurally well suited and reasonably cheap, concrete especially lends itself to aesthetically pleasing designs for shore-based lighthouses.

Modern construction methods have considerably facilitated the building of lighthouses in the open sea. On soft ground, the submerged caisson method is used, a system applied first in 1885 to the building of the Roter Sand Lighthouse in the estuary of the Weser River in Germany and then to the Fourteen Foot Bank light in the Delaware Bay, U.S. With this method, a steel caisson or open-ended cylinder, perhaps 40 feet in diameter, is positioned on the seabed. By excavation of sand, it is sunk into the seabed to a depth of possibly 50 feet. At the same time, extra sections are added to the top as necessary so that it remains above high water level. The caisson is finally pumped dry and filled with concrete to form a solid base on which the lighthouse proper is built.

"Float out" lighthouses

Where the seabed is suitable, it is possible to build a "float out" lighthouse, consisting of a cylindrical tower on a large concrete base that can be 50 feet in diameter. The tower is constructed in a shore berth, towed out to position, and then sunk to the seabed, where the base is finally filled with sand. Weighing 5,000 tons (4.5 million kilograms) or more, these towers rely on their weight for stability and require a leveled, prepared seabed. For greater stability during towing, the cylindrical tower itself often consists of two or more telescopic sections, raised to full height by hydraulic jacks after being founded on the seabed. This design was pioneered largely in Sweden.

Another design, which is more independent of seabed conditions, is the conventional steel-piled structure used for offshore oil and gas rigs. Piles may be driven as deep as 150 feet into the seabed, depending on the underlying strata. The United States has built about 15 light towers of this type, one prominent example being Ambrose Light off New York.

Helicopters are widely employed in the servicing and maintenance of offshore towers, so that modern designs normally include a helipad. In fact, older cylindrical masonry structures of the previous era—including the Eddystone tower—have had helipads fitted above their lanterns.

ILLUMINANTS

Wood fires were not discontinued until 1800, though after about 1550 coal, a more compact and longer-burning fuel, was increasingly favoured, particularly in northwestern Europe. A lighthouse in those days could consume 300 tons or more of coal a year. In full blaze, the coal fire was far superior to other forms of lighting, preferred by mariners to oil or candles. The disadvantage of both coal fires and early oil lamps and candles was the prodigious amount of smoke produced, which resulted in rapid blackening of the lantern panes, obscuring the light.

Oil lamps. In 1782 a Swiss scientist, Aimé Argand, invented an oil lamp whose steady smokeless flame revolutionized lighthouse illumination. The basis of his invention was a circular wick with a glass chimney that ensured an adequate current of air up the centre and the outside of the wick for even and proper combustion of the oil. Eventually, Argand lamps with as many as 10 concentric wicks were designed. These lamps originally burned fish oil, later vegetable oil, and by 1860 mineral oil. The Argand lamp became the principal lighthouse illuminant for more than 100 years.

In 1901 the Briton Arthur Kitson invented the vaporized oil burner, which was subsequently improved by David Hood of Trinity House and others. This burner utilized kerosene vaporized under pressure, mixed with air, and burned to heat an incandescent mantle. The effect of the vaporized oil burner was to increase by six times the power of former oil wick lights. (The principle is still widely used for such utensils as camp stoves and pressure lamps.)

Gas lamps. Early proposals to use coal gas at lighthouses did not meet with great success. A gasification plant at the site was usually impracticable, and most of the lights were too remote for a piped supply. However, acetylene gas, generated in situ from calcium carbide and water, came into use around the turn of the 20th century, and its use increased following the introduction of the dissolved acetylene process, which by dissolving the acetylene in acetone made it safe to compress for storage.

Use of acetylene gas

Acetylene gas as a lighthouse illuminant had a profound influence on the advancement of lighthouse technology, mainly through the work of Gustaf Dalén of Sweden, who pioneered its application between 1900 and 1910. Burned either as an open flame or mixed with air in an incandescent mantle, acetylene produced a light equal to that of oil. Its great advantage was that it could be readily controlled; thus, for the first time automatic unattended lights were possible. Dalén devised many ingenious mechanisms and burners, operating from the pressure of the gas itself, to exploit the use of acetylene. Most of the equipment he designed is still in general use today. One device is an automatic mantle exchanger that brings a fresh mantle into use when the previous one burns out. Another, economizing on gas, was the "sun valve," an automatic day-night switch capable of extinguishing the light during the day. The switch utilized the difference in heat-absorbing properties between a dull black surface and a highly polished one, producing a differential expansion arranged by suitable mechanical linkage to control the main gas valve.

The acetylene system facilitated the establishment of many automatic unattended lighthouses in remote and inaccessible locations, normally requiring only an annual visit to replenish the storage cylinders and overhaul the mechanism. Liquefied petroleum gas, such as propane, has also found use as an illuminant, although both oil and gas lamps have largely been superseded by electricity.

Electric lamps. Electric illumination in the form of carbon arc lamps was first employed at lighthouses at an early date, even while oil lamps were still in vogue. The first of these was at Dungeness, Eng., in 1862, followed by a number of others. The majority of these, however, were eventually converted to oil, since the early arc lamps were difficult to control and costly to operate. In 1913 the Helgoland Lighthouse in the North Sea off Germany was equipped with arc lamps and searchlight mirrors to give a light of 38 million candlepower, the most powerful lighthouse in the world at that time.

Electric-filament lamps

The electric-filament lamp, which came into general use in the 1920s, is now the standard illuminant. Power output ranges from about 1,500 watts for the largest structures down to about 5 watts for buoys and minor beacons. Most lamps are of the tungsten-halogen type for better efficiency and longer life. As new types of electric lamps become available—for example, compact source discharge tube lamps—they are adopted for lighthouse use wherever suitable.

OPTICAL EQUIPMENT

Paraboloidal mirrors. With the advent of the Argand lamp, a reliable and steady illuminant, it became possible to develop effective optical apparatuses for increasing the intensity of the light. In the first equipment of this type, known as the catoptric system, paraboloidal reflectors concentrated the light into a beam. In 1777 William Hutchinson of Liverpool, Eng., produced the first practical mirrors for lighthouses, consisting of a large number of small facets of silvered glass set in a plaster cast molded to a paraboloid form. More generally, shaped metal reflectors were used, silvered or highly polished. These were prone, however, to rapid deterioration from heat and corrosion; the glass facet reflector, although not as efficient, lasted longer. The best metallic reflectors available in 1820 were constructed of heavily silvered copper in the proportion of 6 ounces of silver to 16 ounces of copper (compared with the 0.5 ounce of silver to 16 ounces of copper commonly used for plated tableware of the period). With such heavy plating, cleaning cloths were kept for subsequent recovery of the silver. These mirrors could increase the intensity of an Argand lamp, nominally about five candlepower, almost 400 times.

Although the mirror could effectively concentrate the light into an intense beam, it was necessary to rotate it to make it visible from any direction. This produced the now familiar revolving lighthouse beam, with the light appearing as a series of flashes. Mariners were not favourably disposed to these early flashing lights, contending that a

fixed steady light was essential for a satisfactory bearing. However, the greatly increased intensity and the advantage of using a pattern of flashes to identify the light gradually overcame their objections. The first revolving-beam lighthouse was at Carlsten, near Marstrand, Swed., in 1781.

The Fresnel dioptric system

Rectangular and drum lenses. In 1828 Augustin Fresnel of France produced the first apparatus using the refracting properties of glass, now known as the dioptric system. On a lens panel he surrounded a central bull's-eye lens with a series of concentric glass prismatic rings. The panel collected light emitted by the lamp over a wide horizontal angle and also the light that would otherwise escape to the sky or to the sea, concentrating it into a narrow, horizontal pencil beam. With a number of lens panels rotating around the lamp, he was then able to produce several revolving beams from a single light source, an improvement over the mirror that produces only a single beam. To collect more of the light wasted vertically, he added triangular prism sections above and below the main lens, which both refracted and reflected the light. By doing this he considerably steepened the angle of incidence at which rays shining up and down could be collected and made to emerge horizontally. Thus emerged the full Fresnel catadioptric system, the basis of all lighthouse lens systems today. To meet the requirement for a fixed all-around light, Fresnel modified his principle by producing a cylindrical drum lens, which concentrated the light into an all-around fan beam. Although not as efficient as the rectangular panel, it provided a steady, all-around light. Small drum lenses, robust and compact, are widely used today for buoy and beacon work, eliminating the complication of a rotating mechanism; instead of revolving, their lights are flashed on and off by an electronic code unit.

Prior to Fresnel's invention the best mirror systems could produce a light of about 20,000 candlepower with an Argand lamp. The Fresnel lens system increased this to 80,000 candlepower, roughly equivalent to a modern automobile headlamp; with the pressure oil burner, intensities of up to 1,000,000 candlepower could be achieved. For a light of this order, the burner mantle would measure 4 inches (100 millimetres) in diameter. The rotating lens system would have four large Fresnel glass lens panels, 12 feet high, mounted about four feet from the burner on a revolving lens carriage. The lens carriage would probably weigh five tons, about half of it being the weight of the glass alone. The rotating turntable would float in a circular cast-iron trough containing mercury. With this virtually frictionless support bearing, the entire assembly could be smoothly rotated by weight-driven clockwork. If the illuminant was acetylene gas, the lens rotation could be driven by gas pressure.

Modern rotating lamps

Installations of this type are still in common use, although many have been converted to electric lamps with electric-motor drives. Modern lens equipment of the same type is much smaller, perhaps 30 inches (75 centimetres) high, mounted on ball bearings and driven by an electric motor. With a 250-watt lamp, illumination of several hundred thousand candlepower can be readily obtained. Lens panels can be molded in transparent plastic, which is lighter and cheaper. Drum lenses are also molded in plastic. In addition, with modern techniques, high-quality mirrors can be produced easily and cheaply.

INTENSITY, VISIBILITY, AND CHARACTER OF LIGHTS

Geographic range and luminous range. The luminous intensity of a light, or its candlepower, is expressed in international units called candelas. Intensities of lighthouse beams can vary from thousands to millions of candelas. The range at which a light can be seen depends upon atmospheric conditions and elevation. Since the geographic horizon is limited by the curvature of the Earth, it can be readily calculated for any elevation by standard geometric methods. In lighthouse work the observer is always assumed to be at a height of 15 feet, although on large ships he may be 40 feet above the sea. Assuming a light at a height of 100 feet, the range to an observer at 15 feet above the horizon will be about 16 nautical miles. This is known as the geographic range of the light. (One nautical mile, the distance on the Earth's surface traversed by one minute of arc longitude or latitude, is equivalent to 1.15 statute miles or 1.85 kilometres.)

The luminous range of a light is the limiting range at which the light is visible under prevailing atmospheric conditions and disregarding limitations caused by its height and the Earth's curvature. A very powerful light, low in position, can thus have a clear-weather luminous range greater than that when first seen by the mariner on the horizon. Powerful lights can usually be seen over the horizon because the light is scattered upward by particles of water vapour in the atmosphere; this phenomenon is known as the loom of the light.

Atmospheric conditions have a marked effect on the luminous range of lights. They are defined in terms of a transmission factor, which is expressed as a percentage up to a maximum of 100 percent (representing a perfectly clear atmosphere, never attained in practice). Clear weather in the British Isles corresponds to about 80 percent transmission, but in tropical regions it can rise to 90 percent, increasing the luminous range of a 10,000-candela light from 18 to 28 nautical miles. Conversely, in mist or haze at about 60 percent transmission, a light of 1,000,000 candelas would be necessary to maintain a luminous range of 18 nautical miles. In dense fog, with visibility down to 100 yards or metres, a light of 10,000,000,000 candelas could scarcely be seen at half a nautical mile. Because average clear-weather conditions vary considerably from one region of the world to another, luminous ranges of all lighthouses by international agreement are quoted in an arbitrary standard clear-weather condition corresponding to a daytime meteorological visibility of 10 nautical miles, or 74 percent transmission. This is known as the nominal range of a light. Mariners use conversion tables to determine the actual luminous range in the prevailing visibility.

Because lights of very great intensity yield diminishing returns in operational effectiveness, most very high-powered lights have been abandoned. A maximum of 100,000 candelas, with a clear-weather range of 20 nautical miles, is generally considered adequate. Nevertheless, there are still some very high-powered lights, which for special reasons may have to be visible at a distance in daylight.

Characteristic flash patterns

Identification. Most lighthouses rhythmically flash or eclipse their lights to provide an identification signal. The particular pattern of flashes or eclipses is known as the character of the light, and the interval at which it repeats itself is called the period. The number of different characters that can be used is restricted by international agreement through the International Association of Lighthouse Authorities in Paris, to which the majority of maritime nations belong. The regulations are too lengthy to quote in full, but essentially a lighthouse may display a single flash, regularly repeated at perhaps 5-, 10-, or 15-second intervals. This is known as a flashing light. Alternatively, it may exhibit groups of two, three, or four flashes, with a short eclipse between individual flashes and a long eclipse of several seconds between successive groups. The whole pattern is repeated at regular intervals of 10 or 20 seconds. These are known as group-flashing lights. In another category, "occulting" lights are normally on and momentarily extinguished, with short eclipses interrupting longer periods of light. Analogous to the flashing mode are occulting and group-occulting characters. A special class of light is the isophase, which alternates eclipses and flashes of exactly equal duration.

Steadily burning lights are called fixed lights. For giving mariners accurate directional information in ports, harbours, and estuarial approaches, fixed directional lights display sharply defined red and green sectors. Another sensitive and very accurate method of giving directional instruction is by range lights, which are two fixed lights of different elevation located about half a nautical mile apart. The navigator steers the vessel to keep the two lights aligned one above the other. Laser lights are also employed in this role.

Another use for fixed lights is the control of shipping at harbour entrances. A traffic signal consists of a vertical column of high-powered red, green, and yellow projector lights that are visible in daylight.

The daymark requirement of a lighthouse is also impor-

tant; lighthouse structures are painted to stand out against the prevailing background. Shore lighthouses are usually painted white for this purpose, but in the open sea or against a light background conspicuous bands of contrasting colours, usually red or black, are utilized.

SOUND SIGNALS

The limitations of purely visual navigation very early led to the idea of supplementary audible warning in lighthouses. The first sound signals were explosive. At first cannon were used, and later explosive charges were attached to retractable booms above the lantern and detonated electrically. Sometimes the charges contained magnesium in order to provide an accompanying bright flare. Such signals could be heard up to four nautical miles away. Bells also were used, the striker being actuated by weight-driven clockwork or by a piston driven by compressed gas (usually carbon dioxide). Some bells were very large, weighing up to one ton.

The siren and the diaphone

Compressed air. About the turn of the 20th century, compressed-air fog signals, which sounded a series of blasts, were developed. The most widely used were the siren and the diaphone. The siren consisted of a slotted rotor revolving inside a slotted stator that was located at the throat of a horn. The diaphone worked on the same principle but used a slotted piston reciprocating in a cylinder with matching ports. The largest diaphones could be heard under good conditions up to eight nautical miles away. Operating pressures were at 2 to 3 bars (200 to 300 kilopascals), and a large diaphone could consume more than 50 cubic feet (approximately 1.5 cubic metres) of air per second. This required a large and powerful compressing plant, 50 horsepower or more, with associated air-storage tanks.

A later compressed-air signal was the tyfon. Employing a metal diaphragm vibrated by differential air pressure, it was more compact and efficient than its predecessors.

Electricity. Modern fog signals are almost invariably electric. Like the tyfon, they employ a metal diaphragm, but in the electric signal they vibrate between the poles of an electromagnet that is energized by alternating current from an electronic power unit. Powers range from 25 watts to 4 kilowatts, with ranges from half a nautical mile to five nautical miles. Note frequencies lie between 300 and 750 hertz. Emitters can be stacked vertically, half a wavelength apart, in order to enhance the sound horizontally and reduce wasteful vertical dispersion.

Effective range. Propagation of sound in the open air is extremely haphazard, owing to the vagaries of atmospheric conditions. Wind direction, humidity, and turbulence all have an effect. Vertical wind and temperature gradients can bend the sound up or down; in the latter case it can be reflected off the sea, resulting in shadow zones of silence. The range of audibility of a sound signal is therefore extremely unpredictable. Also, it is difficult to determine with any precision the direction of a signal, especially from the bridge of a ship in fog.

RADIO AIDS

Sophisticated and complex radio navigation systems such as Decca and Loran, and satellite-based global positioning systems such as Navstar, are not properly within the field of lighthouses (see NAVIGATION). Radio and radar beacons, on the other hand, provide the equivalent of a visual seamark that is unaffected by visibility conditions.

Radio beacons. Radio beacons, which first appeared in the 1920s, transmit in the frequency band of 285–315 kilohertz. In a characteristic signal lasting one minute, the station identification, in Morse code, is transmitted two or three times, followed by a period of continuous transmission during which a bearing can be taken by a ship's direction-finding receiver. Bearing accuracy averages better than 3°. The frequency of transmission varies in different parts of the world. In the busy waters of Europe, radio beacons transmit continuously on a number of different channels within the allotted frequency band.

Since the development of satellite-based positioning systems in the 1970s and '80s, the early importance of radio beacons as an aid for marine navigators has diminished considerably—although they have acquired a second important role in broadcasting corrections for improving the accuracy of the satellite systems. The principal users of radio beacons are now small-craft operators, particularly recreational sailors.

Radar responders. Radar-responder beacons are employed in other fields, such as aviation; in marine navigation they are called racons. A racon transmits only in response to an interrogation signal from a ship's radar, at the time when the latter's rotating scanner bears on it. During this brief period, the racon receives some 10 radar pulses, in reaction to which it transmits back a coded reply pulse that is received and displayed on the ship's radar screen. Racons operate on both marine radar bands of 9,300–9,500 megahertz and 2,900–3,100 megahertz. A racon can greatly increase the strength of the echo from a poor radar target, such as a small buoy; it is also helpful in ranging on and identifying positions on inconspicuous and featureless coastlines and in identifying offshore oil and gas rigs. Racons

The first racons came into use in 1966, and there are now many hundreds in service. Early racons, employing vacuum-tube technology, were large and required several hundred watts of power. Modern racons, using solid-state electronics, are compact and light, typically 16 by 24 inches in area and 20 to 35 pounds (10 to 15 kilograms) in weight. They draw an average of one watt in power from low-voltage batteries.

Passive radar echo enhancers are also used on poor targets, such as buoys. They are made up of flat metal sheets joined into polyhedral shapes whose geometry is such as to reflect as much of the radar pulse as possible. A typical array, some 28 by 24 inches overall, can have an echoing area equivalent to that of a flat sheet with an area of some 1,600 square feet (150 square metres).

AUTOMATION

The acetylene-gas illumination system, being fully automatic and reliable, enabled automatic lights to be operated early on. Its main use today is in buoys, which inherently have to operate unattended. Automation on a large scale, bringing considerable savings in operating costs, came after the advent of electrical equipment and technology and the demise of compressed-air fog signals. Unattended lights are now designed to be automatic and self-sustaining, with backup plant brought on-line automatically upon failure of any component of the system. The status of the station is monitored from a remote control centre via landline, radio, or satellite link. Power is provided from public electricity supplies (where practicable), with backup provided by diesel generators or storage batteries. Where solar power with storage batteries is used, the batteries must have sufficient capacity to operate the light during the hours of darkness. In tropical and subtropical regions, day and night are of approximately equal duration throughout the year, but in temperate and polar regions the days become longer and the nights shorter during the summer, and vice versa in winter. In these areas, solar power has to operate on an annual "balance sheet" basis, with excess charge being generated and stored in large batteries during the summer so that a reserve can be drawn upon in winter. Canada and Norway successfully operate solar-powered lights of this type in their Arctic regions.

FLOATING LIGHTS

Floating lights (*i.e.*, lightships and buoys) have an important function in coastal waters, guiding both passing ships and those making for or leaving harbour. They have the great advantage of mobility and can readily be redeployed to meet changed conditions. For example, submerged hazards such as sandbanks can move over the years under the influence of the sea, and, for vessels of very deep draft, safe channels must be correctly marked around these hazards at all times.

Lightships. Lightships originated in the early 17th century, arising from the need to establish seamarks in positions where lighthouses were at that time impracticable. The first lightship, established in 1732 at Nore Sand in the Thames estuary, was rapidly followed by others. These

early vessels were small converted merchant or fishing vessels showing lanterns suspended from crossarms at the masthead. Not until 1820 were vessels built specifically as lightships.

Modern lightships are in most cases an alternative to costly seabed structures. Used to mark the more important hazards and key positions in traffic patterns, they are capable of providing a range of powerful aids. Power is provided by diesel generators. Lightships vary in size but can be up to 150 feet in length, 25 feet in beam, and 500 tons in displacement. They are not normally self-propelled, most often being towed into position and moored by a single chain and anchor. They need to be withdrawn for overhaul every two or three years.

Modern unattended lightships

Lightships are costly items to operate and to maintain and are therefore prime candidates for automation. All lightships are now unattended, and the power of their lights and fog signals has been downgraded to a more appropriate level—*e.g.,* 10,000 candelas for the lights, giving a luminous range of approximately 15 nautical miles, and sound signals with a range of two nautical miles. New lightships are of similar construction to the older type but are smaller—60 feet or less in length. The smallest sizes, 30 feet or less, are intended for sheltered waters and are often known as light floats.

Buoys. Buoys are used to mark safe channels, important reference points, approaches to harbours, isolated dangers and wrecks, and areas of special significance. They also mark traffic lanes in narrow and congested waters where traffic routing is in force (*i.e.,* where ships are being routed into designated lanes with entry and exit points).

Structure and operation. Usually constructed of quarter-inch steel plate, buoys vary in diameter from three to six feet and can weigh as much as eight tons. They are moored to a two- or three-ton concrete or cast-iron sinker by a single length of chain, which is normally about three times as long as the depth of water. Size and type depend on the application. Buoys in deep water need to be large in order to provide buoyancy sufficient to support the extra weight of chain. Stability is provided by a circular "skirt keel," with possibly a ballast weight. The light is mounted on a superstructure some 10 feet above the waterline. For more powerful lights on the open sea or deep water, the light may be 15 or 20 feet above the waterline in order to increase the range. Known as "high focal plane" buoys, for stability they require a cylindrical tail tube extending downward some 10 feet from the bottom of the hull.

Buoys are also manufactured from fibreglass-reinforced plastic. They have the advantage of light weight, hence lighter moorings (often of synthetic cable), ease of handling, and resistance to corrosion. Fibreglass buoys are generally confined to sheltered waters.

In addition to the light, a buoy may be fitted with a racon, radar reflector, and low-power fog signal. In earlier times acetylene gas was the only practicable illuminant, which restricted the power of the light. Modern electric buoy lights range in power from a few hundred candelas up to the region of 1,000 candelas, giving ranges of eight nautical miles or so. The lighting equipment consists of a drum lens, usually made of plastic and between 4 and 12 inches in diameter, along with a low-voltage lamp of 5 to 60 watts. Power can be provided by expendable primary batteries, which need to be replaced every year or two. In order to increase the service interval and also to accommodate more powerful lights, rechargeable batteries with onboard generators are used. Some tail-tube buoys, which tend to oscillate vertically with the motion of the sea, generate power from the oscillating water column in the tube. The water column produces an oscillating air column, which in turn drives a small air-turbine generator. However, the vast majority of buoys are solar-powered, and buoys constitute the major part of solar-powered lights.

Buoyage systems. A buoy's colour and profile and the colour and flash pattern of its light convey information to navigators. The beginnings of a unified system of buoy marking emerged in 1889, but it was not until 1936 that a worldwide unified system was agreed upon at the League of Nations in Geneva. World War II began before the agreement could be fully ratified, and in the aftermath of the war little progress was made. At one time there were nine different systems in use. By 1973, however, a number of spectacular marine accidents urged the international community into action, and by 1980 a new unified system had been agreed upon by 50 maritime countries.

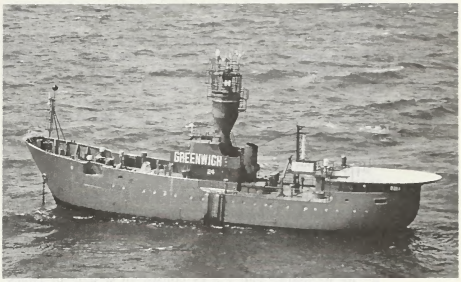


Figure 45: An automated, unattended lightship in the English Channel.

© FotoFlite

Maintained by the International Association of Lighthouse Authorities, the Maritime Buoyage System applies two nearly identical standards to two regions. Region A comprises Europe, Australia, New Zealand, Africa, the Persian Gulf, and most Asian states. Region B includes the Americas, Japan, Korea, and the Philippines. In both regions, the buoyage systems divide buoys into Lateral, Cardinal, and associated classes. Lateral buoys are used to

The Maritime Buoyage System

Reproduced from Kenneth Sutton-Jones, *Pharos: The Lighthouse Yesterday, Today, and Tomorrow*, copyright © 1985 Kenneth Sutton-Jones, used by permission from Michael Russell Publishing Ltd.

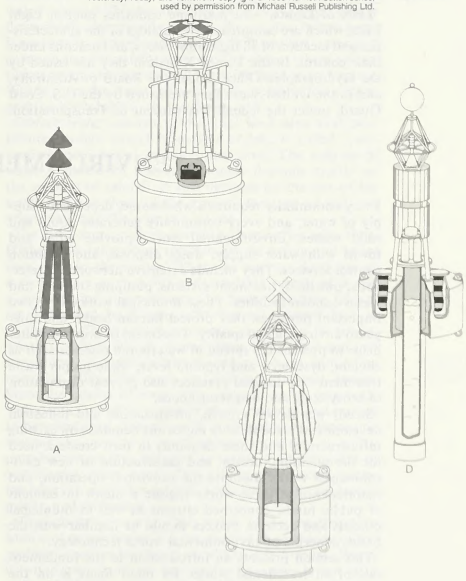


Figure 46: *Four lighted buoys and their power systems.* (A) Gas-operated buoy, with acetylene cylinders located within the buoy body; (B) solar-powered buoy, with electricity generated by solar panels and stored in a secondary battery; (C) electric-powered buoy, with batteries stored in the buoy body; (D) wave-powered buoy, with an air-driven generator activated by an oscillating column of water.

mark channels. In region A a can-profile (*i.e.*, cylindrical) red buoy with a red light indicates the port (left) side of the channel when proceeding in the direction of buoyage, while a conical green buoy indicates the starboard (right) side. The direction of buoyage is the direction taken when approaching a harbour from seaward or when following a clockwise direction around a landmass. Where necessary, it is indicated by arrows on charts. In region B the marking is reversed—*i.e.*, red is to starboard when returning to harbour, and green is to port.

Cardinal buoys indicate the deepest water in an area or the safest side around a hazard. They have a plain hull with a lattice or tubular superstructure that is surmounted by a distinctive "topmark." There are four possible topmarks, corresponding to the four cardinal points of the compass—*i.e.*, north, south, east, and west. The safest navigable water lies to the side of the buoy indicated by its topmark. The marks consist of a pair of cones arranged one above the other in the following configurations: both points upward (north), both points downward (south), base-to-base (east), and point-to-point (west). The cones are painted black, and the buoys are coloured from top to waterline with two or three horizontal bands of black and yellow.

Buoys indicating an isolated danger with safe water all around carry two separated spheres and are painted with alternating horizontal red and black bands. Safe-water buoys, marking an area of safe water, carry a single red sphere and vertical red and white stripes.

Special marks are buoys that indicate other areas or features such as pipelines, prohibited zones, or recreational areas. These buoys are all yellow, with an X-shaped topmark and a yellow light. They may also have a hull shape indicating the safe side of passage.

NATIONAL LIGHTHOUSE SYSTEMS

Lists of Lights. All maritime countries publish Light Lists, which are comprehensive catalogs of the characteristics and location of all lightships, buoys, and beacons under their control. In the United Kingdom they are issued by the Hydrographic Office, under the Board of Admiralty, and in the United States they are issued by the U.S. Coast Guard, under the federal Department of Transportation. Changes in status are disseminated by Notices to Mariners, which update lights lists and charts. Urgent information is broadcast at scheduled times on dedicated radio channels or satellite link. All this information is promulgated in a standard format recommended by the International Hydrographic Organization, based in Monaco.

Lighthouse administration. In most countries lighthouse administration comes under a department of central government. It is usually financed out of general taxation, but it is sometimes funded from a levy on shipping that may on occasion be supplemented by the central government. This type of funding applies to lights intended for general navigation. Lights provided by ports and harbours specifically for ships using the port are paid for separately by port and harbour dues.

In the United States, lighthouses and other aids are administered by the Aids to Navigation Division of the U.S. Coast Guard. The Coast Guard has at least 3,000 lighthouses and a further 6,000 minor lights, plus 4,000 lighted buoys. In 1989, expenditure, derived from general taxation, was some $450 million (£280 million).

The Corporation of Trinity House

In England and Wales, lighthouses are administered by the Corporation of Trinity House, an autonomous non-governmental agency. Trinity House evolved from a royal charter granted in 1514 to a medieval guild or fraternity of Thames river pilots based in the parish of Deptford Stronde. Its charter was later extended to the provision of seamarks. At that time most lights were operated by private owners, who, under concessions purchased from the crown, had the necessary authority to collect payment in the form of light dues. Because of increasing dissatisfaction with the level of charges and with poor service, by act of Parliament in 1836 all privately operated lights in England and Wales were bought and transferred to Trinity House. Seamarks off Scotland were placed under the authority of the Northern Lighthouse Board in Edinburgh, and Irish seamarks were placed under the Commissioners for Irish Lights in Dublin. These three authorities, which today operate a total of 1,100 lighthouses, lightships, buoys, and beacons, still share the pooled user charges, known as the General Lighthouse Fund, for the whole of the British Isles. In 1990 the fund was in the region of £55 million ($105 million). (I.C.C.)

ENVIRONMENTAL WORKS

Every community requires a wholesome, dependable supply of water, and every community generates liquid and solid wastes. Environmental works provide cities and towns with water supply, waste disposal, and pollution control services. They include extensive networks of reservoirs, pipelines, treatment systems, pumping stations, and waste disposal facilities. These municipal works serve two important purposes: they protect human health and safeguard environmental quality. Treatment of drinking water helps to prevent the spread of waterborne diseases such as cholera, dysentery, and typhoid fever, while proper waste treatment and disposal practices also prevent degradation of ecosystems and neighbourhoods.

Steady population growth, urbanization, and industrial development place steadily increasing demands on existing infrastructure, and these demands in turn create a need for the planning, design, and construction of new environmental works. Because the provision, operation, and maintenance of these works require a major investment of public funds, concerned citizens as well as municipal officials and decision makers should be familiar with the basics concepts of environmental works technology.

This section presents an introduction to the fundamentals of environmental works. Its main focus is on the modern facilities and systems that provide communities with water, dispose of waste, and prevent pollution.

Municipal water supply

Of all municipal services, provision of potable water is perhaps the most vital. All people depend on water for drinking, cooking, washing, carrying away wastes, and other domestic needs. Water supply systems must also meet requirements for public, commercial, and industrial activities. During times of droughts, floods, earthquakes, or other emergencies, vigorous efforts must be made to maintain public water supplies.

HISTORICAL BACKGROUND

Developments in supply systems. Water was an important factor in the location of the earliest settled communities, and the evolution of public water supply systems is tied directly to the growth of cities. In the development of water resources beyond their natural condition in rivers, lakes, and springs, the digging of shallow wells was probably the earliest innovation. As the need for water increased and tools were developed, wells were made deeper. Brick-lined wells were built by city dwellers in the Indus River basin as early as 2500 BC, and wells more than 1,600 feet (almost 500 metres) deep are known to have been used in ancient China. Construction of *qanāt*s, slightly sloping tunnels driven into hillside that contain groundwater, probably originated in northwestern Persia (now Armenia) about 700 BC. From the hillsides the water was conveyed by gravity in open channels to nearby towns or cities. The use of *qanāt*s became widespread throughout the region, and some are still in existence. Until 1933 the Iranian capital city, Tehrān, drew its entire water supply from a system of *qanāt*s.

The need to channel water supplies from distant sources was an outcome of the growth of urban communities. Among the most notable of ancient water conveyance sys-

The Roman aqueducts

tems are the aqueducts built between 312 BC and AD 455 throughout the Roman Empire. Some of these impressive works are still in existence. The writings of Sextus Julius Frontinus (who was appointed superintendent of Roman aqueducts in AD 97) provide information about the design and construction of the 11 major aqueducts that supplied Rome itself. Extending from a distant spring-fed area, a lake, or a river, a typical Roman aqueduct included a series of underground and aboveground channels. The longest was the Aqua Marcia, built in 144 BC. Its source was about 23 miles (37 kilometres) from Rome. The aqueduct itself was 57 miles long, however, because it had to meander along land contours in order to maintain a steady flow of water. For about 50 miles the aqueduct was underground in a covered trench, and only for the last 7 miles was it carried above ground on an arcade. In fact, most of the combined length of the aqueducts supplying Rome (about 260 miles) were built as covered trenches or tunnels. When crossing a valley they were supported by arcades comprising one or more levels of massive granite piers and impressive arches.

The aqueducts ended in Rome at distribution reservoirs, from which the water was conveyed to public baths or fountains. A few very wealthy or privileged citizens had water piped directly into their homes, but most of the people carried water in containers from a public fountain. Water was running constantly, the excess being used to clean the streets and flush the sewers. Ancient aqueducts and pipelines were not capable of withstanding much pressure. Channels were constructed of cut stone, brick, rubble, or rough concrete. Pipes were made of drilled stone or of hollowed wooden logs; clay and lead pipes were also used.

During the Middle Ages there was no notable progress in the methods or materials used to convey and distribute water. Cast-iron pipes with joints capable of withstanding high pressures were not used very much until the early 19th century. The steam engine was first applied to water pumping operations at about that time, making it possible for all but the smallest communities to have drinking water supplied directly to individual homes. Asbestos cement, ductile iron, reinforced concrete, and steel came into use as materials for water-supply pipelines in the 20th century.

Developments in water treatment. In addition to quantity of supply, water quality is also of concern. Even the ancients had an appreciation for the importance of water purity. Sanskrit writings from as early as 2000 BC tell how to purify foul water by boiling and filtering. But it was not until the middle of the 19th century that a direct link between polluted water and disease (cholera) was proved. And it was not until the end of that same century that the German bacteriologist Robert Koch proved the germ theory of disease, establishing a scientific basis for the treatment and sanitation of drinking water.

Elimination of waterborne disease

Water treatment is the alteration of a water source in order to achieve a quality that meets specified goals. At the end of the 19th century and the beginning of the 20th, the main goal was elimination of deadly waterborne diseases. The treatment of public drinking water to remove pathogenic, or disease-causing, microorganisms began about that time. Treatment methods included sand filtration as well as the use of chlorine for disinfection. The virtual elimination of diseases such as cholera and typhoid in developed countries proved the success of this water treatment technology.

In developing countries, waterborne disease is still the principal water quality concern. In industrialized nations, however, concern has shifted to the chronic health effects related to chemical contamination. For example, trace amounts of certain synthetic organic substances in drinking water are suspected of causing cancer in humans. The added goal of reducing such health risks is seen in the continually increasing number of factors included in drinking water standards.

WATER SOURCES

Global distribution. Water is present in abundant quantities on and under the Earth's surface, but less than 1 percent of it is liquid fresh water. Most of the Earth's estimated 326 million cubic miles (1.4 billion cubic kilometres) of water is in the oceans or is frozen in polar ice caps and glaciers. Ocean water contains about 4.5 ounces per gallon (35 grams per litre) of dissolved minerals or salts, making it unfit for drinking and for most industrial or agricultural uses.

There is ample fresh water (containing less than one-eighth ounce of salt per gallon) to satisfy all human needs. It is not always available, though, at the times and places where it is needed, and it is not uniformly distributed over the Earth. In many locations, the availability of good-quality water is further reduced because of urban development, industrial growth, and environmental pollution.

Surface water and groundwater. Surface water and groundwater are both important sources for community water supply needs. Groundwater is a common source for single homes and small towns, while rivers and lakes are the usual sources for large cities. Although approximately 98 percent of liquid fresh water exists as groundwater, much of it occurs very deep in the Earth. This makes pumping very expensive, preventing the full development and use of all groundwater resources.

The hydrologic cycle. Water and water vapour are in constant circulation, powered by the energy from sunlight and gravity in a natural process called the hydrologic cycle. Water evaporates from the ocean and land surfaces, is held temporarily as vapour in the atmosphere, and falls back to the Earth's surface as precipitation. Surface water is the residue of precipitation and melted snow, called runoff. Where the average rate of precipitation exceeds the rate at which runoff seeps into the soil, evaporates, or is absorbed by vegetation, bodies of surface water such as streams, rivers, and lakes are formed. Water that infiltrates the Earth's surface becomes groundwater, slowly seeping downward into extensive layers of porous soil and rock called aquifers. Under the pull of gravity, groundwater flows slowly and steadily through the aquifer. In low areas it emerges in springs and streams. Both surface water and groundwater eventually return to the ocean, where evaporation replenishes the supply of atmospheric water vapour. Winds carry the moist air over land, precipitation occurs, and the hydrologic cycle continues.

Replenishing of aquifers

Surface water sources. The total land area that contributes surface runoff to a river or lake is called a watershed, river basin, or catchment area. The volume of water available for municipal supply depends mostly on the amount of rainfall. It also depends on the size of the watershed, the slope of the ground, the type of soil and vegetation, and the type of land use.

The flow rate or discharge of a river varies with time. Higher flow rates occur typically in the spring and lower flow rates in the winter. When the average discharge of a river is not enough for a dependable supply of water, a conservation reservoir may be built. The flow of water is blocked by a dam, allowing an artificial lake to be formed. Conservation reservoirs store water from wet-weather periods for use during times of drought and low streamflow. A water intake structure is built within the reservoir, with inlet ports and valves at several depths. Since the quality of water in a reservoir varies seasonally with depth, a multilevel intake allows water of best quality to be withdrawn. Sometimes it is advisable, for economic reasons, to provide a multipurpose reservoir. A multipurpose reservoir is designed to satisfy a combination of community water needs. In addition to drinking water, the reservoir may also provide flood control, hydroelectric power, and recreation.

Groundwater sources. The value of an aquifer as a source of groundwater is a function of the porosity of the geologic stratum, or layer, of which it is formed. Water is withdrawn from an aquifer by pumping it out of a well or infiltration gallery. An infiltration gallery typically includes several horizontal perforated pipes radiating outward from the bottom of a large-diameter vertical shaft. Wells are constructed in several ways, depending on the depth and nature of the aquifer. Wells used for public water supplies, usually more than 100 feet deep and from 4 to 12 inches (10 to 30 centimetres) in diameter, must

Wells for public water supply

penetrate large aquifers that can provide dependable yields of good-quality water. They are drilled using impact or rotary techniques and are usually lined with a metal pipe or casing to prevent contamination. The annular space around the outside of the upper portion of the casing is filled with cement grout, and a special "sanitary seal" is installed at the top to provide further protection. At the bottom of the casing, a slotted screen is attached to strain silt and sand out of the groundwater. A submersible pump driven by an electric motor can be used to raise the water to the surface. Sometimes a deep well may penetrate a confined artesian aquifer, in which case natural hydrostatic pressure can raise the water to the surface.

WATER REQUIREMENTS

Municipal water supply systems include facilities for storage, transmission, treatment, and distribution. The design of these facilities depends on the quality of the water, on the particular needs of the user or consumer, and on the quantities of water that must be processed.

Drinking water quality. Water has such a strong tendency to dissolve other substances that it is rarely found in nature in a pure condition. When it falls as rain, small amounts of oxygen and carbon dioxide gas become dissolved in it; raindrops also carry tiny dust particles. As it flows over the ground, water picks up fine soil particles, microbes, organic material, and soluble minerals. In lakes, bogs, and swamps, water may gain colour, taste, and odour from decaying vegetation and other natural organic matter. Groundwater usually acquires more dissolved minerals than does surface runoff, because of its longer direct contact with soil and rock. It may also absorb gases such as hydrogen sulfide and methane. In populated areas, the quality of surface water as well as groundwater is directly influenced by human activities and the effects of pollution.

Health concerns. Five general types of impurities are of public health concern. These are organic chemicals, inorganic chemicals, turbidity, microorganisms, and radioactive substances. Organic contaminants include various pesticides, industrial solvents, and trihalomethanes such as chloroform. Inorganic contaminants of major concern include arsenic, nitrate, fluoride, and toxic metals such as lead and mercury. All these substances can harm human health when present above certain concentrations in drinking water. A low concentration of fluoride, however, has been proved to promote dental health. Some communities add fluoride to their water for this purpose.

Turbidity refers to cloudiness caused by very small particles of silt, clay, and other substances suspended in water. Even a slight degree of turbidity in drinking water is objectionable to most people. Turbidity also interferes with disinfection by creating a possible shield for pathogenic organisms. Groundwater normally has very low turbidity owing to the natural filtration that occurs as it percolates through the soil. Surface waters, though, are often high in turbidity.

Coliform bacteria

The most important microbiological measure of drinking water quality is a group of bacteria called coliforms. Coliform bacteria normally are not pathogenic, but they are always present in the intestinal tract of humans and are excreted in very large numbers with human waste. Water contaminated with human waste always contains coliforms, and it is also likely to contain pathogens excreted by infected individuals in the community. Since it is easier to test for the presence of coliforms rather than for specific types of pathogens, coliforms are used as indicator organisms for measuring the biological quality of water. If coliforms are not found in the water, it can be assumed that the water is also free of pathogens. The coliform count thus reflects the chance of pathogens being present: the lower the coliform count, the less likely it is that pathogens are in the water.

Radioactive materials from natural as well as industrial sources can be harmful water contaminants. Wastes from uranium mining, nuclear power plants, and medical research are possible pollutants. Strontium-90 and tritium are radioactive contaminants that have been found in water as a result of nuclear weapons testing. Naturally occurring substances such as radium and radon gas are found in some groundwater sources. The danger from dissolved radon gas arises not from drinking the water but from breathing the gas after it is released into the air.

Aesthetic concerns. Colour, taste, and odour are physical characteristics of drinking water that are important for aesthetic reasons rather than for health reasons. Colour in water may be caused by decaying leaves or algae, giving it a brownish yellow hue. Taste and odour may be caused by naturally occurring dissolved organics or gases. Some well-water supplies, for example, have a rotten-egg odour caused by hydrogen sulfide gas. Chemical impurities associated with the aesthetic quality of drinking water include iron, manganese, copper, zinc, and chloride. Dissolved metals impart a bitter taste to water and may stain laundry and plumbing fixtures. Excessive chlorides give the water an objectionable salty taste.

Hardness. Another parameter of water quality is "hardness." This is a term used to describe the effect of dissolved minerals (mostly calcium and magnesium). Minerals cause deposits of scale in hot water pipes, and they also interfere with the lathering action of soap. Hard water does not harm human health, but the economic problems it causes make it objectionable to most people.

Standards. Water quality standards set limits on the concentrations of impurities allowed in water. Standards also affect the selection of raw water sources and the choice of treatment processes. The development of water quality standards began in the United States in the early 20th century. Since that time, the total number of regulated contaminants has increased as toxicological knowledge and analytical measurement techniques have improved. Modern testing methods now allow the detection of contaminants in extremely low concentrations—as low as one part contaminant per one billion parts of water or even, in some cases, per one trillion parts of water. Water quality standards are continually evolving, usually becoming more stringent. As a result, the number of regulated contaminants increases over time, and their allowable concentrations in water are lowered.

Primary and secondary standards of water quality

Drinking water regulations include two types of standards: primary and secondary. Primary standards are designed to protect public health, while secondary standards are based on aesthetic factors rather than on health effects. Primary standards specify maximum contaminant levels for many chemical, microbiological, and radiological parameters of water quality. They reflect the best available scientific and engineering judgment and take into account exposure from other sources in the environment and from foods. Turbidity is also included in the primary standards because of its tendency to interfere with disinfection. Secondary standards are guidelines or suggested maximum levels of colour, taste, odour, hardness, corrosiveness, and certain other factors.

Municipal water consumption. Water consumption in a community is characterized by several types of demand, including domestic, public, commercial, and industrial uses. Domestic demand includes water for drinking, cooking, washing, laundering, and other household functions. Public demand includes water for fire protection, street cleaning, and use in schools and other public buildings. Commercial and industrial demands include water for stores, offices, hotels, laundries, restaurants, and most manufacturing plants. There is usually a wide variation in total water demand among different communities. This variation depends on population, geographic location, climate, the extent of local commercial and industrial activity, and the cost of water.

Water use or demand is expressed numerically by average daily consumption per capita (per person). In the United States the average demand is approximately 100 gallons (380 litres) per capita per day for domestic and public needs. Overall, the average total demand is about 180 gallons per capita per day, when commercial and industrial water uses are included. (These figures do not include withdrawals from freshwater sources for such purposes as crop irrigation or cooling operations at electric power generating facilities.) Water consumption in some developing countries may average as little as 4 gallons per capita per

day; the world average is estimated to be approximately 16 gallons per person per day.

Average and peak demand

In any community, water demand varies on a seasonal, daily, and hourly basis. On a hot summer day, for example, it is not unusual for total water consumption to be as much as 200 percent of the average demand. Water consumption also varies hourly throughout the day. The peak demands in residential areas usually occur in the morning as well as early evening hours (just before and after the normal workday). Water demands in commercial and industrial districts, though, are usually uniform during the working day. Minimum water demands typically occur in the very early or predawn morning hours, when very few people use water. Civil and environmental engineers must carefully study each community's water use patterns in order to design efficient pumping and distribution systems.

WATER TREATMENT

Water in rivers or lakes is rarely clean enough for human consumption if it is not first treated or purified. Groundwater, too, often needs some level of treatment to render it potable. The primary objective of water treatment is to protect the health of the community. Potable water must, of course, be free of harmful microorganisms and chemicals, but public supplies should also be aesthetically desirable, so that consumers will not be tempted to use water from another more attractive but unprotected source. The water should be crystal clear, with almost no turbidity, and it should be free of objectionable colour, odour, or taste. For domestic supplies, water should not be corrosive, nor should it deposit troublesome amounts of scale and stains on plumbing fixtures. Industrial requirements may be even more stringent; many industries provide special treatment on their own premises.

The type and extent of treatment required to obtain potable water depends on the quality of the source. The better the quality, the less treatment is needed. Surface water usually needs more extensive treatment than does groundwater, because most streams, rivers, and lakes are polluted to some extent. Even in areas remote from human populations, surface water contains suspended silt, organic material, decaying vegetation, and microbes from animal wastes. Groundwater, on the other hand, is usually free of microbes and suspended solids because of natural filtration as the water moves through soil, though it often contains relatively high concentrations of dissolved minerals from its direct contact with soil and rock.

Water is treated in a variety of physical and chemical methods. Treatment of surface water begins with intake screens to prevent fish and debris from entering the plant and damaging pumps and other components. Conventional treatment of water primarily involves clarification and disinfection (Figure 47). Clarification removes most of the turbidity, making the water crystal clear. Disinfection, usually the final step in the treatment of drinking water, destroys pathogenic microbes. Groundwater does not often need clarification, but it should be disinfected as a precaution to protect public health. In addition to clarification and disinfection, the processes of softening, aeration, carbon adsorption, and fluoridation may be used for certain public water sources. Desalination processes are used in areas where freshwater supplies are not readily available.

Clarification. *Sedimentation.* Impurities in water are either dissolved or suspended. The suspended material reduces clarity, and the easiest way to remove it is to rely on gravity. Under quiescent (still) conditions, suspended particles that are denser than water gradually settle to the bottom of a basin or tank. This is called plain sedimentation. Long-term water storage (for more than one month) in reservoirs reduces the amount of suspended sediment and bacteria. Nevertheless, additional clarification is usually needed. In a treatment plant, sedimentation (settling) tanks are built to provide a few hours of storage or detention time as the water slowly flows from tank inlet to outlet. It is impractical to keep water in the tanks for longer periods because of the large volumes that must be treated.

Sedimentation tanks

Sedimentation tanks may be rectangular or circular in shape and are typically about 10 feet deep. Several tanks are usually provided and arranged for parallel (side-by-side) operation. Influent (water flowing in) is uniformly distributed as it enters the tank. Clarified effluent (water flowing out) is skimmed from the surface as it flows over special baffles called weirs. The layer of concentrated solids that collects at the bottom of the tank is called sludge. Modern sedimentation tanks are equipped with mechanical scrapers that continuously push the sludge toward a collection hopper, where it is pumped out.

The efficiency of a sedimentation tank for removing suspended solids depends more on its surface area than on its depth or volume. A relatively shallow tank with a large surface area will be more effective than a very deep tank that holds the same volume but has a smaller surface area. Most sedimentation tanks, though, are not less than 10

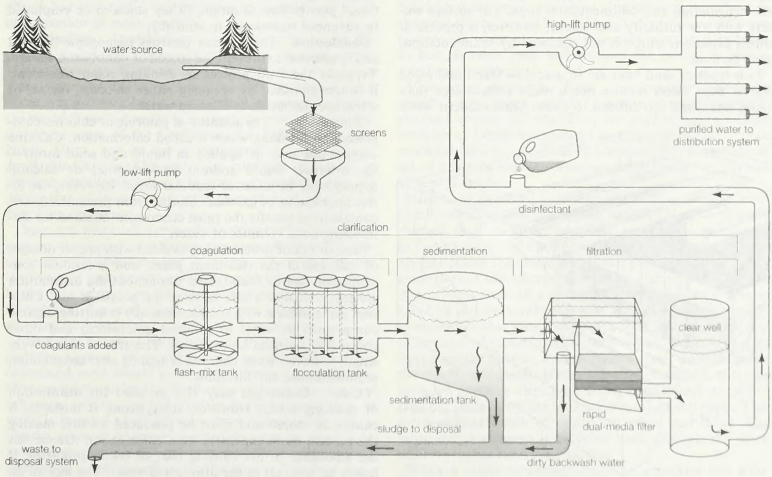


Figure 47: Basic steps in the treatment of municipal water.

feet deep, in order to provide enough room for a sludge layer and a scraper mechanism.

A technique called shallow-depth sedimentation is often applied in modern treatment plants. In this method, several prefabricated units or modules of "tube settlers" are installed near the tops of tanks in order to increase their effective surface area.

Coagulation and flocculation. Suspended particles cannot be removed completely by plain settling. Large, heavy particles settle out readily, but smaller and lighter particles settle very slowly or in some cases do not settle at all. Because of this, the sedimentation step is usually preceded by a chemical process known as coagulation. Chemicals (coagulants) are added to the water to bring the nonsettling particles together into larger, heavier masses of solids called floc. Aluminum sulfate (alum) is the most common coagulant used for water purification. Other chemicals, such as ferric sulfate or sodium aluminate, may also be used.

Coagulation is usually accomplished in two stages: rapid mixing and slow mixing. Rapid mixing serves to disperse the coagulants evenly throughout the water and to ensure a complete chemical reaction. Sometimes this is accomplished by adding the chemicals just before the pumps, allowing the pump impellers to do the mixing. Usually, though, a small flash-mix tank provides about one minute of detention time. After the flash mix, a longer period of gentle agitation is needed to promote particle collisions and enhance the growth of floc. This gentle agitation, or slow mixing, is called flocculation; it is accomplished in a tank that provides at least a half-hour of detention time. The flocculation tank has wooden paddle-type mixers that slowly rotate on a horizontal motor-driven shaft. After flocculation, the water flows into the sedimentation tanks. Some small water treatment plants combine coagulation and sedimentation in a single prefabricated steel unit called a solids-contact tank.

Filtration. Even after coagulation and flocculation, sedimentation does not remove enough suspended impurities from water to make it crystal clear. The remaining nonsettling floc causes noticeable turbidity in the water and can shield microbes from disinfection. Filtration is a physical process that removes these impurities from water by percolating it downward through a layer or bed of porous, granular material such as sand. Suspended particles become trapped within the pore spaces of the filter media, which also remove harmful protozoa and natural colour. Most surface water supplies require filtration after the coagulation and sedimentation steps. For surface waters with low turbidity and colour, however, a process of direct filtration, which is not preceded by sedimentation, may be used.

Rapid dual-media filters

Two types of sand filter are in use: slow filters and rapid filters. Slow filters require much more surface area than rapid filters and are difficult to clean. Most modern water treatment plants now use rapid dual-media filters following coagulation and sedimentation. A dual-media filter consists of a layer of anthracite coal above a layer of fine sand. The upper layer of coal traps most of the large floc, and the finer sand grains in the lower layer trap smaller impurities. This process is called in-depth filtration, as the impurities are not simply screened out or removed at the surface of the filter bed, as is the case in slow sand filters. In order to enhance in-depth filtration, so-called mixed-media filters are used in some treatment plants. These have a third layer of a fine-grained, dense mineral called garnet at the bottom of the bed.

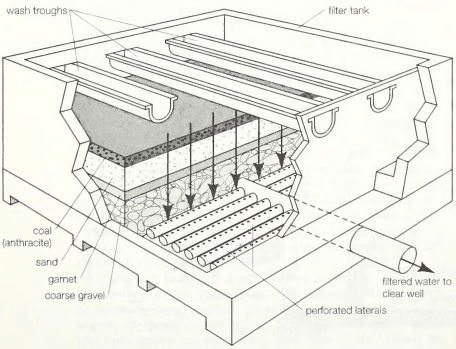


Figure 48: Schematic diagram of a rapid-filter facility.

Rapid filters are housed in boxlike concrete structures, with multiple boxes arranged on both sides of a piping gallery (Figure 48). A large tank called a clear well is usually built under the filters to hold the clarified water temporarily. A layer of coarse gravel usually supports the filter media. When clogged by particles removed from the water, the filter bed must be cleaned by backwashing. In the backwash process, the direction of flow through the filter is reversed. Clean water is forced upward through the media, expanding the filter bed slightly and carrying away the impurities in wash troughs. The backwash water is distributed uniformly across the filter bottom by an underdrain system of perforated pipes or porous tile blocks.

Because of its reliability, the rapid filter is the most common type of filter used to treat public water supplies. However, other types of filters may be used, including pressure filters, diatomaceous earth filters, and microstrainers. A pressure filter has a granular media bed, but, instead of being open at the top like a gravity-flow rapid filter, it is enclosed in a cylindrical steel tank. Water is pumped through the filter under pressure. In diatomaceous earth filters a natural powderlike material composed of the shells of microscopic organisms called diatoms is used as a filter media. The powder is supported in a thin layer on a metal screen or fabric, and water is pumped through the layer. Pressure filters and diatomaceous earth filters are used most often for industrial applications or for public swimming pools.

Microstrainers

Microstrainers consist of a finely woven stainless steel wire cloth mounted on a revolving drum that is partially submerged in the water. Water enters through an open end of the drum and flows out through the screen, leaving suspended solids behind. Captured solids are washed into a hopper when they are carried up out of the water by the rotating drum. Microstrainers are used mainly to remove algae from surface water supplies before conventional gravity-flow filtration. (They can also be employed in advanced wastewater treatment.)

Disinfection. Disinfection destroys pathogenic bacteria and is essential to prevent the spread of waterborne disease. Typically the final process in drinking water treatment, it is accomplished by applying either chlorine, ozone, or ultraviolet radiation to clarified water.

Chlorination. The addition of chlorine or chlorine compounds to drinking water is called chlorination. Chlorine compounds may be applied in liquid and solid forms—for instance, liquid sodium hypochlorite, or calcium hypochlorite in tablet or granular form. However, the direct application of gaseous chlorine from pressurized steel containers is usually the most economical method for disinfecting large volumes of water.

Taste or odour problems are avoided with proper dosages of chlorine at the treatment plant, and a residual concentration can be maintained throughout the distribution system to ensure a safe level at the points of use. Chlorine can combine with certain naturally occurring organic compounds in water to produce chloroform and other potentially harmful by-products. The risk of this is very small, however, when chlorine is applied after coagulation, sedimentation, and filtration.

Ozone. Ozone gas may also be used for disinfection of drinking water. However, since ozone is unstable, it cannot be stored and must be produced on-site, making the process more expensive than chlorination. Ozone has the advantage of not causing taste or odour problems; it leaves no residual in the disinfected water. The lack of an ozone residual, however, makes it difficult to monitor its

continued effectiveness as water flows through the distribution system.

Ultraviolet radiation. Ultraviolet radiation destroys pathogens, and its use as a disinfecting agent eliminates the need to handle chemicals. It leaves no residual, and it does not cause taste or odour problems. But the high cost of its application makes it a poor competitor with either chlorine or ozone as a disinfectant.

Additional treatment. Clarification and disinfection are the conventional processes for purifying surface water supplies. Other techniques may be used in addition, or separately, to remove certain impurities, depending on the quality of the raw water.

The lime-soda method

Water softening. Softening is the process of removing the dissolved calcium and magnesium salts that cause hardness in water. It is achieved either by adding chemicals that form insoluble precipitates or by ion exchange. Chemicals used for softening include calcium hydroxide (slaked lime) and sodium carbonate (soda ash). This lime-soda method of water softening must be followed by sedimentation and filtration in order to remove the precipitates. Ion exchange is accomplished by passing the water through columns of a natural or synthetic resin that trades sodium ions for calcium and magnesium ions. Ion exchange columns must eventually be regenerated by washing with a sodium chloride solution.

Aeration. Aeration is a physical treatment process used for taste and odour control and for removal of dissolved iron and manganese. It consists of spraying water into the air or cascading it downward through stacks of perforated trays. Dissolved gases that cause tastes and odours are transferred from the water to the air. Oxygen from the air, meanwhile, reacts with any iron and manganese in the water, forming a precipitate that is removed by sedimentation and filtration.

Carbon adsorption. An effective method for removing dissolved organic substances that cause tastes, odours, or colours is adsorption by activated carbon. Adsorption is the capacity of a solid particle to attract molecules to its surface. Powdered carbon mixed with water can adsorb and hold many different organic impurities. When the carbon is saturated with impurities, it is cleaned or reactivated by heating to a high temperature in a special furnace.

Fluoridation. Many communities reduce the incidence of tooth decay in young children by adding sodium fluoride or other fluorine compounds to filtered water. The dosage of fluoride must be carefully controlled. Low concentrations are beneficial and cause no harmful side effects, but very high concentrations of fluoride may cause discoloration of tooth enamel.

Desalination. Desalination, or desalting, is the separation of fresh water from salt water or brackish water. Major advances in desalination technology have taken place since the 1950s, as the need for supplies of fresh water has grown in arid and densely populated areas of the world. Desalted water is the main source of municipal supply in areas of the Caribbean, the Middle East, and North Africa, and its use is increasing in the southeastern United States. Although it is relatively expensive to produce, desalted water can be more economical than the alternative of transporting large quantities of fresh water over long distances.

There are two basic types of desalting techniques: thermal processes and membrane processes. Both types consume considerable amounts of energy. Thermal methods involve heat transfer and a phase change of the water from liquid into vapour or ice. Membrane methods use very thin sheets of special plastic that act as selective barriers, allowing pure water to be separated from the salt.

Thermal processes. Distillation, a thermal process that includes heating, evaporation, and condensation, is the oldest and most widely used of desalination technologies. Modern methods for the distillation of large quantities of salt water rely on the fact that the boiling temperature of water is lowered as air pressure drops, significantly reducing the amount of energy needed to vaporize the water. Systems that utilize this principle include multistage flash distillation, multiple effect distillation, and vapour compression distillation.

Multistage flash distillation

Multistage flash distillation plants account for more than half of the world production of desalted water. The process is carried out in a series of closed vessels (stages) set at progressively lower internal pressures. Heat is added to the system from a boiler. When preheated salt water enters a low-pressure chamber, some of it rapidly boils, or flashes, into water vapour. The vapour is condensed into fresh water on heat-exchange tubes that run through each stage. These tubes carry incoming seawater, thereby reducing the heat required from the boiler. Fresh water collects in trays under the tubes. The remaining brine flows into the next stage at even lower pressure, where some of it again flashes into vapour. A multistage flash plant may have as many as 40 stages, permitting salt water to boil repeatedly without supplying additional heat.

Multiple effect distillation also takes place in a series of low-pressure vessels (effects), but it differs from multistage distillation in that preheated salt water is sprayed onto evaporator tubes in order to promote rapid evaporation in each vessel. This process requires pumping the salt water from one effect to the next.

In the vapour compression system, heat is provided by the compression of vapour rather than by direct heat input from a boiler. When the vapour is rapidly compressed, its temperature rises. Some of the compressed and heated vapour is then recycled through a series of tubes passing through a reduced-pressure chamber, where evaporation of salt water occurs. Electricity is the main source of energy for this process. It is used for small-scale desalting applications—for example, at coastal resorts.

Two other thermal processes are solar humidification and freezing. In solar humidification, salt water is collected in shallow basins in a "still," a structure similar to a greenhouse. The water is warmed as sunlight enters through inclined glass or plastic covers. Water vapour rises, condenses on the cooler covers, and trickles down to a collecting trough. Thermal energy from the sun is free, but a solar still is expensive to build, requires a large land area, and needs additional energy for pumping water to and from the facility. Solar humidification units are suitable for providing desalted water to individual families or for very small villages where sunlight is abundant.

The freezing process, also called crystallization, involves cooling salt water to form crystals of pure ice. The ice crystals are separated from the unfrozen brine, rinsed to remove residual salt, and then melted to produce fresh water. Freezing is theoretically more efficient than distillation, and scaling as well as corrosion problems are lessened at the lower operating temperatures, but the mechanical difficulties of handling mixtures of ice and water prevent the construction of large-scale commercial plants. In hot climates, heat leakage into the facility is also a significant problem.

Membrane processes. Two commercially important membrane processes used for desalination are electrodialysis and reverse osmosis. They are used mainly to desalt brackish or highly mineralized water supplies rather than much saltier seawater. In both methods, thin plastic sheets act as selective barriers, allowing fresh water but not salt to flow through.

Desalination by electrodialysis

Most salts dissolved in water exist in the form of electrically charged particles called ions. Half are positively charged (*e.g.,* sodium), and half are negatively charged (*e.g.,* chloride). In electrodialysis, an electric voltage is applied across the saline solution. This causes ions to migrate toward the electrode that has a charge opposite to that of their own. In a typical electrodialysis unit, several hundred plastic membranes that are selectively permeable to either positive ions or negative ions, but not both, are closely spaced in alternation and bound together with electrodes on the outside. Incoming salt water flows between the membrane sheets. Under the applied voltage, the ions move in opposite directions through the membranes, but they are trapped by the next membrane in the stack. This forms alternate cells of dilute salt water and brine. The more dilute solution is recycled back through the stack until it reaches freshwater quality.

When a semipermeable membrane separates two solutions of different concentrations, there is a natural ten-

dency for the concentrations to become equalized. Water flows from the dilute side to the concentrated side. This process is called osmosis. However, pressure applied to the concentrated side can reverse the direction of this flow. In reverse osmosis (also called ultrafiltration), salty water is pumped into a vessel and pressurized against the membrane. Fresh water diffuses through the membrane, leaving a more concentrated salt solution behind.

Next to multistage flash distillation, reverse osmosis is the second-ranking desalting process. It will play a greater role in the desalting of seawater and brackish water as more durable membranes are developed. It can also be applied to the advanced treatment of municipal sewage and industrial wastewater.

Cogeneration and hybrid processes. Desalting costs are reduced by using cogeneration and hybrid processes. Cogeneration (or dual-purpose) desalination plants are large-scale facilities that produce both electric power and desalted seawater. Distillation methods, in particular, are suitable for cogeneration. The high-pressure steam that runs electric generators can be recycled in the distillation unit's brine heater. This significantly reduces fuel consumption compared to what is required if separate facilities are built. Cogeneration is very common in the Middle East and North Africa.

Hybrid systems are units that operate with two or more different desalting processes (*e.g.*, distillation and reverse osmosis). They offer further economic benefits when employed in cogeneration plants, productively combining the operation of each process.

Effluent disposal. Desalination produces fresh water but also a significant volume of waste effluent, called brine. Since the primary pollutant in the brine is salt, disposal in the ocean is generally not a problem for facilities located near a coastline. At inland desalination facilities, care must be taken to prevent pollution of ground or surface waters. Methods of brine disposal include dilution, evaporation, injection into a saline aquifer, or pipeline transport to a suitable disposal point.

WATER DISTRIBUTION

A water distribution system is a network of pumps, pipelines, storage tanks, and other appurtenances. It must deliver adequate quantities of water at pressures sufficient for operating plumbing fixtures and fire fighting equipment, yet it must not deliver water at such high pressures as to increase the occurrence of leaks and pipeline breaks. Pressure regulating valves may be installed to reduce pressure levels in low-lying service areas. More than half the cost of a municipal water supply system is for the distribution network.

Water mains and feeders

Pipelines. The pipeline system of a municipal water distribution network consists of arterial water mains or primary feeders, which convey water from the treatment plant to areas of major water use in the community, and smaller-diameter pipelines called secondary feeders, which tie in to the mains. Usually not less than 6 inches (150 millimetres) in diameter, these pipelines are placed within the public right-of-way so that service connections can be made for all potential water users. The pipelines are usually arranged in a gridiron pattern that allows water to circulate in interconnected loops; this permits any broken sections of pipe to be isolated for repair without disrupting service to large areas of the community. "Dead-end" patterns may also be used, but they do not permit circulation, and the water they provide is more susceptible to taste and odour problems due to stagnation.

A water distribution pipeline must be able to resist internal and external forces, as well as corrosion. Pipes are placed under stress by internal water pressure, by the weight of the overlying soil, and by vehicles passing above. They may also have to withstand water-hammer forces; these occur when valves are closed too rapidly, causing pressure waves to surge through the system. In addition, metal pipes may rust internally if the water supply is corrosive or externally because of corrosive soil conditions.

Materials. Distribution pipes are made of asbestos cement, cast iron, ductile iron, plastic, reinforced concrete, or steel. Although not as strong as iron, asbestos cement, because of its corrosion resistance and ease of installation, is a desirable material for secondary feeders up to 16 inches in diameter. Pipe sections are easily joined with a coupling sleeve and rubber-ring gasket. Cast iron has an excellent record of service, with many installations still functioning after 100 years. Ductile iron, a stronger and more elastic type of cast iron, is used in newer installations. Iron pipes are provided in diameters up to 48 inches and are usually coated to prevent corrosion. Underground sections are connected with bell-and-spigot joints, the spigot end of one pipe section being pushed into the bell end of an adjacent section. A rubber-ring gasket in the bell end is compressed when the two sections are joined, creating a watertight, flexible connection. Flanged and bolted joints are used for above-ground installations.

The use of plastic pipes is increasing. Available in diameters up to 24 inches, they are lightweight and easily installed. They are also corrosion-resistant, and their smoothness provides good hydraulic characteristics. Plastic pipes are connected either by a bell-and-spigot compression-type joint or by threaded screw couplings.

Precast reinforced concrete pipe sections up to 12 feet in diameter are used for arterial mains. Reinforced concrete pipes are strong and durable. They are joined using a bell-and-spigot-type connection that is sealed with cement mortar. Steel pipe is sometimes used for arterial mains in above-ground installations. It is very strong and lighter than concrete pipe, but it must be protected against corrosion by lining the interior and by painting and wrapping the exterior. Sections of steel pipe are joined by welding or with mechanical coupling devices.

Fittings. In order to function properly, a water distribution system requires several types of fittings, including hydrants, shut-off valves, and other appurtenances. The main purpose of hydrants is to provide water for fire fighting. They also are used for flushing water mains, pressure testing, water sampling, and washing debris off public streets.

Many types of valves are used to control the quantity and direction of water flow. Gate valves are usually installed throughout the pipe network. They allow sections to be shut off and isolated during the repair of broken mains, pumps, or hydrants. A type of valve commonly used for throttling and controlling the rate of flow is the butterfly valve. Other valves used in water distribution systems include pressure-reducing valves, check valves, and air-release valves.

Installation. Water mains must be placed three to six feet below the ground surface to protect against traffic loads and to prevent freezing. Since the water in a distribution system is under pressure, pipelines can follow the shape of the land, uphill as well as downhill. They must be installed with proper bedding and backfill. Compaction of soil layers under the pipe (bedding) as well as above the pipe (backfill) is necessary to provide proper support. A water main should never be installed in the same trench with a sewer line. Where the two must cross, the water main should be placed above the sewer line.

Low-lift and high-lift pumps

Pumps. Many kinds of pumps are used in distribution systems. Pumps that lift surface water and move it to a nearby treatment plant are called low-lift pumps. These move large volumes of water at relatively low discharge pressures. Pumps that discharge treated water into arterial mains are called high-lift pumps; these operate under higher pressures. Pumps that increase the pressure within the distribution system or raise water into an elevated storage tank are called booster pumps. Well pumps lift water from underground and discharge it directly into a distribution system.

Most water distribution pumps are of the centrifugal type, in which a rapidly rotating impeller adds energy to the water and raises the pressure inside the pump casing. The flow rate through a centrifugal pump depends on the pressure against which it operates. The higher the pressure, the lower the flow or discharge. Another kind of pump is the positive-displacement type. This pump delivers a fixed quantity of water with each cycle of a piston or rotor. The water is literally pushed or displaced from the pump casing. The flow capacity of a positive-

displacement pump is unaffected by the pressure of the system in which it operates.

Storage tanks. Distribution storage tanks, familiar sights in many communities, serve two basic purposes: equalizing storage and emergency storage. Equalizing storage is the volume of water needed to satisfy peak hourly demands in the community. During the late-night and early-morning hours, when water demand is very low, high-lift pumps fill the tank. During the day, when the water demand is high, water flows out of the tank to help satisfy the peak hourly water needs. This allows for a uniform flow rate at the treatment plant and pumping station. The capacity of a distribution storage tank is about equal to the average daily water demand.

Distribution storage tanks are built at ground level on hilltops higher than the service area. In areas with flat topography, the tanks may be elevated above ground on towers in order to provide adequate water pressures, or ground-level storage tanks with booster pumping may be provided.

Water-pollution control

Water is called the "universal solvent" because of its strong tendency to dissolve other substances. Since pure water is not found in nature (*i.e.,* outside chemical laboratories), any distinction between clean water and polluted water depends on the type and concentration of impurities found in the water as well as on its intended use. In broad terms, water is said to be polluted when it contains enough impurities to make it unfit for a particular use, such as drinking, swimming, or fishing. Although water quality is affected by natural conditions, the word pollution usually implies human activity as the source of contamination. Water pollution is caused primarily by the drainage of contaminated waters into surface water or groundwater. Water-pollution control, therefore, primarily involves the removal of impurities before they reach natural bodies of water or aquifers.

HISTORICAL BACKGROUND

The Roman sewers

Direct discharge of sewage. Many ancient cities had drainage systems, but they were primarily intended to carry rainwater away from roofs and pavements. A notable example is the drainage system of ancient Rome. It included many surface conduits that were connected to a large vaulted channel, called the Cloaca Maxima ("Great Sewer"), which carried drainage water to the Tiber River. Built of stone and on a grand scale, the Cloaca Maxima is one of the oldest existing monuments of Roman engineering.

There was little progress in urban drainage or sewerage during the Middle Ages. Privy vaults and cesspools were used, but most wastes were simply dumped into gutters to be flushed through the drains by floods. Toilets or water closets were installed in houses in the early 19th century, but they were usually connected to cesspools, not to sewers. In densely populated areas, local conditions soon became intolerable, because the cesspools were seldom emptied and frequently overflowed. The threat to public health became apparent. In England in the middle of the 19th century, outbreaks of cholera were traced directly to well-water supplies contaminated with human wastes from privy vaults and cesspools. It soon became necessary for all water closets in the larger towns to be connected directly to the storm sewers. This transferred sewage from the ground near where houses stood to nearby bodies of water. Thus, a new problem emerged: surface water pollution.

Developments in sewage treatment. It used to be said that "the solution to pollution is dilution." When small amounts of sewage are discharged into a flowing body of water, a natural process of stream self-purification occurs. Densely populated communities generate such large quantities of sewage, however, that dilution alone does not prevent pollution. This makes it necessary to treat or purify wastewater to some degree before disposal.

The construction of centralized sewage treatment or water pollution control plants began in the late 19th and early 20th centuries, principally in the United Kingdom and the United States. Instead of discharging sewage directly into a nearby body of water, it was first passed through a combination of physical, biological, and chemical processes that removed some or most of the pollutants. Also beginning in the 1900s, new sewage-collection systems were designed to separate storm water from domestic wastewater, so that treatment plants did not become overloaded during periods of wet weather.

After the middle of the 20th century, increasing public concern for environmental quality led to broader and more stringent regulation of wastewater disposal practices. Higher levels of treatment were required. For example, pretreatment of industrial wastewater, with the aim of preventing toxic chemicals from interfering with the biological processes used at sewage treatment plants, often became a necessity. In fact, wastewater treatment technology advanced to the point where it became possible to remove virtually all pollutants from sewage. This was so expensive, however, that such high levels of treatment were not usually justified.

Modern water pollution control plants became large, complex facilities that required considerable amounts of energy for their operation. After the rise of oil prices in the 1970s, concern for energy conservation became a more important factor in the design of new pollution control systems. Consequently, land disposal and subsurface disposal of sewage began to receive increased attention where feasible. Such low-tech pollution control methods not only would help to conserve energy but also would serve to recycle nutrients and replenish groundwater supplies.

SOURCES OF POLLUTION

Point sources and dispersed sources

Water pollutants may originate from a point source or from a dispersed source. A point-source pollutant is one that reaches water from a single pipeline or channel, such as a sewage-discharge or outfall pipe. Dispersed sources are broad, unconfined areas from which pollutants enter a body of water. Surface runoff from farms, for example, is a dispersed source of pollution, carrying animal wastes, fertilizers, pesticides, and silt into nearby streams. Urban storm water drainage is also considered a dispersed source because of the many locations at which it enters local streams or lakes. Point-source pollutants are easier to control than dispersed-source pollutants, since they flow to a single location where treatment processes can remove them from the water. Such control is not usually possible over pollutants from dispersed sources, which cause a large part of the overall water pollution problem. Dispersed-source water pollution is best reduced by enforcing proper land-use plans and development standards.

General types of water pollutants include pathogenic organisms, oxygen-demanding wastes, plant nutrients, synthetic organic chemicals, inorganic chemicals, sediments, radioactive substances, oil, and heat. Sewage is the primary source of the first three types. Farms and industrial facilities are also sources of some of them. Sediment from eroded topsoil is considered a pollutant because it can damage aquatic ecosystems, and heat (particularly from power-plant cooling water) is considered a pollutant because of the adverse effect it has on dissolved oxygen levels and aquatic life in rivers and lakes.

SEWAGE CHARACTERISTICS

Types of sewage. There are three types of wastewater or sewage: domestic sewage, industrial sewage, and storm sewage. Domestic sewage carries used water from houses and apartments; it is also called sanitary sewage. Industrial sewage is used water from manufacturing or chemical processes. Storm sewage, or storm water, is runoff from precipitation that is collected in a system of pipes or open channels.

Domestic sewage is slightly more than 99.9 percent pure water by weight. The rest, less than 0.1 percent, contains a wide variety of dissolved and suspended impurities. Although amounting to a very small fraction of the sewage by weight, the nature of these impurities and the large volumes of sewage in which they are carried make disposal of domestic wastewater a significant technical problem. The principal impurities are putrescible organic materials and

plant nutrients, but domestic sewage is also very likely to contain disease-causing microbes. Industrial wastewater usually contains specific and readily identifiable chemical compounds, depending on the nature of the industrial process. Storm sewage carries organic materials, suspended and dissolved solids, and other substances picked up as it travels over the ground.

Principal pollutants. *Organic material.* The amount of putrescible organic material in sewage is measured by the biochemical oxygen demand, or BOD; the more organic material there is in the sewage, the higher the BOD. BOD is the amount of oxygen required by microorganisms to decompose the organic substances in sewage. It is among the most important parameters for the design and operation of sewage treatment plants. Industrial sewage may have BOD levels many times that of domestic sewage. The BOD of storm sewage is of particular concern when it is mixed with domestic sewage in combined sewer systems.

Biochemical oxygen demand (BOD)

Dissolved oxygen is an important water quality factor for lakes and rivers. The higher the concentration of dissolved oxygen, the better the water quality. When sewage enters a lake or stream, decomposition of the organic materials begins. Oxygen is consumed as microorganisms use it in their metabolism. This can quickly deplete the available oxygen in the water. When the dissolved oxygen levels drop too low, trout and other aquatic species soon perish. In fact, if the oxygen level drops to zero, the water will become septic. Decomposition of organic compounds without oxygen causes the undesirable odours usually associated with septic or putrid conditions.

Suspended solids. Another important characteristic of sewage is suspended solids. The volume of sludge produced in a treatment plant is directly related to the total suspended solids present in the sewage. Industrial and storm sewage may contain higher concentrations of suspended solids than domestic sewage. The extent to which a treatment plant removes suspended solids, as well as BOD, determines the efficiency of the treatment process.

Plant nutrients. Domestic sewage contains compounds of nitrogen and phosphorus, two elements that are basic nutrients essential for the growth of plants. In lakes, excessive amounts of nitrates and phosphates can cause the rapid growth of algae. Algal blooms, often caused by sewage discharges, accelerate the natural aging of lakes in a process called eutrophication.

Microbes. Domestic sewage contains many millions of microorganisms per gallon. Most are harmless coliform bacteria from the human intestinal tract, but domestic sewage is also likely to carry pathogenic microbes. Coliforms are used as indicators of sewage pollution; a high coliform count usually indicates recent sewage pollution.

WASTEWATER COLLECTION SYSTEMS

A wastewater collection system is a network of pipes, pumping stations, and appurtenances that convey sewage from its points of origin to a point of treatment and disposal.

Combined systems. Systems that carry a mixture of both domestic sewage and storm sewage are called combined sewers. Combined sewers typically consist of large-diameter pipes or tunnels, because of the large volumes of storm water that must be carried during wet-weather periods. They are very common in older cities but are no longer designed and built as part of new sewerage facilities. Because wastewater treatment plants cannot handle large volumes of storm water, sewage must bypass the treatment plants during wet weather and be discharged directly into the receiving water. These combined-sewer overflows, containing untreated domestic sewage, cause recurring water pollution problems and are very troublesome sources of pollution.

In some large cities, the combined-sewer overflow problem has been reduced by diverting the first flush of combined sewage into a large basin or underground tunnel. After temporary storage, it can be treated by settling and disinfection before being discharged into a receiving body of water, or it can be treated in a nearby wastewater treatment plant at a rate that will not overload the facility. Another method for controlling combined sewage involves the use of swirl concentrators. These direct sewage through cylindrically shaped devices that create a vortex, or whirlpool, effect. The vortex helps concentrate impurities in a much smaller volume of water for treatment.

Managing combined-sewer overflow

Separate systems. New wastewater collection facilities are designed as separate systems, carrying either domestic sewage or storm sewage but not both. Storm sewers usually carry surface runoff to a point of disposal in a stream or river. Small detention basins may be built as part of the system, storing stormwater temporarily and reducing the magnitude of the peak flow rate. Sanitary sewers, on the other hand, carry domestic wastewater to a sewage treatment plant. Pretreated industrial wastewater may be allowed into municipal sanitary sewer systems, but storm water is excluded.

Storm sewers are usually built with sections of reinforced concrete pipe. Corrugated metal pipes may be used in some cases. Storm water inlets or catch basins are located at suitable intervals in a street right-of-way or in easements across private property. The pipelines are usually located to allow downhill gravity flow to a nearby stream or to a detention basin. Storm water pumping stations are avoided, if possible, because of the very large pump capacities that would be needed to handle the intermittent flows.

A sanitary sewer system includes laterals, submains, and interceptors. Except for individual house connections, laterals are the smallest sewers in the network. They usually are not less than eight inches or 200 millimetres in diameter and carry sewage by gravity into larger submains, or collector sewers. The collector sewers tie in to a main interceptor, or trunk line, which carries the sewage to a treatment plant. Interceptors are usually built with precast sections of reinforced concrete pipe, up to 15 feet or 4 metres in diameter. Other materials used for sanitary sewers include vitrified clay, asbestos cement, plastic, steel, or

© John Edwards/Tony Stone Images

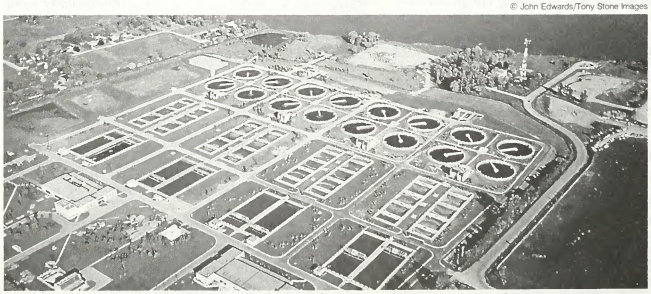


Figure 49: Sewage treatment plant on the shores of Lake Ontario, Pickering, Ont., Can. In the foreground are primary clarifiers, in the centre are aeration tanks, and in the background are secondary clarifiers.

ductile iron. The use of plastic for laterals is increasing because of its lightness and ease of installation. Iron and steel pipes are used for force mains or in pumping stations. (Force mains are pipelines that carry sewage under pressure when it must be pumped.)

Pumps. Pumping stations are built when sewage must be raised from a low point to a point of higher elevation, or where the topography prevents downhill gravity flow. Special nonclogging pumps are available to handle raw sewage. They are installed in structures called lift stations.

Lift stations

There are two basic types of lift station: dry well and wet well. A wet-well installation has only one chamber or tank to receive and hold the sewage until it is pumped out. Specially designed submersible pumps and motors can be located at the bottom of the chamber, completely below the water level. Dry-well installations have two separate chambers, one to receive the wastewater and one to enclose and protect the pumps and controls. The protective dry chamber allows easy access for inspection and maintenance. All sewage lift stations, whether of the wet-well or dry-well type, should include at least two pumps. One pump can operate while the other is removed for repair.

Flow rates. There is a wide variation in sewage flow rates over the course of a day. A sewer system must accommodate this variation. In most cities, domestic sewage flow rates are highest in the morning and evening hours. They are lowest during the middle of the night. Flow quantities depend upon population density, water consumption, and the extent of commercial or industrial activity in the community. The average sewage flow rate is usually about the same as the average water use in the community. In a lateral sewer, short-term peak flow rates can be roughly four times the average flow rate. In a trunk sewer, peak flow rates may be two-and-a-half times the average.

WASTEWATER TREATMENT AND DISPOSAL

The predominant method of wastewater disposal in large cities and towns is discharge into a body of surface water. Suburban and rural areas rely more on subsurface disposal. In either case, wastewater must be purified or treated to some degree, in order to protect both public health and water quality. Suspended particulates and biodegradable organics must be removed to varying extents. Pathogenic bacteria must be destroyed. It may also be necessary to remove nitrates and phosphates (plant nutrients) and to neutralize or remove industrial wastes and toxic chemicals.

Stream standards and effluent standards

The degree to which wastewater must be treated varies, depending on local environmental conditions and governmental standards. Two pertinent types of standards are stream standards and effluent standards. Stream standards, designed to prevent the deterioration of existing water quality, set limits on the amounts of specific pollutants allowed in streams, rivers, and lakes. The limits depend on a classification of the "maximum beneficial use" of the water. Water quality parameters that are regulated by stream standards include dissolved oxygen, coliforms, turbidity, acidity, and toxic substances. Effluent standards, on the other hand, pertain directly to the quality of the treated wastewater discharged from a sewage treatment plant. The factors controlled under these standards usually include biochemical oxygen demand, suspended solids, acidity, and coliforms.

There are three levels of wastewater treatment: primary, secondary, and tertiary (or advanced). Primary treatment removes about 60 percent of total suspended solids and about 35 percent of BOD; dissolved impurities are not removed. It is usually used as a first step before secondary treatment. Secondary treatment removes more than 85 percent of both suspended solids and BOD. A minimum level of secondary treatment is usually required in the United States and other developed countries. When more than 85 percent of total solids and BOD must be removed, or when dissolved nitrate and phosphate levels must be reduced, tertiary (advanced) treatment methods are used. Tertiary processes can remove more than 99 percent of all the impurities from sewage, producing an effluent of almost drinking-water quality. Tertiary treatment can be very expensive, often doubling the cost of secondary treatment. It is used only under special circumstances.

For all levels of wastewater treatment, the last step prior to discharge of the sewage effluent into a body of surface water is disinfection. Disinfection is usually accomplished by mixing the effluent with chlorine gas in a contact tank for at least 15 minutes. Because chlorine residuals in the effluent may have adverse effects on aquatic life, an additional chemical may be added to dechlorinate the effluent. Ultraviolet radiation, which can disinfect without leaving any residual in the effluent, is becoming more competitive with chlorine as a wastewater disinfectant.

Primary treatment. Primary treatment removes material that will either float or readily settle out by gravity. It includes the physical processes of screening, comminution, grit removal, and sedimentation. Screens are made of long, closely spaced narrow metal bars. They block floating debris such as wood, rags, and other bulky objects that could clog pipes or pumps. In modern plants the screens are cleaned mechanically, and the material is promptly disposed of by burial on the plant grounds. A comminutor may be used to grind and shred debris that passes through the screens. The shredded material is removed later by sedimentation or flotation processes.

Grit chambers are long narrow tanks that are designed to slow down the flow so that solids such as sand, coffee grounds, and eggshells will settle out of the water. Grit causes excessive wear and tear on pumps and other plant equipment. Its removal is particularly important in cities with combined sewer systems, which carry a good deal of silt, sand, and gravel that wash off streets or land during a storm.

Primary clarifiers

Suspended solids that pass through screens and grit chambers are removed from the sewage in sedimentation tanks. These tanks, also called primary clarifiers, provide about two hours of detention time for gravity settling to take place. As the sewage flows through them slowly, the solids gradually sink to the bottom. The settled solids (known as raw or primary sludge) are moved along the tank bottom by mechanical scrapers. Sludge is collected in a hopper, where it is pumped out for removal. Mechanical surface-skimming devices remove grease and other floating materials.

Secondary treatment. Secondary treatment removes the soluble organic matter that escapes primary treatment. It also removes more of the suspended solids. Removal is usually accomplished by biological processes in which microbes consume the organic impurities as food, converting them into carbon dioxide, water, and energy for their own growth and reproduction. The sewage treatment plant provides a suitable environment, albeit of steel and concrete, for this natural biological process. Removal of soluble organic matter at the treatment plant helps to protect the dissolved oxygen balance of a receiving stream, river, or lake.

There are three basic biological treatment methods: the trickling filter, the activated sludge process, and the oxidation pond. A fourth, less common method is the rotating biological contacter.

Trickling filter. A trickling filter is simply a tank filled with a deep bed of stones. Settled sewage is sprayed continuously over the top of the stones and trickles to the bottom, where it is collected for further treatment. As the wastewater trickles down, bacteria gather and multiply on the stones. The steady flow of sewage over these growths allows the microbes to absorb the dissolved organics, thus lowering the BOD of the sewage. Air circulating upward through the spaces among the stones provides sufficient oxygen for the metabolic processes.

Settling tanks, called secondary clarifiers, follow the trickling filters. These clarifiers remove microbes that are washed off the rocks by the flow of wastewater. Two or more trickling filters may be connected in series, and sewage can be recirculated in order to increase treatment efficiencies.

Activated sludge. The activated sludge treatment system consists of an aeration tank followed by a secondary clarifier (Figure 50). Settled sewage, mixed with fresh sludge that is recirculated from the secondary clarifier, is introduced into the aeration tank. Compressed air is then injected into the mixture through porous diffusers located

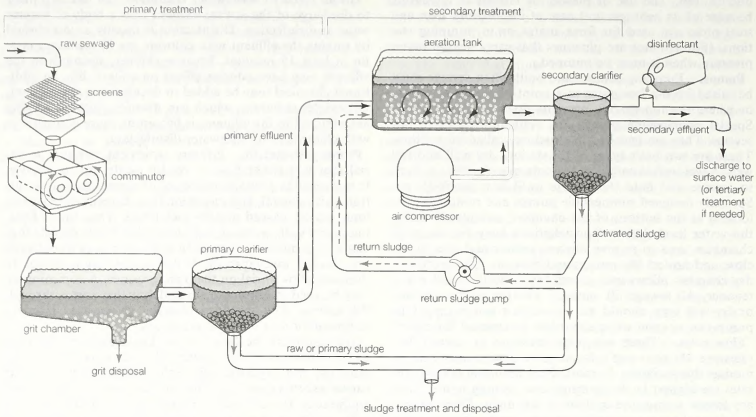


Figure 50: Primary and secondary treatment of sewage, using the activated sludge process.

at the bottom of the tank. As it bubbles to the surface, the diffused air provides oxygen and a rapid mixing action. Air can also be added by the churning action of mechanical propeller-like mixers located at the tank surface.

Under such oxygenated conditions, microorganisms thrive, forming an active, healthy suspension of biological solids (mostly bacteria) called activated sludge. About six hours of detention is provided in the aeration tank. This gives the microbes enough time to absorb dissolved organics from the sewage, reducing the BOD. The mixture then flows from the aeration tank into the secondary clarifier, where activated sludge settles out by gravity. Clear water is skimmed from the surface of the clarifier, disinfected, and discharged as secondary effluent. The sludge is pumped out from a hopper at the bottom of the tank. About 30 percent of the sludge is recirculated back into the aeration tank, where it is mixed with the primary effluent.

Recirculation of microbes

This recirculation is a key feature of the activated sludge process. The recycled microbes are well acclimated to the sewage environment and readily metabolize the organic materials in the primary effluent. The remaining 70 percent of the secondary sludge must be treated and disposed of in an acceptable manner (see below *Sludge treatment and disposal*).

Variations of the activated sludge process include extended aeration, contact stabilization, and high-purity oxygen aeration. Extended-aeration and contact-stabilization systems omit the primary settling step. They are efficient for treating small sewage flows from motels, schools, and other relatively isolated wastewater sources. Both of these treatments are usually provided in prefabricated steel tanks called package plants (Figure 51). Oxygen aeration systems mix pure oxygen with activated sludge. A richer concentration of oxygen allows the aeration time to be shortened from six to two hours, reducing the required tank volume.

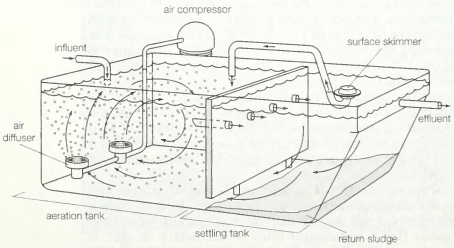


Figure 51: Schematic diagram of a prefabricated "package plant" for the aeration treatment of small sewage flows.

Oxidation pond. Oxidation ponds (also called lagoons or stabilization ponds) are large, shallow ponds designed to treat wastewater through the interaction of sunlight, bacteria, and algae. Algae grow using energy from the sun and carbon dioxide and inorganic compounds released by bacteria in water. During the process of photosynthesis, the algae release oxygen needed by aerobic bacteria. Mechanical aerators are sometimes installed to supply yet more oxygen, thereby reducing the required size of the pond. Sludge deposits in the pond must eventually be removed by dredging. Algae remaining in the pond effluent can be removed by filtration or by a combination of chemical treatment and settling.

Rotating biological contacter. In this treatment system, a series of large plastic disks mounted on a horizontal shaft are partially submerged in primary effluent. As the shaft rotates, the disks are exposed alternately to air and wastewater, allowing a layer of bacteria to grow on the disks and to metabolize the organics in the wastewater.

Tertiary treatment. When the intended receiving water is very vulnerable to the effects of pollution, secondary effluent may be treated further by several tertiary processes.

Effluent polishing. For the removal of additional suspended solids and BOD from secondary effluent, effluent polishing is an effective treatment. It is most often accomplished using granular-media filters, much like the filters used to purify drinking water. Polishing filters are usually built as prefabricated units, with tanks placed directly above the filters for storing backwash water. Effluent polishing of wastewater may also be achieved using microstrainers of the type used in municipal water treatment (see above *Water treatment: Clarification: Filtration*).

Removal of plant nutrients. When treatment standards require the removal of plant nutrients from the sewage, it is often done as a tertiary step. Phosphorus in wastewater is usually present in the form of organic compounds and phosphates that can easily be removed by chemical precipitation. This process, however, increases the volume and weight of sludge. Nitrogen, another important plant nutrient, is present in sewage in the form of ammonia and nitrates. Ammonia is toxic to fish, and it also exerts an oxygen demand in receiving waters as it is converted to nitrates. Nitrates, like phosphates, promote the growth of algae and the eutrophication of lakes. A method called nitrification-denitrification can be used to remove the ni-

Removing phosphorus and nitrogen

trates. It is a two-step biological process in which ammonia nitrogen is first converted into nitrates by microorganisms. The nitrates are further metabolized by another species of bacteria, forming nitrogen gas that escapes into the air. This process requires the construction of more aeration and settling tanks and significantly increases the cost of treatment.

A physical-chemical process called ammonia stripping may be used to remove ammonia from sewage. Chemicals are added to convert ammonium ions into ammonia gas. The sewage is then cascaded down through a tower, allowing the gas to come out of solution and escape into the air. Stripping is less expensive than nitrification-denitrification, but it does not work very efficiently in cold weather.

Land treatment. In some locations, secondary effluent can be applied directly to the ground and a polished effluent obtained by natural processes as the wastewater flows over vegetation and percolates through the soil. There are three types of land treatment: slow rate, rapid infiltration, and overland flow.

In the slow-rate, or irrigation, method, effluent is applied onto the land by ridge-and-furrow spreading (in ditches) or by sprinkler systems. Most of the water and nutrients are absorbed by the roots of growing vegetation. In the rapid infiltration method, the wastewater is stored in large ponds called recharge basins. Most of it percolates to the groundwater, and very little is absorbed by vegetation. For this method to work, soils must be highly permeable.

In overland flow, wastewater is sprayed onto an inclined vegetated terrace and slowly flows to a collection ditch. Purification is achieved by physical, chemical, and biological processes, and the collected water is usually discharged into a nearby stream.

Land treatment of sewage can provide moisture and nutrients for the growth of vegetation, such as corn or grain for animal feed. It also can recharge, or replenish, groundwater aquifers. Land treatment, in effect, allows sewage to be recycled for beneficial use. Large land areas are required, however, and the feasibility of this kind of treatment may be limited further by soil texture and climate.

Subsurface disposal. In sparsely populated suburban or rural areas, it is usually not economical to build sewage collection systems and a centrally located treatment plant. Instead, a separate subsurface disposal system is provided for each home. For subsurface disposal to succeed, the permeability, or hydraulic conductivity, of the soil must be within an acceptable range. The capacity of the soil to absorb settled wastewater is determined by a "percolation test."

Septic tanks

A subsurface disposal system consists of a buried septic tank and either a leaching field or seepage pits. A septic tank serves as a settling tank and sludge storage chamber. Although the sludge decomposes anaerobically, it eventually accumulates and must be pumped out periodically. Floating solids and grease are trapped by a baffle at the tank outlet, and settled sewage flows out into the leaching field or seepage pits. A leaching field includes several perforated pipelines placed in shallow trenches. The pipes distribute the effluent over a sizable area as it seeps into the soil. If the site is too small for a conventional leaching field, deeper seepage pits may be used instead of shallow trenches. Both leaching fields and seepage pits must be placed above seasonally high groundwater levels.

Wastewater reuse. Wastewater can be a valuable resource in cities or towns where population is growing and water supplies are limited. In addition to easing the strain on limited fresh water supplies, the reuse of wastewater can improve the quality of streams and lakes by reducing the effluent discharges that they receive. Wastewater may be reclaimed and reused for crop and landscape irrigation, groundwater recharge, or recreational purposes. Reclamation for drinking is technically possible, but this reuse faces significant public resistance.

There are two types of wastewater reuse: direct and indirect. In direct reuse, treated wastewater is piped into some type of water system without first being diluted in a natural stream or lake or in groundwater. One example is the irrigation of a golf course with effluent from a municipal wastewater treatment plant. Indirect reuse involves the mixing of reclaimed wastewater with another body of water before reuse. In effect, any community that uses a surface water supply downstream from the treatment plant discharge pipe of another community is indirectly reusing wastewater. Indirect reuse is also accomplished by discharging reclaimed wastewater into a groundwater aquifer and later withdrawing the water for use. Discharge into an aquifer (called artificial recharge) is done by either deep-well injection or shallow surface spreading.

Quality and treatment requirements for reclaimed wastewater become more stringent as the chances for direct human contact and ingestion increase. The impurities that must be removed depend on the intended use of the water. For example, removal of phosphates or nitrates is not necessary if the intended use is landscape irrigation. If direct reuse as a potable supply is intended, tertiary treatment with multiple barriers against contaminants is required. This may include secondary treatment followed by granular-media filtration, ultraviolet radiation, granular activated carbon adsorption, reverse osmosis, air stripping, ozonation, and chlorination.

Recycling nonpotable water

The use of gray-water recycling systems in new commercial buildings offers a method of saving water and reducing total sewage volumes. These systems filter and chlorinate drainage from tubs and sinks and reuse the water for nonpotable purposes (*e.g.,* flushing toilets and urinals). Recycled water can be marked with a blue dye to ensure that it is not used for potable purposes.

SLUDGE TREATMENT AND DISPOSAL

Characteristics. The residue that accumulates in sewage treatment plants is called sludge. Treatment and disposal of sewage sludge are major factors in the design and operation of all water pollution control plants. Two basic goals of treating sludge before final disposal are to reduce its volume and to stabilize the organic materials. Stabilized sludge does not have an offensive odour and can be handled without causing a nuisance or health hazard. Smaller sludge volume reduces the costs of pumping and storage.

Treatment methods. Treatment of sewage sludge may include a combination of thickening, digestion, dewatering, and disposal processes.

Thickening. Thickening is usually the first step in sludge treatment, because it is impractical to handle thin sludge, a slurry of solids suspended in water. Thickening is usually accomplished in a tank called a gravity thickener. A thickener can reduce the total volume of sludge to less than half the original volume. An alternative to gravity thickening is dissolved-air flotation. In this method, air bubbles carry the solids to the surface, where a layer of thickened sludge forms.

Digestion. Sludge digestion is a biological process in which organic solids are decomposed into stable substances. Digestion reduces the total mass of solids, destroys pathogens, and makes it easier to dewater or dry the sludge. Digested sludge is inoffensive, having the appearance and characteristics of a rich potting soil.

Most large sewage treatment plants use a two-stage digestion system in which organics are metabolized by bacteria anaerobically (in the absence of oxygen). In the first stage, the sludge is heated and mixed in a closed tank for about 15 days, while digestion takes place. The sludge then flows into a second tank, which serves primarily for storage and settling. As the organic solids are broken down by anaerobic bacteria, carbon dioxide gas and methane gas are formed. Methane is combustible and is used as a fuel to heat the first digestion tank as well as to generate electricity for the plant. Anaerobic digestion is very sensitive to temperature, acidity, and other factors. It requires careful monitoring and control.

Sludge digestion may also take place aerobically—that is, in the presence of oxygen. The sludge is vigorously aerated in an open tank for about 20 days. Methane gas is not formed in this process. Although aerobic systems are easier to operate than anaerobic systems, they usually cost more to operate because of the power needed for aeration. Aerobic digestion is often combined with small extended-aeration or contact-stabilization systems.

Both aerobic and anaerobic digestion convert about half of the organic sludge solids to liquids and gases.

Dewatering. Digested sewage sludge is usually dewatered before disposal. Dewatered sludge still contains a significant amount of water—often as much as 70 percent—but even at that moisture content, sludge no longer behaves as a liquid and can be handled as a solid material. Sludge drying beds provide the simplest method of dewatering. A digested sludge slurry is spread on an open bed of sand and allowed to remain until dry. Drying takes place by a combination of evaporation and gravity drainage through the sand. A piping network built under the sand collects the water, which is pumped back to the head of the plant. After about six weeks of drying, the sludge cake, as it is called, may have a solids content of about 40 percent. It can then be removed from the sand with a pitchfork or a front-end loader. In order to reduce drying time in wet or cold weather, a glass enclosure may be built over the sand beds. Since a good deal of land area is needed for drying beds, this method of dewatering is commonly used in rural or suburban towns rather than in densely populated cities.

Sludge drying beds

Alternatives to sludge drying beds include the rotary-drum vacuum filter, the centrifuge, and the belt filter press. These mechanical systems require less space than do sludge drying beds, and they offer a greater degree of operational control. However, they usually have to be preceded by a step called sludge conditioning, in which chemicals are added to the liquid sludge to coagulate solids and improve drainability.

Disposal. The final destination of treated sewage sludge usually is the land. Dewatered sludge can be buried underground in a sanitary landfill. It also may be spread on agricultural land in order to make use of its value as a soil conditioner and fertilizer. Since sludge may contain toxic industrial chemicals, it is not spread on land where crops are grown for human consumption.

Where a suitable site for land disposal is not available, as in urban areas, sludge may be incinerated. Incineration completely evaporates the moisture and converts the organic solids into inert ash. The ash must be disposed of, but the reduced volume makes disposal more economical. Air pollution control is a very important consideration when sewage sludge is incinerated; appropriate air cleaning devices such as scrubbers and filters must be used.

Dumping of sludge in the ocean, once an economical disposal method for many coastal communities, is no longer considered a viable option. It is now prohibited in the United States.

Solid-waste management

Material that is discarded because it has served its purpose or is no longer useful is called solid waste. Improper disposal of municipal solid waste can create unsanitary conditions, and these conditions in turn can lead to pollution of the environment and to outbreaks of vector-borne disease (that is, diseases spread by rodents and insects). The tasks of collecting, treating, and disposing of solid wastes present complex technical challenges. They also pose a wide variety of administrative, economic, and social problems that must be managed and solved.

HISTORICAL BACKGROUND

Early waste disposal. In ancient cities, wastes were thrown into the unpaved streets and roadways, where they were left to accumulate. It was not until 320 BC, in Athens, that the first known law forbidding this practice was established. At that time, a system for waste removal began to evolve in Greece and in the Greek-dominated cities of the eastern Mediterranean. In ancient Rome, property owners were responsible for cleaning the streets fronting their property. But organized waste collection was associated only with state-sponsored events, such as parades. Disposal methods were very crude, involving open pits located just outside the city walls. As populations increased, efforts were made to transport wastes farther out from the cities.

After the fall of Rome, waste collection and municipal sanitation began a decline that lasted throughout the Middle Ages. Near the end of the 14th century, scavengers were given the task of carting wastes to dumps outside city walls. But this was not the case in smaller towns, where most people still threw wastes into the streets. It was not until 1714 that every city in England was required to have an official scavenger. Toward the end of the 18th century in America, municipal collection of garbage was begun in Boston, New York City, and Philadelphia. Waste disposal methods were still very crude, however. Garbage collected in Philadelphia, for example, was simply dumped into the Delaware River downstream from the city.

Developments in waste management. A technological approach to solid-waste management began to develop in the latter part of the 19th century. Watertight garbage cans were first introduced in the United States, and sturdier vehicles were used to collect and transport wastes. A significant development in solid-waste treatment and disposal practices was marked by the construction of the first refuse incinerator in England in 1874. By the beginning of the 20th century, 15 percent of major American cities were incinerating solid wastes. Even then, however, most of the largest cities were still using primitive disposal methods such as open dumping on land or in water.

The first incinerators

Technological advances continued during the first half of the 20th century, including the development of garbage grinders, compaction trucks, and pneumatic collection systems. By mid-century, however, it had become evident that open dumping and improper incineration of solid waste were causing problems of pollution and public health. As a result, sanitary landfills were developed to replace the practice of open dumping and to reduce the reliance on waste incineration. In many countries waste was divided into two categories, hazardous and nonhazardous, and separate regulations were developed for their disposal. Landfills were designed and operated in a manner that minimized risks to public health and the environment. New refuse incinerators were designed to recover heat energy from the waste and were provided with extensive air-pollution control devices to satisfy stringent standards of air quality. Modern solid-waste management plants in most developed countries now emphasize the practice of recycling and waste reduction at the source, rather than incineration and land disposal.

SOLID-WASTE CHARACTERISTICS

Composition and properties. The sources of solid waste include residential, commercial, institutional, and industrial activities. Certain types of wastes that cause immediate danger to exposed individuals or environments are classified as hazardous; these are discussed below (see *Hazardous-waste management*). All nonhazardous solid waste from a community that requires collection and transport to a processing or disposal site is called refuse or municipal solid waste. Refuse includes garbage and rubbish. Garbage is mostly decomposable food wastes; rubbish is mostly dry material such as glass, paper, cloth, or wood. Garbage is highly putrescible or decomposable, while rubbish is not. Trash is rubbish that includes bulky items such as old refrigerators, couches, or large tree stumps. Trash requires special collection and handling.

Garbage, rubbish, and trash

Solid-waste characteristics vary considerably among communities and nations. American refuse is usually lighter, for example, than European or Japanese refuse. In the United States paper and paperboard products make up close to 40 percent of the total weight of municipal solid waste; food waste accounts for less than 10 percent. The rest is a mixture of yard trimmings, wood, glass, metal, plastic, leather, cloth, and other miscellaneous materials. In a loose or uncompacted state, municipal solid waste of this type weighs approximately 200 pounds per cubic yard (120 kilograms per cubic metre). These figures vary with geographic location, economic conditions, season of the year, and many other factors. Waste characteristics from each community must be studied carefully before any treatment or disposal facility is designed and built.

Generation and storage. Rates of solid-waste generation also vary widely. In the United States, for example, municipal refuse is generated at an average rate of ap-

proximately 4.4 pounds (2 kilograms) per person per day. Japan generates roughly half this amount, yet in Canada the rate is almost seven pounds per person per day. In some developing countries (*e.g.,* India) the average rate can be lower than one pound per person per day. These data include refuse from commercial, institutional, and industrial sources, as well as from residential sources. The actual rates of refuse generation must be carefully determined when a community plans a solid-waste management project.

Most communities require household refuse to be stored in durable, easily cleaned containers with tight-fitting covers in order to minimize rodent or insect infestation and offensive odours. Galvanized metal or plastic containers of about 30-gallon (115-litre) capacity are commonly used, although some communities employ larger containers that can be lifted mechanically and emptied into collection trucks. Plastic bags are frequently used as liners or as disposable containers for curbside collection. Where large quantities of refuse are generated—such as at shopping centres, hotels, or apartment buildings—dumpsters may be used for temporary storage until the waste is collected. Some office and commercial buildings use on-site compactors to reduce the waste volume.

SOLID-WASTE COLLECTION

Collecting and transporting. Proper solid-waste collection is important for the protection of public health, safety, and environmental quality. It is a labour-intensive activity, accounting for approximately three-quarters of the total cost of solid-waste management. Public employees are often assigned to the task, but sometimes it is more economical for private companies to do the work under contract to the municipality or for private collectors to be paid by individual homeowners. A driver and one or two loaders serve each collection vehicle. These are typically trucks of the enclosed, compacting type, with capacities up to 40 cubic yards (30 cubic metres). Loading can be done from the front, rear, or side. Compaction reduces the volume of refuse in the truck to less than half of its loose volume.

Collection vehicles

The task of selecting an optimal collection route is a complex problem, especially for large and densely populated cities. An optimal route is one that results in the most efficient use of labour and equipment, and selecting such a route requires the application of computer analyses that account for all of the many design variables in a large and complex network. Variables include frequency of collection, haulage distance, type of service, and climate. Collection of refuse in rural areas can present a special problem, since the population densities are low, leading to high unit costs.

Refuse collection usually occurs at least once per week because of the rapid decomposition of food wastes. The amount of garbage in the refuse of an individual home can be reduced by garbage grinders, or garbage disposals. Ground garbage puts an extra load on sewerage systems, but this can usually be accommodated. Many communities now conduct source separation and recycling programs, in which homeowners and businesses separate recyclable materials from garbage and place them in separate containers for collection. In addition, some communities have drop-off centres where residents can bring recyclables.

Transfer stations. If the final destination of the refuse is not near the community in which it is generated, one or more transfer stations may be necessary. A transfer station is a central facility where refuse from many collection vehicles is combined into a larger vehicle, such as a tractor-trailer unit. Open-top trailers are designed to carry about 100 cubic yards of uncompacted waste to a regional processing or disposal location. Closed compacter-type trailers are also available, but they must be equipped with ejector mechanisms. In a direct discharge type of station, several collection trucks empty directly into the transport vehicle. In a storage discharge type of station, refuse is first emptied into a storage pit or onto a platform, and then machinery is used to hoist or push the solid waste into the transport vehicle. Large transfer stations can handle more than 500 tons of refuse per day.

SOLID-WASTE TREATMENT

Once collected, municipal solid waste may be treated in order to reduce the total volume and weight of material that requires final disposal. Treatment changes the form of the waste and makes it easier to handle. It can also serve to recover certain materials, as well as heat energy, for recycling or reuse.

Incineration. *Furnace operation.* Burning is a very effective method of reducing the volume and weight of solid waste. In modern incinerators, the waste is burned inside a properly designed furnace under very carefully controlled conditions. The combustible portion of the waste combines with oxygen, releasing mostly carbon dioxide, water vapour, and heat. Incineration can reduce the volume of uncompacted waste by more than 90 percent, leaving an inert residue of ash, glass, metal, and other solid materials called bottom ash. The gaseous by-products of incomplete combustion, along with finely divided particulate material called fly ash, are carried along in the incinerator air stream. Fly ash includes cinders, dust, and soot. In order to remove fly ash and gaseous by-products before they are exhausted into the atmosphere, modern incinerators must be equipped with extensive emission-control devices. Such devices include fabric baghouse filters, acid gas scrubbers, and electrostatic precipitators. Bottom ash and fly ash are usually combined and disposed of in a landfill. If the ash is found to contain toxic metals, then it must be managed as a hazardous waste.

Bottom ash and fly ash

Municipal solid-waste incinerators are designed to receive and burn a continuous supply of refuse. A deep refuse storage pit, or tipping area, provides enough space for about one day of waste storage. The refuse is lifted from the pit by a crane equipped with a bucket or grapple device. It is then deposited into a hopper and chute above the furnace and released onto a charging grate or stoker. The grate shakes and moves waste through the furnace, allowing air to circulate around the burning material. Modern incinerators are usually built with a rectangular furnace, although rotary kiln furnaces and vertical circular furnaces are available. Furnaces are constructed of refractory bricks that can withstand the high combustion temperatures.

Combustion in a furnace occurs in two stages: primary and secondary. In primary combustion, moisture is driven off and the waste is ignited and volatilized. In secondary combustion, the remaining unburned gases and particulates are oxidized, eliminating odours and reducing the amount of fly ash in the exhaust. When the refuse is very moist, auxiliary gas or fuel oil is sometimes burned to start the primary combustion.

In order to provide enough oxygen for both primary and secondary combustion, air must be thoroughly mixed with the burning refuse. Air is supplied from openings beneath the grates or is admitted to the area above. The relative amounts of this underfire air and "overfire air" must be determined by the plant operator to achieve good combustion efficiency. A continuous flow of air can be maintained by a natural draft in a tall chimney or by mechanical forced-draft fans.

Energy recovery. The energy value of refuse can be as much as one-third that of coal, depending on the paper content, and the heat given off during incineration can be recovered by the use of a refractory-lined furnace coupled to a boiler. Boilers convert the heat of combustion into steam or hot water, thus allowing the energy content of the refuse to be recycled. Incinerators that recycle heat energy in this way are called waste-to-energy plants. Instead of a separate furnace and boiler, a water-tube wall furnace may also be used for energy recovery. Such a furnace is lined with vertical steel tubes spaced closely enough to form continuous sections of wall. The walls are insulated on the outside in order to reduce heat loss. Water circulating through the tubes absorbs heat to produce steam, and it also helps to control combustion temperatures without the need for excessive air, thus lowering air pollution control costs.

Waste-to-energy plants

Waste-to-energy plants operate as either mass burn or refuse-derived fuel systems. A mass burn system uses all the refuse, without prior treatment or preparation. A refuse-

derived fuel system separates combustible wastes from noncombustibles such as glass and metal before burning. If a turbine is installed at the plant, both steam and electricity can be produced in a process called cogeneration.

Waste-to-energy systems are more expensive to build and operate than plain incinerators because of the need for special equipment and controls, highly skilled technical personnel, and auxiliary fuel systems. On the other hand, the sale of generated steam or electricity offsets much of the extra cost, and recovery of heat energy from refuse is a viable solid-waste management option from both an engineering and an economic point of view. About 80 percent of municipal refuse incinerators in the United States are waste-to-energy facilities.

Composting. Another method of treating municipal solid waste is composting, a biological process in which the organic portion of refuse is allowed to decompose under carefully controlled conditions. Microbes metabolize the organic waste material and reduce its volume by as much as 50 percent. The stabilized product is called compost or humus; it resembles potting soil in texture and odour and may be used as a soil conditioner or mulch.

Composting offers a method of processing and recycling both garbage and sewage sludge in one operation. As more stringent environmental rules and siting constraints limit the use of solid-waste incineration and landfill options, the application of composting is likely to increase. The steps involved in the process include sorting and separating, size reduction, and digestion of the refuse.

Sorting and shredding. The decomposable materials in refuse are isolated from glass, metal, and other inorganic items through sorting and separating operations. These are carried out mechanically, using differences in such physical characteristics of the refuse as size, density, and magnetic properties. Shredding or pulverizing reduces the size of the waste articles, resulting in a uniform mass of material. It is accomplished with hammermills and rotary shredders.

Digesting and processing. Pulverized waste is ready for composting either by the open windrow method or in an enclosed mechanical facility. Windrows are long, low mounds of refuse. They are turned or mixed every few days to provide air for the microbes digesting the organics. Depending on moisture conditions, it may take five to eight weeks for complete digestion of the waste. Because of the metabolic action of aerobic bacteria, temperatures in an active compost pile reach about 150° F (65° C), killing pathogenic organisms that may be in the waste material.

Open windrow composting

Open windrow composting requires relatively large land areas. Enclosed mechanical composting facilities can reduce land requirements by about 85 percent. Mechanical composting systems employ one or more closed tanks or digesters equipped with rotating vanes that mix and aerate the shredded waste. Complete digestion of the waste takes about one week.

Digested compost must be processed before it can be used as a mulch or soil conditioner. Processing includes drying, screening, and granulating or pelletizing. These steps improve the market value of the compost, which is the most serious constraint to the success of composting as a waste management option. Agricultural demand for digested compost is usually low because of the high cost of transporting it and because of competition with inorganic chemical fertilizers.

Sanitary landfill. Land disposal is the most common management strategy for municipal solid waste. Refuse can be safely deposited in a sanitary landfill, a disposal site that is carefully selected, designed, constructed, and operated to protect the environment and public health. One of the most important factors relating to landfilling is that the buried waste never comes in contact with surface water or groundwater. Engineering design requirements include a minimum distance between the bottom of the landfill and the seasonally high groundwater table. Most new landfills are required to have an impermeable liner or barrier at the bottom, as well as a system of groundwater monitoring wells. Completed landfill sections also must be "capped" with an impermeable cover to keep precipitation or surface runoff away from the buried waste. Bottom and cap liners may be made of flexible plastic membranes, layers of clay soil, or a combination of both.

Constructing the landfill. The basic element of a sanitary landfill is the refuse cell. This is a confined portion of the site in which refuse is spread and compacted in thin layers; several layers may be compacted on top of one another to a maximum depth of about 10 feet. The compacted refuse occupies about one-quarter of its original loose volume. At the end of each day's operation, the refuse is covered with a layer of soil to eliminate windblown litter, odours, and insect or rodent problems. One refuse cell thus contains the daily volume of compacted refuse and soil cover. Several adjacent refuse cells make up a lift, and eventually a landfill may comprise two or more lifts stacked one on top of the other. The final cap for a completed landfill may also be covered with a layer of topsoil that can support vegetative growth.

Daily cover soil may be available on-site, or it may be hauled in and stockpiled from off-site sources. Various types of heavy machinery, such as crawler tractors or rubber-tired dozers, are used to spread and compact the refuse and soil. Heavy steel-wheeled compactors may also be employed to achieve high-density compaction of the refuse.

The area and depth of a new landfill is carefully staked out, and the base is prepared for construction of any required liner and leachate collection system. Where a plastic liner is used, at least 12 inches of sand is carefully spread over it to provide protection from landfill vehicles. At sites where excavations can be made below grade, the trench method of construction may be followed (Figure 52). Where this is not feasible because of topography or groundwater conditions, the area method may be practiced, resulting in a mound or hill rising above the original ground. Since no ground is excavated in the area method, soil usually must be hauled to the site from some other location. Variations of the area method may be employed where a landfill site is located on sloping ground, in a valley, or in a ravine; the completed landfill eventually blends in with the landscape.

Trench method and area method of landfilling

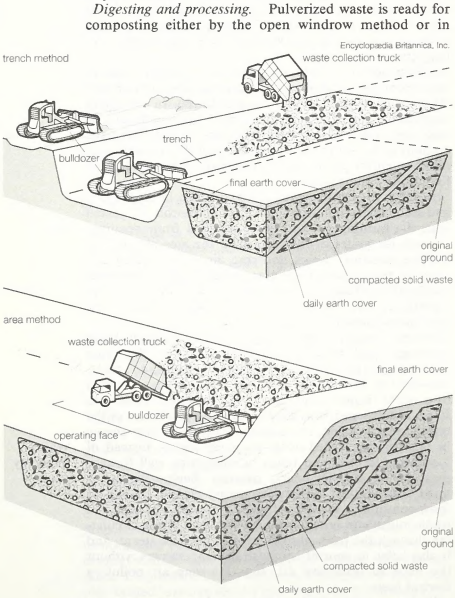


Figure 52: Two methods of constructing a sanitary landfill. The top and bottom liners and the leachate collection systems are not shown.

Controlling by-products. Organic material buried in a landfill decomposes by anaerobic microbial action. Complete decomposition usually takes more than 20 years. One of the by-products of this decomposition is methane gas. Methane is poisonous and explosive when diluted in the air, and it can flow long distances through porous layers of soil. If it is allowed to collect in basements or other confined areas, dangerous conditions may arise. In modern landfills, methane movement is controlled by impermeable barriers and by gas venting systems. In some landfills, the methane gas is collected and recovered for use as a fuel.

A highly contaminated liquid called leachate is another by-product of decomposition in sanitary landfills. Most leachate is the result of runoff that infiltrates the refuse cells and comes in contact with decomposing garbage. If leachate reaches the groundwater or seeps out onto the ground surface, serious environmental pollution problems can occur, including the possible contamination of drinking water supplies. Methods of controlling leachate include the interception of surface water in order to prevent it from entering the landfill and the use of impermeable liners or barriers between the waste and the groundwater. New landfill sites should also be provided with groundwater monitoring wells and leachate collection and treatment systems.

Importance in waste management. In communities where appropriate sites are available, sanitary landfills usually provide the most economical option for refuse disposal. However, it is becoming increasingly difficult to find sites that offer adequate capacity, accessibility, and environmental conditions. Nevertheless, landfills will always play a key role in solid-waste management. It is not possible to recycle all components of solid waste, and there will always be residues from incineration and other treatment processes that will eventually require disposal underground. In addition, landfills can actually improve poor-quality land. In some communities, properly completed landfills are converted into recreational parks, playgrounds, or golf courses.

Recycling. Separating, recovering, and reusing components of solid waste that may still have economic value is called recycling. One type of recycling is the recovery and reuse of heat energy, a practice discussed above (see *Incineration*). Composting can also be considered a recycling process, since it reclaims the organic parts of solid waste for reuse as mulch or soil conditioner. Still other waste materials have potential for reuse; these include paper, metal, glass, plastic, and rubber, and their recovery is discussed here.

Separation. Before any material can be recycled, it must be separated from the raw waste and sorted. Separation can be accomplished at the source of the waste or at a central processing facility. Source separation, also called curbside separation, is done by individual citizens who collect newspapers, bottles, cans, and garbage separately and place them at the curb for collection. Many communities allow "commingling" of nonpaper recyclables (glass, metal, and plastic). In either case, municipal collection of source-separated refuse is more expensive than ordinary refuse collection.

In lieu of source separation, recyclable materials can be separated from garbage at centralized mechanical processing plants. Experience has shown that the quality of recyclables recovered from such facilities is lowered by contamination with moist garbage and broken glass. The best practice, as now recognized, is to have citizens separate refuse into a limited number of categories, including: newspaper; magazines and other wastepaper; commingled metals, glass, and plastics; and, finally, garbage and other nonrecyclables. The newspaper, other paper wastes, and commingled recyclables are collected separately from the other refuse and are processed at a centralized material recycling facility, or MRF (prounced "murf" in waste-management jargon). A modern MRF can process about 300 tons of recyclable wastes per day.

Material recycling facilities (MRFs)

At a typical MRF, commingled recyclables are loaded onto a conveyor. Steel cans ("tin" cans are actually steel with only a thin coating of tin) are removed by an electromagnetic separator, and the remaining material passes over a vibrating screen in order to remove broken glass. Next, the conveyor passes through an air classifier, which separates aluminum and plastic containers from heavier glass containers. Glass is hand sorted by colour, and aluminum cans are separated from plastics by an eddy-current separator, which repels the aluminum from the conveyor belt.

Reuse. Recovered broken glass can be crushed and used in asphalt pavement while colour-sorted glass is also crushed and sold to glass manufacturers as cullet, an essential ingredient in glassmaking. Steel cans are baled and shipped to steel mills as scrap, and aluminum is baled or compacted for reuse by smelters. Aluminum is one of the smallest components of municipal solid waste, but it has the highest value as a recyclable material. Recycling of plastic is a challenge, mostly because of the many different polymeric materials used in its production. Mixed thermoplastics can be used only to make lower quality products, such as "plastic lumber."

In the paper stream, old newspapers are hand sorted on a conveyor belt in order to remove corrugated materials and mixed papers. It is then baled or loose-loaded into trailers for shipment to paper mills, where it is reused in the making of more newspaper. Mixed paper is separated from corrugated paper for sale to tissue mills. Although the processes of pulping, de-inking, and screening wastepaper is generally more expensive than making paper from virgin wood fibres, the market for recycled paper should improve as more processing plants are established.

Rubber is sometimes reclaimed from solid waste and shredded, reformed, and remolded in a process called revulcanization, but it is usually not as strong as the original material. Shredded rubber can be used as an additive in asphalt pavements, and discarded tires may be employed in "tire playgrounds." In general, the most difficult problem associated with the recycling of any solid-waste material is finding applications and suitable markets. Recycling by itself will not solve the growing problem of solid-waste management and disposal. There will always be some unusable and completely valueless solid residue requiring final disposal.

© Craig Hammell/The Stock Market



Figure 53: Hand sorting of paper waste at a recycling facility in Elizabeth, N.J.

Hazardous-waste management

Hazardous waste is any waste material that, when improperly handled, can cause substantial harm to human health and safety or to the environment. Hazardous wastes can take the form of solids, liquids, sludges, or contained gases, and they are generated primarily by chemical production, manufacturing, and other industrial activities. They may cause damage during inadequate storage, transportation, treatment, or disposal operations. Improper waste storage or disposal frequently contaminates surface and groundwater supplies. People living in homes built near old and abandoned waste disposal sites may be in a particularly vulnerable position. In an effort to remedy existing problems and to prevent future harm from hazardous wastes, governments closely regulate the practice of hazardous-waste management.

HAZARDOUS-WASTE CHARACTERISTICS

Hazardous wastes are classified on the basis of their biological, chemical, and physical properties. These properties generate materials that are either toxic, reactive, ignitable, corrosive, infectious, or radioactive.

Toxic wastes

Toxic wastes are poisons, even in very small or trace amounts. They may have acute effects, causing death or violent illness, or they may have chronic effects, slowly causing irreparable harm. Some are carcinogenic, causing cancer after many years of exposure. Others are mutagenic, causing major biological changes in the offspring of exposed humans and wildlife.

Reactive wastes are chemically unstable and react violently with air or water. They cause explosions or form toxic vapours. Ignitable wastes burn at relatively low temperatures and may cause an immediate fire hazard. Corrosive wastes include strong acidic or alkaline substances. They destroy solid material and living tissue upon contact, by chemical reaction.

Infectious wastes include used bandages, hypodermic needles, and other materials from hospitals or biological research facilities. Radioactive wastes emit ionizing energy that can harm living organisms. Because some radioactive materials can persist in the environment for many thousands of years before fully decaying, there is much concern over the control of these wastes. However, the handling and disposal of radioactive material is not a responsibility of local municipal government. Owing to the scope and complexity of the problem, the management of radioactive waste (particularly nuclear-fission waste) is usually considered to be a separate engineering task from other forms of hazardous-waste management and is not discussed in this article.

TRANSPORT OF HAZARDOUS WASTE

Hazardous waste generated at a particular site often requires transport to an approved treatment, storage, or disposal facility (TSDF). Because of potential threats to public safety and the environment, transport is given special attention by governmental agencies. In addition to the occasional accidental spill, hazardous waste has, in the past, been intentionally spilled or abandoned at random locations in a practice known as "midnight dumping." This practice has been greatly curtailed by the enactment of laws that require proper labeling, transport, and tracking of all hazardous wastes.

Transport vehicles. Hazardous waste is generally transported by truck over public highways. Only a very small amount is transported by rail, and almost none is moved by air or inland waterway. Highway shipment is the most common because road vehicles can gain access to most industrial sites and approved TSDFs. Railroad trains require expensive siding facilities and are suitable only for very large waste shipments.

Hazardous wastes can be shipped in tank trucks made of steel or aluminum alloy, with capacities up to about 9,000 gallons (34,000 litres). They also can be containerized and shipped in 55-gallon (200-litre) drums. Specifications and standards for cargo tank trucks and shipping containers are included in governmental regulations.

The manifest system. In the United States a key feature of regulations pertaining to waste transport is the "cradle-to-grave" manifest system, which monitors the journey of hazardous waste from its point of origin to the point of final disposal. The manifest system helps to eliminate the problem of midnight dumping. It also provides a means for determining the type and quantity of hazardous waste being generated as well as the recommended emergency procedures in case of an accidental spill. A manifest is a record-keeping document that must be prepared by the generator of the hazardous waste, such as a chemical manufacturer. The generator has primary responsibility for the ultimate disposal of the waste and must give the manifest, along with the waste itself, to a licensed waste transporter. A copy of the manifest must be delivered by the transporter to the recipient of the waste at an authorized TSDF. Each time the waste changes hands, a copy of the manifest must be signed. Copies of the manifest are kept by each party involved, and additional copies are sent to appropriate environmental agencies.

Monitoring transport of hazardous waste

In the event of a leak or accidental spill of hazardous waste during its transport, the transporter must take immediate and appropriate actions, including notifying local authorities of the discharge. An area may have to be diked to contain the wastes, and efforts must be undertaken to remove the wastes and reduce environmental or public health hazards.

TREATMENT, STORAGE, AND DISPOSAL

Several options are available for hazardous-waste management. The most desirable is to reduce the quantity of waste at its source or to recycle the materials for some other productive use. Nevertheless, while reduction and recycling are desirable options, they are not regarded as the final remedy to the problem of hazardous-waste disposal. There will always be a need for treatment and for storage or disposal of some amount of hazardous waste.

Treatment. Hazardous waste can be treated by chemical, thermal, biological, and physical methods. Chemical methods include ion exchange, precipitation, oxidation and reduction, and neutralization. Among thermal methods is high-temperature incineration, which can not only detoxify certain organic wastes but also destroy them as well. Special types of thermal equipment are used for burning waste in either a solid, liquid, or sludge form. These include the fluidized-bed incinerator, multiple-hearth furnace, rotary kiln, and liquid-injection incinerator. One problem posed by hazardous waste incineration is the potential for air pollution.

Biological treatment of certain organic wastes, such as those from the petroleum industry, is also an option. One method used to treat hazardous waste biologically is called landfarming. In this technique the waste is carefully mixed with surface soil on a suitable tract of land. Microbes that can metabolize the waste may be added, along with nutrients. In some cases, a genetically engineered species of bacteria is used. Food or forage crops are not grown on the same site. Microbes can also be used for stabilizing hazardous wastes on previously contaminated sites; in that case the process is called bioremediation.

The chemical, thermal, and biological treatment methods outlined above change the molecular form of the waste material. Physical treatment, on the other hand, concentrates, solidifies, or reduces the volume of the waste. Physical processes include evaporation, sedimentation, flotation, and filtration. Yet another process is solidification, which is achieved by encapsulating the waste in concrete, asphalt, or plastic. Encapsulation produces a solid mass of material that is resistant to leaching. Waste can also be mixed with lime, fly ash, and water to form a solid, cementlike product.

Surface storage and land disposal. Hazardous wastes that are not destroyed by incineration or other chemical processes need to be disposed of properly. For most of such wastes, land disposal is the ultimate destination, although it is not an attractive practice because of the inherent environmental risks involved. Two basic methods of land disposal include landfilling and underground injection. Prior to land disposal surface storage or containment systems are often employed as a temporary method.

Surface storage. Temporary on-site waste storage facilities include open waste piles and ponds or lagoons. New waste piles must be carefully constructed over an impervious base and must comply with regulatory requirements similar to those for landfills. The piles must be protected from wind dispersion or erosion; if leachate is generated, monitoring and control systems must be provided. Only noncontainerized solid, nonflowing waste material can be stored in a new waste pile, and the material must be landfilled when the size of the pile becomes unmanageable.

Hazardous waste lagoons

A common type of temporary storage impoundment for hazardous liquid waste is an open pit or holding pond, called a lagoon. New lagoons must be lined with impervious clay soils and flexible membrane liners in order to protect groundwater. Leachate collection systems must be installed between the liners, and groundwater monitoring wells are required. Except for some sedimentation, evaporation of volatile organics, and possibly some surface aeration, open lagoons provide no treatment of the waste. Accumulated sludge must be removed periodically and subjected to further handling as a hazardous waste.

Many older unlined waste piles and lagoons are located above aquifers used for public water supply, thus posing significant risks to public health and environmental quality. A large number of these old sites have been identified and scheduled for clean up, or "remediation."

Secure landfills. Landfilling of hazardous solid or containerized waste is regulated more stringently than landfilling of municipal solid waste. Hazardous wastes must be deposited in so-called secure landfills, which provide at least 10 feet (3 metres) of separation between the bottom of the landfill and the underlying bedrock or groundwater table. A secure hazardous-waste landfill must have two impermeable liners and leachate collection systems (Figure 54). The double leachate collection system consists of a network of perforated pipes placed above each liner. The upper system prevents the accumulation of leachate trapped in the fill, and the lower serves as a backup. Collected leachate is pumped to a treatment plant. In order to reduce the amount of leachate in the fill and minimize the potential for environmental damage, an impermeable cap or cover is placed over a finished landfill. A groundwater monitoring system that includes a series of deep wells drilled in and around the site is also required. The wells allow a routine program of sampling and testing to detect any leaks or groundwater contamination. If a leak does occur, the wells can be pumped to intercept the polluted water and bring it to the surface for treatment.

Deep-well injection. One option for the disposal of liquid hazardous waste is deep-well injection, a procedure that involves pumping liquid waste through a steel casing into a porous layer of limestone or sandstone. High pressures are applied to force the liquid into the pores and fissures of the rock, where it is to be permanently stored. The injection zone must lie below a layer of impervious rock or clay, and it may extend more than one-half mile below the surface. Deep-well injection is relatively inexpensive and requires little or no pretreatment of the waste. But it poses a danger of leaking hazardous waste and eventually polluting subsurface water supplies.

REMEDIAL ACTION

Disposal of hazardous waste in unlined pits, ponds, or lagoons poses a threat to human health and environmental quality. Many such uncontrolled disposal sites were used in the past and have been abandoned; depending on a determination of the level of risk, it may be necessary to remediate these sites. In some cases, the risk may require emergency action. In other instances, engineering studies may be required to assess the situation thoroughly before remedial action is undertaken.

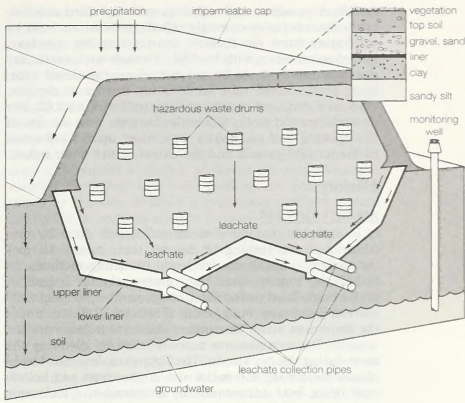


Figure 54: Schematic diagram of a secure hazardous-waste landfill with a double leachate collection system.

One option for remediation is to remove completely all the waste material from the site and transport it to another location for treatment and proper disposal. This so-called off-site solution is usually the most expensive option. An alternative is on-site remediation, which reduces the production of leachate and lessens the chance of groundwater contamination. On-site remediation may include temporary removal of the hazardous waste, construction of a secure landfill on the same site, and proper replacement of the waste. It may also include treatment of any contaminated soil or groundwater. Treated soil may be replaced on-site and treated groundwater returned to the aquifer by deep-well injection.

Containment of old waste sites

A less costly alternative is full containment of the waste. This is done by placing an impermeable cover over the hazardous waste site and by blocking the lateral flow of groundwater with subsurface cutoff walls. It is possible to use cutoff walls for this purpose when there is a natural layer of impervious soil or rock below the site. The walls are constructed around the perimeter of the site, deep enough to penetrate to the impervious layer. They can be excavated as trenches around the site without moving or disturbing the waste material. The trenches are filled with a bentonite-clay slurry to prevent their collapse during construction, and they are backfilled with a mixture of soil and cement that solidifies to form an impermeable barrier. Cutoff walls thus serve as vertical barriers to the flow of water, while the impervious layer serves as a barrier at the bottom. (J.A.N.)

TUNNELS AND UNDERGROUND EXCAVATIONS

A tunnel is an essentially horizontal underground passageway produced by excavation or occasionally by nature's action in dissolving a soluble rock, such as limestone. A vertical opening is usually called a shaft. Tunnels have many uses: for mining ores, for transportation—including road vehicles, trains, subways, and canals—and for conducting water and sewage. Underground chambers, often associated with a complex of connecting tunnels and shafts, increasingly are being used for such things as underground hydroelectric-power plants, ore-processing plants, pumping stations, vehicle parking, storage of oil and water, water-treatment plants, warehouses, and light manufacturing; also command centres and other special military needs.

True tunnels and chambers are excavated from the inside—with the overlying material left in place—and then lined as necessary to support the adjacent ground. A hillside tunnel entrance is called a portal; tunnels may also be started from the bottom of a vertical shaft or from the end of a horizontal tunnel driven principally for

construction access and called an adit. So-called cut-and-cover tunnels (more correctly called conduits) are built by excavating from the surface, constructing the structure, and then covering with backfill. Tunnels underwater are now commonly built by the use of an immersed tube: long, prefabricated tube sections are floated to the site, sunk in a prepared trench, and covered with backfill. For all underground work, difficulties increase with the size of the opening and are greatly dependent upon weaknesses of the natural ground and the extent of the water inflow.

History

ANCIENT TUNNELS

It is probable that the first tunneling was done by prehistoric people seeking to enlarge their caves. All major ancient civilizations developed tunneling methods. In Babylonia, tunnels were used extensively for irrigation; and a brick-lined pedestrian passage some 3,000 feet (900 metres) long was built about 2180 to 2160 BC under the Euphrates River to connect the royal palace with the temple. Construction was accomplished by diverting the river during the dry season. The Egyptians developed techniques for cutting soft rocks with copper saws and hollow reed drills, both surrounded by an abrasive, a technique probably used first for quarrying stone blocks and later in excavating temple rooms inside rock cliffs. Abu Simbel Temple on the Nile, for instance, was built in sandstone about 1250 BC for Ramses II (in the 1960s it was cut apart and moved to higher ground for preservation before flooding from the Aswān High Dam). Even more elaborate temples were later excavated within solid rock in Ethiopia and India.

The Greeks and Romans both made extensive use of tunnels: to reclaim marshes by drainage and for water aqueducts, such as the 6th-century-BC Greek water tunnel on the isle of Samos driven some 3,400 feet through limestone with a cross section about 6 feet square. Perhaps the largest tunnel in ancient times was a 4,800-foot-long, 25-foot-wide, 30-foot-high road tunnel (the Pausilippo) between Naples and Pozzuoli, executed in 36 BC. By that time surveying methods (commonly by string line and plumb bobs) had been introduced, and tunnels were advanced from a succession of closely spaced shafts to provide ventilation. To save the need for a lining, most ancient tunnels were located in reasonably strong rock, which was broken off (spalled) by so-called fire quenching, a method involving heating the rock with fire and suddenly cooling it by dousing with water. Ventilation methods were primitive, often limited to waving a canvas at the mouth of the shaft, and most tunnels claimed the lives of hundreds or even thousands of the slaves used as workers. In AD 41 the Romans used some 30,000 men for 10 years to push a 3.5-mile (6-kilometre) tunnel to drain Lacus Fucinus. They worked from shafts 120 feet apart and up to 400 feet deep. Far more attention was paid to ventilation and safety measures when workers were freemen, as shown by archaeological diggings at Hallstatt, Austria, where saltmine tunnels have been worked since 2500 BC.

Largest ancient tunnel

FROM THE MIDDLE AGES TO THE PRESENT

Canal and railroad tunnels. Because the limited tunneling in the Middle Ages was principally for mining and military engineering, the next major advance was to meet Europe's growing transportation needs in the 17th century. The first of many major canal tunnels was the Canal du Midi (also known as Languedoc) tunnel in France, built in 1666–81 by Pierre Riquet as part of the first canal linking the Atlantic and the Mediterranean. With a length of 515 feet and a cross section of 22 by 27 feet, it involved what was probably the first major use of explosives in public-works tunneling, gunpowder placed in holes drilled by handheld iron drills. A notable canal tunnel in England was the Bridgewater Canal Tunnel, built in 1761 by James Brindley to carry coal to Manchester from the Worsley mine. Many more canal tunnels were dug in Europe and North America in the 18th and early 19th centuries. Though the canals fell into disuse with the introduction of railroads about 1830, the new form of transport produced a huge increase in tunneling, which continued for nearly 100 years as railroads expanded over the world. Much pioneer railroad tunneling developed in England. A 3.5-mile tunnel (the Woodhead) of the Manchester-Sheffield Railroad (1839–45) was driven from five shafts up to 600 feet deep. In the United States, the first railroad tunnel was a 701-foot construction on the Allegheny Portage Railroad. Built in 1831–33, it was a combination of canal and railroad systems, carrying canal barges over a summit. Though plans for a transport link from Boston to the Hudson River had first called for a canal tunnel to pass under the Berkshire Mountains, by 1855, when the Hoosac Tunnel was started, railroads had already established their worth, and the plans were changed to a double-track railroad bore 24 by 22 feet and 4.5 miles long. Initial estimates contemplated completion in 3 years; 21 were actually required, partly because the rock proved too hard for either hand drilling or a primitive power saw. When the state of Massachusetts finally took over the project, it completed it in 1876 at five times the originally estimated cost. Despite frustrations, the Hoosac Tunnel contributed notable advances in tunneling, including one of the first uses of dynamite, the first use of electric firing of explosives, and the introduction of power drills, initially steam and later air, from which there ultimately developed a compressed-air industry.

First use of explosives

Simultaneously, more spectacular railroad tunnels were being started through the Alps. The first of these, the Mont Cenis Tunnel (also known as Fréjus), required 14 years (1857–71) to complete its 8.5-mile length. Its engineer, Germain Sommeiller, introduced many pioneering techniques, including rail-mounted drill carriages, hydraulic ram air compressors, and construction camps for workers complete with dormitories, family housing, schools, hospitals, a recreation building, and repair shops. Sommeiller also designed an air drill that eventually made it possible to move the tunnel ahead at the rate of 15 feet per day and was used in several later European tunnels until replaced by more durable drills developed in the United States by Simon Ingersoll and others on the Hoosac Tunnel. As this long tunnel was driven from two headings separated by 7.5 miles of mountainous terrain, surveying techniques had to be refined. Ventilation became a major problem, which was solved by the use of forced air from water-powered fans and a horizontal diaphragm at mid-height, forming an exhaust duct at top of the tunnel. Mont Cenis was soon followed by other notable Alpine railroad tunnels: the 9-mile St. Gotthard (1872–82), which introduced compressed-air locomotives and suffered major problems with water inflow, weak rock, and bankrupt contractors; the 12-mile Simplon (1898–1906); and the 9-mile Lötschberg (1906–11), on a northern continuation of the Simplon railroad line.

Tunnels in the Alps

Nearly 7,000 feet below the mountain crest, Simplon encountered major problems from highly stressed rock flying off the walls in rock bursts; high pressure in weak schists and gypsum, requiring 10-foot-thick masonry lining to resist swelling tendencies in local areas; and from high-temperature water (130° F [54° C]), which was partly treated by spraying from cold springs. Driving Simplon as two parallel tunnels with frequent crosscut connections considerably aided ventilation and drainage.

Lötschberg disaster and its effect

Lötschberg was the site of a major disaster in 1908. When one heading was passing under the Kander River valley, a sudden inflow of water, gravel, and broken rock filled the tunnel for a length of 4,300 feet, burying the entire crew of 25 men. Though a geologic panel had predicted that the tunnel here would be in solid bedrock far below the bottom of the valley fill, subsequent investigation showed that bedrock lay at a depth of 940 feet, so that at 590 feet the tunnel tapped the Kander River, allowing it and soil of the valley fill to pour into the tunnel, creating a huge depression, or sink, at the surface. After this lesson in the need for improved geologic investigation, the tunnel was rerouted about one mile (1.6 kilometres) upstream, where it successfully crossed the Kander Valley in sound rock.

Most long-distance rock tunnels have encountered problems with water inflows. One of the most notorious was the first Japanese Tanna Tunnel, driven through the Takiji

Peak in the 1920s. The engineers and crews had to cope with a long succession of extremely large inflows, the first of which killed 16 men and buried 17 others, who were rescued after seven days of tunneling through the debris. Three years later another major inflow drowned several workers. In the end, Japanese engineers hit on the expedient of digging a parallel drainage tunnel the entire length of the main tunnel. In addition, they resorted to compressed-air tunneling with shield and air lock (see below), a technique almost unheard-of in mountain tunneling.

Subaqueous tunnels. Tunneling under rivers was considered impossible until the protective shield was developed in England by Marc Brunel, a French émigré engineer. The first use of the shield, by Brunel and his son Isambard, was in 1825 on the Wapping-Rotherhithe Tunnel through clay under the Thames River. The tunnel was of horseshoe section 22¼ by 37½ feet and brick-lined. After several floodings from hitting sand pockets and a seven-year shutdown for refinancing and building a second shield, the Brunels succeeded in completing the world's first true subaqueous tunnel in 1841, essentially nine years' work for a 1,200-foot-long tunnel. In 1869 by reducing to a small size (8 feet) and by changing to a circular shield plus a lining of cast-iron segments, Peter W. Barlow and his field engineer, James Henry Greathead, were able to complete a second Thames tunnel in only one year as a pedestrian walkway from Tower Hill. In 1874, Greathead made the subaqueous technique really practical by refinements and mechanization of the Brunel-Barlow shield and by adding compressed air pressure inside the tunnel to hold back the outside water pressure. Compressed air alone was used to hold back the water in 1880 in a first attempt to tunnel under New York's Hudson River; major difficulties and the loss of 20 lives forced abandonment after only 1,600 feet had been excavated.

Greathead technique

The first major application of the shield-plus-compressed-air technique occurred in 1886 on the London subway with an 11-foot bore, where it accomplished the unheard-of record of seven miles of tunneling without a single fatality. So thoroughly did Greathead develop his procedure that it was used successfully for the next 75 years with no significant change. A modern Greathead shield illustrates his original developments as pictured in Figure 55: miners working under a hood in individual small pockets that can be quickly closed against inflow; shield propelled forward by jacks; permanent lining segments erected under protection of the shield tail; and the whole tunnel pressurized to resist water inflow.

Once subaqueous tunneling became practical, many railroad and subway crossings were constructed with the Greathead shield, and the technique later proved adaptable for the much larger tunnels required for automobiles. A new problem, noxious gases from internal-combustion engines, was successfully solved by Clifford Holland for the world's first vehicular tunnel, completed in 1927 under the Hudson River and now bearing his name. Holland and his chief engineer, Ole Singstad, solved the ventilation problem with huge-capacity fans in ventilating buildings at each end, forcing air through a supply duct below the roadway, with an exhaust duct above the ceiling. Such ventilation provisions significantly increased the tunnel size, requiring about a 30-foot diameter for a two-lane vehicular tunnel.

Many similar vehicular tunnels were built by shield-and-compressed-air methods—including Lincoln and Queens tunnels in New York City, Sumner and Callahan in Boston, and Mersey in Liverpool. Since 1950, however, most subaqueous tunnelers preferred the immersed-tube method, in which long tube sections are prefabricated, towed to the site, sunk in a previously dredged trench, connected to sections already in place, and then covered with backfill. This basic procedure was first used in its present form on the Detroit River Railroad Tunnel between Detroit and Windsor, Ontario (1906–10). A prime advantage is the avoidance of high costs and the risks of operating a shield under high air pressure, since work inside the sunken tube is at atmospheric pressure (free air).

Machine-mined tunnels. Sporadic attempts to realize the tunnel engineer's dream of a mechanical rotary excavator culminated in 1954 at Oahe Dam on the Missouri River near Pierre, in South Dakota. With ground conditions being favourable (a readily cuttable clay-shale), success resulted from a team effort: Jerome O. Ackerman as chief engineer, F.K. Mittry as initial contractor, and James S. Robbins as builder of the first machine—the "Mittry Mole." Later contracts developed three other Oahe-type moles, so that all the various tunnels here were machine-mined—totaling eight miles of 25- to 30-foot diameter. These were the first of the modern moles that since 1960 have been rapidly adopted for many of the world's tunnels as a means of increasing speeds from the previous range of 25 to 50 feet per day to a range of several hundred feet per day. The Oahe mole was partly inspired by work on a pilot tunnel in chalk started under the English Channel for which an air-powered rotary cutting arm, the Beaumont borer, had been invented. A 1947 coal-mining version followed, and in 1949 a coal saw was used to cut a circumferential slot in chalk for 33-foot-diameter tunnels at Fort Randall Dam in South Dakota. In 1962 a comparable breakthrough for the more difficult excavation of vertical shafts was achieved in the American development of the mechanical raise borer, profiting from earlier trials in Germany.

Tunneling techniques

BASIC TUNNELING SYSTEM

Four categories of tunnels

Tunnels are generally grouped in four broad categories, depending on the material through which they pass: soft ground, consisting of soil and very weak rock; hard rock; soft rock, such as shale, chalk, and friable sandstone; and subaqueous. While these four broad types of ground condition require very different methods of excavation and ground support, nevertheless, nearly all tunneling

By courtesy of (right) Port of New York Authority, from (left) Richardson and Mayo, *Practical Tunnel Driving* (copyright 1941); used with permission of McGraw-Hill Book Company

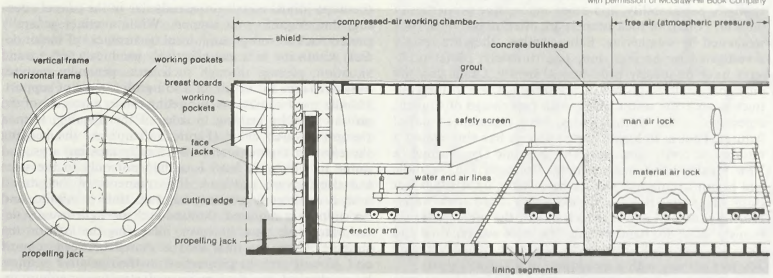


Figure 55: *Tunneling shield, a basic engineering tool for underwater and soft-ground tunneling.* (Left) Front view. (Right) Cross section, showing shield advancing, with lining being emplaced behind it, and air locks for entry into compressed-air working chamber.

operations involve certain basic procedures: investigation, excavation and materials transport, ground support, and environmental control. Similarly, tunnels for mining and for civil-engineering projects share the basic procedures but differ greatly in the design approach toward permanence, owing to their differing purposes. Many mining tunnels have been planned only for minimum-cost temporary use during ore extraction, although the growing desire of surface owners for legal protection against subsequent tunnel collapse may cause this to change. By contrast, most civil-engineering or public-works tunnels involve continued human occupancy plus full protection of adjacent owners and are much more conservatively designed for permanent safety. In all tunnels, geologic conditions play the dominant role in governing the acceptability of construction methods and the practicality of different designs. Indeed, tunneling history is filled with instances in which a sudden encounter with unanticipated conditions caused long stoppages for changes in construction methods, in design, or in both, with resulting great increases in cost and time. At the Awali Tunnel in Lebanon in 1960, for example, a huge flow of water and sand filled over 2 miles of the bore and more than doubled construction time to eight years for its 10-mile length.

Geologic investigation. Thorough geologic analysis is essential in order to assess the relative risks of different locations and to reduce the uncertainties of ground and water conditions at the location chosen. In addition to soil and rock types, key factors include the initial defects controlling behaviour of the rock mass; size of rock block between joints; weak beds and zones, including faults, shear zones, and altered areas weakened by weathering or thermal action; groundwater, including flow pattern and pressure; plus several special hazards, such as heat, gas, and earthquake risk. For mountain regions the large cost and long time required for deep borings generally limit their number; but much can be learned from thorough aerial and surface surveys, plus well-logging and geophysical techniques developed in the oil industry. Often the problem is approached with flexibility toward changes in design and in construction methods and with continuous exploration ahead of the tunnel face, done in older tunnels by mining a pilot bore ahead and now by drilling. Japanese engineers have pioneered methods for prelocating troublesome rock and water conditions.

Geologic problems in rock chambers

For large rock chambers and also particularly large tunnels, the problems increase so rapidly with increasing opening size that adverse geology can make the project impractical or at least tremendously costly. Hence, the concentrated opening areas of these projects are invariably investigated during the design stage by a series of small exploratory tunnels called drifts, which also provide for in-place field tests to investigate engineering properties of the rock mass and can often be located so their later enlargement affords access for construction.

Since shallow tunnels are more often in soft ground, borings become more practical. Hence, most subways involve borings at intervals of 100–500 feet to observe the water table and to obtain undisturbed samples for testing strength, permeability, and other engineering properties of the soil. Portals of rock tunnels are often in soil or in rock weakened by weathering. Being shallow, they are readily investigated by borings, but, unfortunately, portal problems have frequently been treated lightly. Often they are only marginally explored or the design is left to the contractor, with the result that a high percentage of tunnels, especially in the United States, have experienced portal failures. Failure to locate buried valleys has also caused a number of costly surprises. The five-mile Oso Tunnel in New Mexico offers one example. There, in 1967, a mole had begun to progress well in hard shale, until 1,000 feet from the portal it hit a buried valley filled with water-bearing sand and gravel, which buried the mole. After six months' delay for hand mining, the mole was repaired and soon set new world records for advance rate—averaging 240 feet per day with a maximum of 420 feet per day.

Excavation and materials handling. Excavation of the ground within the tunnel bore may be either semicontinuous, as by handheld power tools or mining machine, or cyclic, as by drilling and blasting methods for harder rock. Here each cycle involves drilling, loading explosive, blasting, ventilating fumes, and excavation of the blasted rock (called mucking). Commonly, the mucker is a type of front-end loader that moves the broken rock onto a belt conveyor that dumps it into a hauling system of cars or trucks. As all operations are concentrated at the heading, congestion is chronic, and much ingenuity has gone into designing equipment able to work in a small space. Since progress depends on the rate of heading advance, it is often facilitated by mining several headings simultaneously, as opening up intermediate headings from shafts or from adits driven to provide extra points of access for longer tunnels.

Removing waste material

For smaller diameters and longer tunnels, a narrow-gauge railroad is commonly employed to take out the muck and bring in workers and construction material. For larger-size bores of short to moderate length, trucks are generally preferred. For underground use these require diesel engines with scrubbers to eliminate dangerous gases from the exhaust. While existing truck and rail systems are adequate for tunnels progressing in the range of 40–60 feet (12–18 metres) per day, their capacity is inadequate to keep up with fast-moving moles progressing at the rate of several hundred feet per day. Hence, considerable attention is being devoted to developing high-capacity transport systems—continuous-belt conveyors, pipelines, and innovative rail systems (high-capacity cars on high-speed trains). Muck disposal and its transport on the surface can also be a problem in congested urban areas. One solution successfully applied in Japan is to convey it by pipeline to sites where it can be used for reclamation by landfill.

For survey control, high-accuracy transit-level work (from base lines established by mountaintop triangulation) has generally been adequate; long tunnels from opposite sides of the mountain commonly meet with an error of one foot or less. Further improvements are likely from the recent introduction of the laser, the pencil-size light beam of which supplies a reference line readily interpreted by workers. Most moles in the United States now use a laser beam to guide steering, and some experimental machines employ electronic steering actuated by the laser beam.

Ground support. The dominant factor in all phases of the tunneling system is the extent of support needed to hold the surrounding ground safely. Engineers must consider the type of support, its strength, and how soon it must be installed after excavation. The key factor in timing support installation is so-called stand-up time—*i.e.,* how long the ground will safely stand by itself at the heading, thus providing a period for installing supports. In soft ground, stand-up time can vary from seconds in such soils as loose sand up to hours in such ground as cohesive clay and even drops to zero in flowing ground below the water table, where inward seepage moves loose sand into the tunnel. Stand-up time in rock may vary from minutes in raveling ground (closely fractured rock where pieces gradually loosen and fall) up to days in moderately jointed rock (joint spacing in feet) and may even be measured in centuries in nearly intact rock, where the rock-block size (between joints) equals or exceeds size of the tunnel opening, thus requiring no support. While a miner generally prefers rock to soft ground, local occurrences of major defects within the rock can effectively produce a soft-ground situation; passage through such areas generally requires radical change to the use of a soft-ground type of support.

Stand-up time

Under most conditions, tunneling causes a transfer of the ground load by arching to sides of the opening, termed the ground-arch effect (Figure 56, top). At the heading the effect is three-dimensional, locally creating a ground dome in which the load is arched not only to the sides but also forward and back. If permanence of the ground arch is completely assured, stand-up time is infinite, and no support is required. Ground-arch strength usually deteriorates with time, however, increasing the load on the support. Thus, the total load is shared between support and ground arch in proportion to their relative stiffness by a physical mechanism termed structure-medium interaction. The support load increases greatly when the inherent ground strength is much reduced by allowing

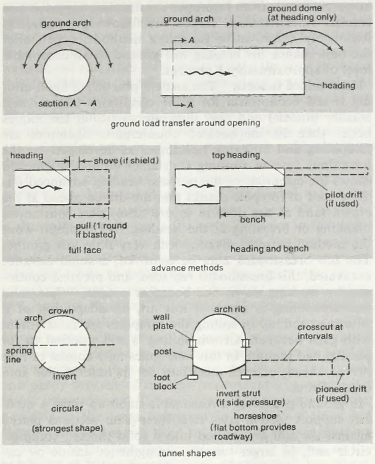


Figure 56: Tunnel terminology (see text).

excessive yield to loosen the rock mass. Because this may occur when installation of support is delayed too long, or because it may result from blast damage, good practice is based on the need to preserve the strength of the ground arch as the strongest load-carrying member of the system, by prompt installation of proper support and by preventing blast damage and movement from water inflow that has a tendency to loosen the ground.

Because stand-up time drops rapidly as size of the opening increases, the full-face method of advance (Figure 56, centre), in which the entire diameter of the tunnel is excavated at one time, it is most suitable for strong ground or for smaller tunnels. The effect of weak ground can be offset by decreasing the size of opening initially mined and supported, as in the top heading and bench method of advance. For the extreme case of very soft ground, this approach results in the multiple-drift method of advance (Figure 57), in which the individual drifts are reduced to a small size that is safe for excavation and portions of the support are placed in each drift and progressively connected as the drifts are expanded. The central core is left unexcavated until sides and crown are safely supported, thus providing a convenient central buttress for bracing the temporary support in each individual drift. While this obviously slow multidrift method is an old technique for very weak ground, such conditions still force its adoption as a last resort in some modern tunnels. In 1971, for example, on the Straight Creek interstate highway tunnel in Colorado, a very complex pattern of multiple drifts was found necessary to advance this large horseshoe-shaped tunnel 42 by 45 feet high through a weak shear zone more than 1,000 feet wide, after unsuccessful trials with full-face operation of a shield.

Multiple-drift method

In early tunnels, timber was used for the initial or temporary support, followed by a permanent lining of brick or stone masonry. Since steel became available, it has been widely used as the first temporary stage or primary support. For protection against corrosion, it is nearly always encased in concrete as a second stage or final lining. Steel-rib support with timber blocking outside has been widely employed in rock tunnels. The horseshoe shape is common for all but the weakest rocks, since the flat bottom facilitates hauling. By contrast, the stronger and more structurally efficient circular shape is generally required to support the greater loads from soft ground. Figure 56, bottom, compares these two shapes and indicates a number of terms identifying various parts of the cross section and adjacent members for a steel-rib type of support. Here a wall plate is generally used only with a top heading method, where it serves to support arch ribs both in the top heading and also where the bench is being excavated by spanning over this length until posts can be inserted beneath. Newer types of supports are discussed below with more modern tunnel procedures, in which the trend is away from two stages of support toward a single support system, part installed early and gradually strengthened in increments for conversion to the final complete support system.

Environmental control. In all but the shortest tunnels, control of the environment is essential to provide safe working conditions. Ventilation is vital, both to provide fresh air and to remove explosive gases such as methane and noxious gases, including blast fumes. While the problem is reduced by using diesel engines with exhaust scrubbers and by selecting only low-fume explosives for underground use, long tunnels involve a major ventilating plant that employs a forced draft through lightweight pipes up to three feet in diameter and with booster fans at intervals. In smaller tunnels, the fans are frequently reversible, exhausting fumes immediately after blasting, then reversing to supply fresh air to the heading where the work is now concentrated.

High-level noise generated at the heading by drilling equipment and throughout the tunnel by high-velocity air in the vent lines frequently requires the use of earplugs with sign language for communication. In the future, equipment operators may work in sealed cabs, but communication is an unsolved problem. Electronic equipment in tunnels is prohibited, since stray currents may activate blasting circuits. Thunderstorms may also produce stray currents and require special precautions.

Control of dust and heat

Dust is controlled by water sprays, wet drilling, and the use of respirator masks. Since prolonged exposure to dust from rocks containing a high percentage of silica may cause a respiratory ailment known as silicosis, severe conditions require special precautions, such as a vacuum-exhaust hood for each drill.

While excess heat is more common in deep tunnels, it occasionally occurs in fairly shallow tunnels. In 1953, workers in the 6.4-mile Telecote Tunnel near Santa Barbara, Calif., were transported immersed in water-filled mine cars through the hot area (117° F [47° C]). In 1970 a complete refrigeration plant was required to progress through a huge inflow of hot water at 150° F (66° C) in the 7-mile Graton Tunnel, driven under the Andes to drain a copper mine in Peru.

MODERN SOFT-GROUND TUNNELING

Settlement damage and lost ground. Soft-ground tunnels most commonly are used for urban services (subways, sewers, and other utilities) for which the need for quick

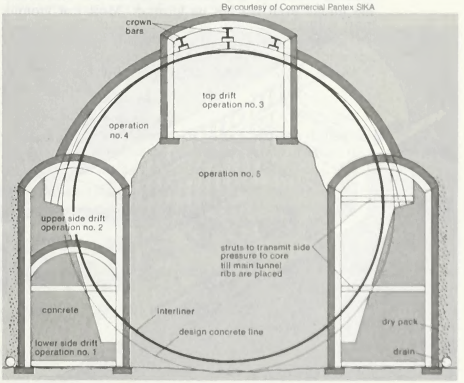


Figure 57: Multiple-drift method of excavation.

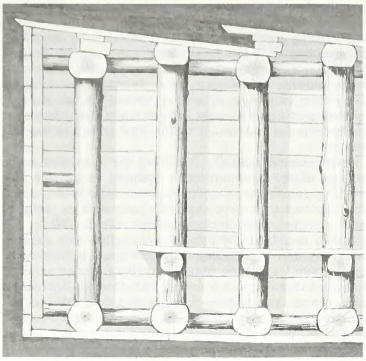


Figure 58: Heading advance by forepoling.

access by passengers or maintenance staff favours a shallow depth. In many cities this means that the tunnels are above bedrock, making tunneling easier but requiring continuous support. The tunnel structure in such cases is generally designed to support the entire load of the ground above it, in part because the ground arch in soil deteriorates with time and in part as an allowance for load changes resulting from future construction of buildings or tunnels. Soft-ground tunnels are typically circular in shape because of this shape's inherently greater strength and ability to readjust to future load changes. In locations within street rights-of-way, the dominant concern in urban tunneling is the need to avoid intolerable settlement damage to adjoining buildings. While this is rarely a problem in the case of modern skyscrapers, which usually have foundations extending to rock and deep basements often extending below the tunnel, it can be a decisive consideration in the presence of moderate-height buildings, whose foundations are usually shallow. In this case the tunnel engineer must choose between underpinning or employing a tunneling method that is sufficiently foolproof that it will prevent settlement damage.

Urban tunnels

Surface settlement results from lost ground—*i.e.*, ground that moves into the tunnel in excess of the tunnel's actual volume. All soft-ground tunneling methods result in a certain amount of lost ground. Some is inevitable, such as the slow lateral squeeze of plastic clay that occurs ahead of the tunnel face as new stresses from doming at the heading cause the clay to move toward the face before the tunnel even reaches its location. Most lost ground, however, results from improper construction methods and careless workmanship. Hence the following emphasizes reasonably conservative tunneling methods, which offer the best chance for holding lost ground to an acceptable level of approximately 1 percent.

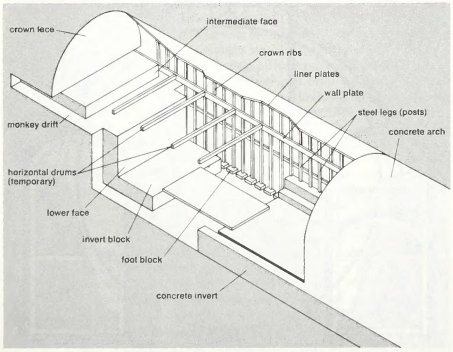


Figure 59: Soft-ground support by ribs and liner plates.

Hand-mined tunnels. The ancient practice of hand mining is still economical for some conditions (shorter and smaller tunnels) and may illustrate particular techniques better than its mechanized counterpart. Examples are forepoling and breasting techniques as developed for the hazardous case of running (unstable) ground. Figure 58 shows the essentials of the process: heading advanced under a roof of forepole planks that are driven ahead at the crown (and at the sides in severe cases) plus continuous planking or breasting at the heading. With careful work the method permits advance with very little lost ground. The top breastboard may be removed, a small advance excavated, this breastboard replaced, and progress continued by working down one board at a time. While solid wall forepoling is nearly a lost art, an adaptation of it is termed spiling. In spiling the forepoles are intermittent with gaps between. Crown spiling is still resorted to for passing bad ground; in this case spiles may consist of rails driven ahead, or even steel bars set in holes drilled into crushed rock.

In ground providing a reasonable stand-up time, a modern support system uses steel liner-plate sections placed against the soil and bolted into a solid sheeted complete circle and, in larger tunnels, strengthened inside by circular steel ribs. Individual liner plates are light in weight and are easily erected by hand. By employing small drifts (horizontal passageways), braced to a central core, liner-plate technique has been successful in larger tunnels—Figure 59 shows 1940 practice on the 20-foot tunnels of the Chicago subway. The top heading is carried ahead, preceded slightly by a "monkey drift" in which the wall plate is set and serves as a footing for the arch ribs, also to span over as the wall plate is underpinned by erecting posts in small notches at each side of the lower bench. As the ribs and liner plate provide only a light support, they are stiffened by installation of a concrete lining about one day behind the mining. While liner-plate tunnels are more economical than shield tunnels, the risks of lost ground are somewhat greater and require not only very careful workmanship but also thorough soil-mechanics investigation in advance, pioneered in Chicago by Karl V. Terzaghi (founder of modern soil mechanics).

Shield tunnels. The risk of lost ground can also be reduced by using a shield with individual pockets from which workers can mine ahead; these can quickly be closed to stop a run-in (Figure 55). In extremely soft ground the shield may be simply shoved ahead with all its pockets closed, completely displacing the soil ahead of it; or it may be shoved with some of the pockets open, through which the soft soil extrudes like a sausage, cut into chunks for removal by a belt conveyor. The first of these methods was used on the Lincoln Tunnel in Hudson River silt.

Internal support in shields

Support erected inside the tail of the shield consists of large segments, so heavy that they require a power erector arm for positioning while being bolted together. Because of its high resistance to corrosion, cast iron has been the most commonly used material for segments, thus eliminating the need for a secondary lining of concrete. Today, lighter segments are employed. In 1968, for example, the San Francisco subway used welded steel-plate segments, protected outside by a bituminous coating and galvanized inside. British engineers have developed precast concrete segments that are proving popular in Europe.

An inherent problem with the shield method is the existence of a 2- to 5-inch (5- to 13-centimetre) ring-shaped void left outside the segments as the result of the thickness of the skin plate and the clearance needed for segment erection. Movement of soil into this void could result in up to 5 percent lost ground, an amount intolerable in urban work. Lost ground is held to reasonable levels by promptly blowing small-sized gravel into the void, then injecting cement grout (sand-cement-water mixture).

Water control. A soft-ground tunnel below the water table involves a constant risk of a run-in—*i.e.*, soil and

Dropping the water table

water flowing into the tunnel, which often results in complete loss of the heading. One solution is to lower the water table below the tunnel bottom before construction begins. This can be accomplished by pumping from deep wells ahead and from well points within the tunnel. While this benefits the tunneling, dropping the water table increases the loading on deeper soil layers. If these are relatively compressible, the result can be a major settlement of adjacent buildings on shallow foundations, an extreme example being a 15- to 20-foot subsidence in Mexico City due to overpumping.

When soil conditions make it undesirable to drop the water table, compressed air inside the tunnel may offset the outside water pressure. In larger tunnels, air pressure is generally set to balance the water pressure in the lower part of the tunnel, with the result that it then exceeds the smaller water pressure at the crown (upper part). Since air tends to escape through the upper part of the tunnel, constant inspection and repair of leaks with straw and mud are required. Otherwise, a blowout could occur, depressurizing the tunnel and possibly losing the heading as soil enters. Compressed air greatly increases operating costs, partly because a large compressor plant is needed, with standby equipment to insure against loss of pressure and partly because of the slow movement of workers and muck trains through the air locks. The dominant factor, however, is the huge reduction in productive time and lengthy decompression time required for people working under air to prevent the crippling disease known as the bends (or caisson disease), also encountered by divers. Regulations stiffen as pressure increases up to the usual maximum of 45 pounds per square inch (3 atmospheres) where daily time is limited to one hour working and six hours for decompression. This, plus higher hazard pay, makes tunneling under high air pressure very costly. In consequence, many tunneling operations attempt to lower the operating air pressure, either by partially dropping the water table or, especially in Europe, by strengthening the ground through the injection of solidifying chemical grouts. French and British grouting-specialist companies have developed a number of highly engineered chemical grouts, and these are achieving considerable success in advance cementing of weak soil.

Tunnel blowout

Soft-ground moles. Since their first success in 1954, moles (mining machines) have been rapidly adopted worldwide. Close copies of the Oahe moles were used for similar large-diameter tunnels in clay shale at Gardiner Dam in Canada and at Mangla Dam in Pakistan during the mid-1960s, and subsequent moles have succeeded at many other locations involving tunneling through soft rocks. Of the several hundred moles built, most have been designed for the more easily excavated soil tunnel and are now beginning to divide into four broad types (all are similar in that they excavate the earth with drag teeth and discharge the muck onto a belt conveyor, and most operate inside a shield).

Four types of moles

The open-face-wheel type is probably the most common. In Figure 60 a scale model is used to illustrate the operation of the wheel within a shield and the lining of precast concrete segments that are conveyed forward for erection within the tail of the shield. In the wheel in Figure 60 the cutter arm rotates in one direction; in a variant model it oscillates back and forth in a windshield-wiper action that is most suitable in wet, sticky ground. While suitable for firm ground, the open-face mole has sometimes been buried by running or loose ground.

The closed-faced-wheel mole partly offsets this problem, since it can be kept pressed against the face while taking in muck through slots. Since the cutters are changed from the face, changing must be done in firm ground. This kind of mole performed well, beginning in the late 1960s, on the San Francisco subway project in soft to medium clay with some sand layers, averaging 30 feet per day. In this project, mole operation made it cheaper and safer to drive two single-track tunnels than one large double-track tunnel. When adjacent buildings had deep foundations, a partial lowering of the water table permitted operations under low pressure, which succeeded in limiting surface settlement to about one inch. In areas of shallow building foundations, dewatering was not permitted; air pressure was then doubled to 28 pounds per square inch, and settlements were slightly smaller.

A third type is the pressure-on-face mole. Here, only the face is pressurized, and the tunnel proper operates in free air—thus avoiding the high costs of labour under pressure. In 1969 a first major attempt used air pressure on the face of a mole operating in sands and silts for the Paris Metro. A 1970 attempt in volcanic clays of Mexico City used a clay-water mixture as a pressurized slurry (liquid mixture); the technique was novel in that the slurry muck was removed by pipeline, a procedure simultaneously also used in Japan with a 23-foot-diameter pressure-on-face mole. The concept has been further developed in England, where an experimental mole of this type was first constructed in 1971.

The digger-shield type of machine is essentially a hydraulic-powered digger arm excavating ahead of a shield, whose protection can be extended forward by hydraulically operated poling plates, acting as retractable spiles. In 1967–70 in the 26-foot-diameter Saugus-Castaic Tunnel near Los Angeles, a mole of this type produced daily progress in clayey sandstone averaging 113 feet per day and 202 feet maximum, completing five miles of tunnel one-half year ahead of schedule. In 1968 an independently developed device of similar design also worked well in compacted silt for a 12-foot-diameter sewer tunnel in Seattle.

Pipe jacking. For small tunnels in a five- to eight-foot size range, small moles of the open-face-wheel type have been effectively combined with an older technique known as pipe jacking, in which a final lining of precast concrete pipe is jacked forward in sections. Figure 61 illustrates the system as used in 1969 on two miles of sewer in Chicago clay with jacking runs up to 1,400 feet between shafts. A laser-aligned wheel mole cut a bore slightly larger than the lining pipe. Friction was reduced by bentonite lubricant added outside through holes drilled from the surface, which were later used for grouting any voids outside the pipe lining. The original pipe-jacking technique was developed particularly for crossing under railroads and highways as a means of avoiding traffic interruption from the alternate of construction in open trench. Since the Chicago project showed a potential for progress of a few hundred feet per day, the technique has become attractive for small tunnels.

MODERN ROCK TUNNELING

Nature of the rock mass. It is important to distinguish between the high strength of a block of solid or intact rock

By courtesy of Calweld, Division of Smith International Incorporated, Santa Fe Springs, California

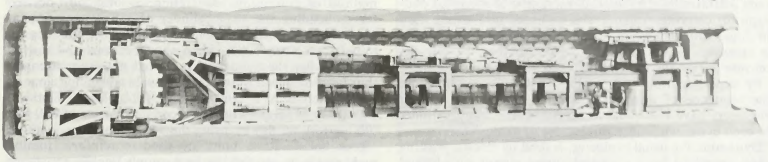


Figure 60: Scale model of a tunneling system (20 feet, 2 inches [6.15 metres] in diameter) for excavation in firm ground. The mechanism includes a wheel-type boring machine, electrohydraulic trailing power unit, 360° revolving drum-type segmented lining erector, segmented lining handling conveyor, and muck-removal conveyor.

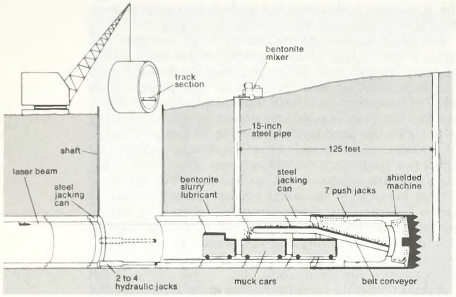


Figure 61: Pipe jacking within a tunnel (see text).
From *Engineering News Record* (1969)

and the much lower strength of the rock mass consisting of strong rock blocks separated by much weaker joints and other rock defects. While the nature of intact rock is significant in quarrying, drilling, and cutting by moles, tunneling and other areas of rock engineering are concerned with the properties of the rock mass. These properties are controlled by the spacing and nature of the defects, including joints (generally fractures caused by tension and sometimes filled with weaker material), faults (shear fractures frequently filled with claylike material called gouge), shear zones (crushed from shear displacement), altered zones (in which heat or chemical action have largely destroyed the original bond cementing the rock crystals), bedding planes, and weak seams (in shale, often altered to clay). Since these geologic details (or hazards) usually can only be generalized in advance predictions, rock-tunneling methods require flexibility for handling conditions as they are encountered. Any of these defects can convert the rock to the more hazardous soft-ground case.

Geostress

Also important is the geostress—*i.e.,* the state of stress existing in situ prior to tunneling. Though conditions are fairly simple in soil, geostress in rock has a wide range because it is influenced by the stresses remaining from past geologic events: mountain building, crustal movements, or load subsequently removed (melting of glacial ice or erosion of former sediment cover). Evaluation of the geostress effects and the rock mass properties are primary objectives of the relatively new field of rock mechanics and are dealt with below with underground chambers since their significance increases with opening size. This section therefore emphasizes the usual rock tunnel, in the size range of 15 to 25 feet.

Conventional blasting. Blasting is carried on in a cycle of drilling, loading, blasting, ventilating fumes, and removing muck. Since only one of these five operations can be conducted at a time in the confined space at the heading, concentrated efforts to improve each have resulted in raising the rate of advance to a range of 40–60 feet per day, or probably near the limit for such a cyclic system. Drilling, which consumes a major part of the time cycle, has been intensely mechanized in the United States. High-speed drills with renewable bits of hard tungsten carbide are positioned by power-operated jib booms located at each platform level of the drilling jumbo (a mounted platform for carrying drills). Truck-mounted jumbos are used in larger tunnels. When rail-mounted, the drilling jumbo is arranged to straddle the mucker so that drilling can resume during the last phase of the mucking operation.

By experimenting with various drill-hole patterns and the sequence of firing explosives in the holes, Swedish engineers have been able to blast a nearly clean cylinder in each cycle, while minimizing use of explosives.

Dynamite, the usual explosive, is fired by electric blasting caps, energized from a separate firing circuit with locked switches. Cartridges are generally loaded individually and seated with a wooden tamping rod; Swedish efforts to expedite loading often employ a pneumatic cartridge loader.

Reducing explosive loading time

American efforts toward reduced loading time have tended to replace dynamite with a free-running blasting agent, such as a mixture of ammonium nitrate and fuel oil (called AN-FO), which in granular form (prills) can be blown into the drill hole by compressed air. While AN-FO-type agents are cheaper, their lower power increases the quantity required, and their fumes usually increase ventilating requirements. For wet holes, the prills must be changed to a slurry requiring special processing and pumping equipment.

Rock support. Most common loading on the support of a tunnel in hard rock is due to the weight of loosened rock below the ground arch, where designers rely particularly on experience with Alpine tunnels as evaluated by two Austrians, Karl V. Terzaghi, the founder of soil mechanics, and Josef Stini, a pioneer in engineering geology. The support load is greatly increased by factors weakening the rock mass, particularly blasting damage. Furthermore, if a delay in placing support allows the zone of rock loosening to propagate upward (*i.e.,* rock falls from the tunnel roof), the rock-mass strength is reduced, and the ground arch is raised. Obviously, the loosened rock load can be greatly altered by a change in joint inclination (orientation of rock fractures) or by the presence of one or more of the rock defects previously mentioned. Less frequent but more severe is the case of high geostress, which in hard, brittle rock may result in dangerous rock bursts (explosive spalling off from the tunnel side) or in a more plastic rock mass may exhibit a slow squeezing into the tunnel. In extreme cases, squeezing ground has been handled by allowing the rock to yield while keeping the process under control, then remining and resetting initial support several times, plus deferring concrete lining until the ground arch becomes stabilized.

Steel rib sets

For many years steel rib sets were the usual first-stage support for rock tunnels, with close spacing of the wood blocking against the rock being important to reduce bending stress in the rib. Advantages are increased flexibility in changing rib spacing plus the ability to handle squeezing ground by resetting the ribs after remining. A disadvantage is that in many cases the system yields excessively, thus inviting weakening of the rock mass. Finally, the rib system serves only as a first-stage or temporary support, requiring a second-stage encasement in a concrete lining for corrosion protection.

Concrete linings. Concrete linings aid fluid flow by providing a smooth surface and insure against rock fragment falling on vehicles using the tunnel. While shallow tunnels are often lined by dropping concrete down holes drilled from the surface, the greater depth of most rock tunnels requires concreting entirely within the tunnel. Operations in such congested space involve special equipment, including agitator cars for transport, pumps or compressed-air devices for placing the concrete, and telescoping arch forms that can be collapsed to move forward inside forms remaining in place. The invert is generally concreted first, followed by the arch where forms must be left in place from 14 to 18 hours for the concrete to gain necessary strength. Voids at the crown are minimized by keeping the discharge pipe buried in fresh concrete. The final operation consists of contact grouting, in which a sand-cement grout is injected to fill any voids and to establish full contact between lining and ground. The method usually produces progress in the range of 40 to 120 feet per day. In the 1960s there was a trend toward an advancing-slope method of continuous concreting, as originally devised for embedding the steel cylinder of a hydropower penstock. In this procedure, several hundred feet of forms are initially set, then collapsed in short sections and moved forward after the concrete has gained necessary strength, thus keeping ahead of the continuously advancing slope of fresh concrete. As a 1968 example, Libby Dam's Flathead Tunnel in Montana attained a concreting rate of 300 feet (90 metres) per day by using the advancing slope method.

Rock bolts. Rock bolts are used to reinforce jointed rock much as reinforcing bars supply tensile resistance in reinforced concrete. After early trials about 1920, they were developed in the 1940s for strengthening laminated roof strata in mines (see INDUSTRIES, EXTRACTION AND

PROCESSING). For public works their use has increased rapidly since 1955, as confidence has developed from two independent pioneering applications, both in the early 1950s. One was the successful change from steel rib sets to cheaper rock bolts on major portions of the 85 miles of tunnels forming New York City's Delaware River Aqueduct. The other was the success of such bolts as the sole rock support in large underground powerhouse chambers of Australia's Snowy Mountains project. Since about 1960, rock bolts have had major success in providing the sole support for large tunnels and rock chambers with spans up to 100 feet. Bolts are commonly sized from 0.75 to 1.5 inches and function to create a compression across rock fissures, both to prevent the joints opening and to create resistance to sliding along the joints. For this they are placed promptly after blasting, anchored at the end, tensioned, and then grouted to resist corrosion and to prevent anchor creep. Rock tendons (prestressed cables or bundled rods, providing higher capacity than rock bolts) up to 250 feet long and prestressed to several hundred tons each have succeeded in stabilizing many sliding rock masses in rock chambers, dam abutments, and high rock slopes. A noted example is their use in reinforcing the abutments of Vaiont Dam in Italy. In 1963 this project experienced disaster when a giant landslide filled the reservoir, causing a huge wave to overtop the dam, with large loss of life. Remarkably, the 875-foot-high arch dam survived this huge overloading; the rock tendons are believed to have supplied a major strengthening.

Pioneering uses of rock bolts

Rock tendons

Shotcrete. Shotcrete is small-aggregate concrete conveyed through a hose and shot from an air gun onto a backup surface on which it is built up in thin layers. Though sand mixes had been so applied for many years, new equipment in the late 1940s made it possible to improve the product by including coarse aggregate up to one inch; strengths of 6,000 to 10,000 pounds per square inch (400 to 700 kilograms per square centimetre) became common. Following initial success as rock-tunnel support in 1951–55 on the Maggia Hydro Project in Switzerland, the technique was further developed in Austria and Sweden. The remarkable ability of a thin shotcrete layer (one to three inches) to bond to and knit fissured rock into a strong arch and to stop raveling of loose pieces soon led to shotcrete largely superseding steel rib support in many European rock tunnels. By 1962 the practice had spread to South America. From this experience plus limited trial at the Hecla Mine in Idaho, the first major use of coarse-aggregate shotcrete for tunnel support in North America developed in 1967 on the Vancouver Railroad Tunnel, with a cross section 20 by 29 feet high and a length of two miles. Here an initial two- to four-inch coat proved so successful in stabilizing hard, blocky shale and in preventing raveling in friable (crumbly) conglomerate and sandstone that the shotcrete was thickened to six inches in the arch and four inches on the walls to form the permanent support, saving about 75 percent of the cost of the original steel ribs and concrete lining.

Applying shotcrete

A key to shotcreting's success is its prompt application before loosening starts to reduce the strength of the rock mass. In Swedish practice this is accomplished by applying immediately after blasting and, while mucking is in progress, utilizing the "Swedish robot" shown in Figure 62, which allows the operator to remain under the protection of the previously supported roof. On the Vancouver tunnel, shotcrete was applied from a platform extending forward from the jumbo while the mucking machine operated below. By taking advantage of several unique properties of shotcrete (flexibility, high bending strength, and ability to increase thickness by successive layers), Swedish practice has developed shotcreting into a single-support system that is strengthened progressively as needed for conversion into the final support.

Preserving rock strength. In rock tunnels, the requirements for support can be significantly decreased to the extent that the construction method can preserve the inherent strength of the rock mass. The opinion has been often expressed that a high percentage of support in rock tunnels in the United States (perhaps over half) has been needed to stabilize rock damaged by blasting rather than

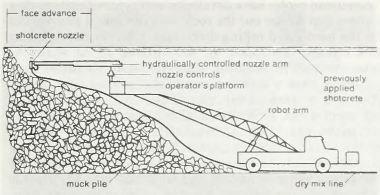


Figure 62: "Swedish robot" used for shotcreting.
From *Civil Engineering* (January 1970), A.S.C.E.

because of an inherently low strength of the rock. As a remedy, two techniques are currently available. First is the Swedish development of sound-wall blasting (to preserve rock strength), treated below under rock chambers, since its importance increases with size of the opening. The second is the American development of rock moles that cut a smooth surface in the tunnel (Figure 63), thus minimizing rock damage and support needs—here limited to rock bolts connected by steel straps for this sandstone tunnel. In stronger rocks (as the 1970 Chicago sewers in dolomite) mole excavation not only largely eliminated need for support but also produced a surface of adequate smoothness for sewer flow, which permitted a major saving by omitting the concrete lining. Since their initial success in clay shale, the use of rock moles has expanded rapidly and has achieved significant success in medium-strength rock such as sandstone, siltstone, limestone, dolomite, rhyolite, and schist. The advance rate has ranged up to 300 to 400 feet per day and has often outpaced other operations in the tunneling system. While experimental moles were used successfully to cut hard rock such as granite and quartzite, such devices were not economical, because cutter life was short, and frequent cutter replacement was costly. This was likely to change, however, as mole manufacturers sought to extend the range of application. Improvement in cutters and progress in reducing the time lost from equipment breakages were producing consistent improvements.

Rock moles

By courtesy of the U.S. Bureau of Reclamation

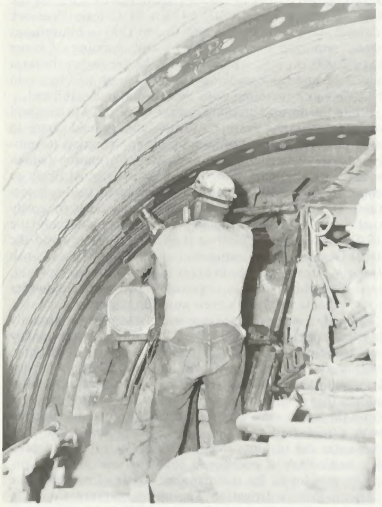


Figure 63: Workman tightening the nut on a rock bolt that holds a steel strap inside a mole-excavated tunnel. Laser beam target at left is used to guide steering of mole.

American moles have developed two types of cutters: disk cutters that wedge out the rock between initial grooves cut by the hard-faced rolling disks, and roller-bit cutters using bits initially developed for fast drilling of oil wells. As later entrants in the field, European manufacturers have generally tried a different approach—milling-type cutters that mill or plane away part of the rock, then shear off undercut areas. Attention is also focusing on broadening the moles' capabilities to function as the primary machine of the whole tunneling system. Thus, future moles are expected not only to cut rock but also to explore ahead for dangerous ground; handle and treat bad ground; provide a capability for prompt erection of support, rock bolting, or shotcreting; change cutters from the rear in loose ground; and produce rock fragments of a size appropriate to capability of the muck removal system. As these problems are solved, the continuous-tunneling system by mole is expected largely to replace the cyclic drilling and blasting system.

Continuous-tunneling system

Water inflows. Exploring ahead of the path of a tunnel is particularly necessary for location of possible high water inflows and permitting their pretreatment by drainage or grouting. When high-pressure flows occur unexpectedly, they result in long stoppages. When huge flows are encountered, one approach is to drive parallel tunnels, advancing them alternately so that one relieves pressure in front of the other. This was done in 1898 in work on the Simplon Tunnel and in 1969 on the Graton Tunnel in Peru, where flow reached 60,000 gallons (230,000 litres) per minute. Another technique is to depressurize ahead by drain holes (or small drainage drifts on each side), an extreme example being the 1968 Japanese handling of extraordinarily difficult water and rock conditions on the Rokko Railroad Tunnel, using approximately three-quarters of a mile of drainage drifts and five miles of drain holes in a one-quarter-mile length of the main tunnel.

Heavy ground. The miner's term for very weak or high geostress ground that causes repeated failures and replacement of support is heavy ground. Ingenuity, patience, and large increases of time and funds are invariably required to deal with it. Special techniques have generally been evolved on the job, as indicated by a few of the numerous examples. On the 7.2-mile Mont Blanc Vehicular Tunnel of 32-foot size under the Alps in 1959–63, a pilot bore ahead helped greatly to reduce rock bursts by relieving the high geostress. The 5-mile, 14-foot El Colegio Penstock Tunnel in Colombia was completed in 1965 in bituminous shale, requiring the replacement and resetting of more than 2,000 rib sets, which buckled as the invert (bottom supports) and sides gradually squeezed in up to 3 feet, and by deferring concreting until the ground arch stabilized.

While the ground arch eventually stabilized in these and numerous similar examples, knowledge is inadequate to establish the point between desirable deformation (to mobilize ground strength) and excessive deformation (which reduces its strength), and improvement is most likely to come from carefully planned and observed field-test sections at prototype scale, but these have been so costly that very few have actually been executed, notably the 1940 test sections in clay on the Chicago subway and the 1950 Garrison Dam test tunnel in the clay-shale of North Dakota. Such prototype field testing has resulted, however, in substantial savings in eventual tunnel cost. For harder rock, reliable results are even more fragmentary.

Desirable and excessive deformation

Unlined tunnels. Numerous modest-size conventionally blasted tunnels have been left unlined if human occupancy was to be rare and the rock was generally good. Initially, only weak zones are lined, and marginal areas are left for later maintenance. Most common is the case of a water tunnel that is built oversized to offset the friction increase from the rough sides and, if a penstock tunnel, is equipped with a rock trap to catch loose rock pieces before they can enter the turbines. Most of these have been successful, particularly if operations could be scheduled for periodic shutdowns for maintenance repair of rockfalls; the Laramie-Poudre Irrigation Tunnel in northern Colorado experienced only two significant rockfalls in 60 years, each easily repaired during a nonirrigation period. In contrast, a progressive rockfall on the 14-mile Kemano penstock tunnel in Canada resulted in shutting down the whole town of Kitimat in British Columbia, and vacationing workers for nine months in 1961 since there were no other electric sources to operate the smelter. Thus, the choice of an unlined tunnel involves a compromise between initial saving and deferred maintenance plus evaluation of the consequences of a tunnel shutdown.

Underground excavations and structures

ROCK CHAMBERS

While chambers in 1971 were being excavated in rock to fulfill a wide variety of functions, the main stimulus to their development had come from hydroelectric-power-plant requirements. Though the basic concept originated in the United States, where the world's first underground hydroplants were built in enlarged tunnels at Snoqualme Falls near Seattle, Wash., in 1898 and at Fairfax Falls, Vt., in 1904, Swedish engineers developed the idea into excavating large chambers to accommodate hydraulic machinery. After an initial trial in 1910–14 at the Porjus Plant north of the Arctic Circle, many underground power plants were subsequently built by the Swedish State Power Board. Swedish success soon popularized the idea through Europe and over the world, particularly to Australia, Scotland, Canada, Mexico, and Japan, where several hundred underground hydroplants have been built since 1950. Sweden, having a long experience with explosives and rock work, with generally favourable strong rock, and with energetic research and development, has even been able to lower the costs for underground work to approximate those for surface construction of such facilities as power plants, warehouses, pumping plants, oil-storage tanks, and water-treatment plants. With costs in the United States being 5 to 10 times greater underground, new construction of underground chambers was not significantly resumed there until 1958, when the Haas underground hydroplant was built in California and the Norad underground air force command centre in Colorado. By 1970 the United States had begun to adopt the Swedish concept and had completed three more hydroplants with several more under construction or being planned.

Swedish work on rock chambers

Favourably located, an underground hydroplant can have several advantages over a surface plant, including lower costs, because certain plant elements are built more simply underground: less risk from avalanches, earthquakes, and bombing; cheaper year-round construction and operation (in cold climates); and preservation of a scenic environment—a dominant factor in Scotland's tourist area and now receiving recognition worldwide. A typical layout involves a complex assembly of tunnels, chambers, and shafts. Figure 64 shows the world's largest underground powerhouse, Churchill Falls in the Labrador wilderness of Canada, with a capacity of five million kilowatts, under construction since 1967 at a total project cost of about $1 billion. By building a dam of modest height well above the falls and by locating the powerhouse at 1,000 feet depth with a one-mile tunnel (the tailrace tunnel) to discharge water from the turbines below downstream rapids, the designers have been able to develop a head (water height) of 1,060 feet while at the same time preserving the scenic 250-foot-high waterfall, expected to be a major tourist attraction once several hundred miles of wilderness-road improvement permits public access. Openings here are of impressive size: machine hall (powerhouse proper), 81-foot span by 154 feet high by 972 feet long; surge chamber, 60 feet by 148 feet high by 763 feet; and two tailrace tunnels, 45 by 60 feet high.

Advantages of underground location for a hydroplant

Large rock chambers are economical only when the rock can essentially support itself through a durable ground arch with the addition of only a modest amount of artificial support. Otherwise, major structural support for a large opening in weak rock is very costly. The Norad project, for example, included an intersecting grid of chambers in granite 45 by 60 feet high, supported by rock bolts except in one local area. Here, one of the chamber intersections coincided with the intersection of two curving shear zones of fractured rock—a happening which added $3.5 million extra cost for a perforated concrete dome 100 feet in

diameter to secure this local area. In some Italian and Portuguese underground powerhouses, weak-rock areas have necessitated comparable costly lining. While significant rock defects are more manageable in the usual 10- to 20-foot rock tunnel, the problem so increases with increasing size of opening that the presence of extensive weak rock can easily place a large-chamber project outside the range of economic practicality. Hence, geologic conditions are very carefully investigated for rock-chamber projects, using many borings plus exploratory drifts to locate rock defects, with a three-dimensional geologic model to aid in visualizing conditions. A chamber location is selected that offers the least risk of support problems. This objective was largely attained in the granite gneiss at Churchill Falls, where the location and chamber configuration were changed several times to avoid rock defects. Rock-chamber projects, furthermore, rely heavily on the relatively new field of rock mechanics to evaluate the engineering properties of the rock mass, in which exploratory drifts are particularly important in affording access for in-place field testing.

Rock-mechanics investigation. The young field of rock mechanics was beginning, early in the 1970s, to develop a rational basis of design for projects in rock; much is already developed for projects in soil by the older field of soil mechanics. Initially, the discipline had been stimulated by such complex projects as arch dams and underground chambers and then increasingly with similar problems with tunnels, rock slopes, and building foundations. In treating the rock mass with its defects as an engineering material, the science of rock mechanics utilizes numerous techniques such as theoretical analysis, laboratory testing, field testing on-site, and instrumentation to monitor performance during construction and operation. Since rock mechanics is a discipline in itself, only the most common field tests are briefly outlined below to give some concept of its role in design, particularly for a rock-chamber project.

Techniques of rock mechanics

Geostress, which can be a significant factor in choice of chamber orientation, shape, and support design, is usually determined in exploratory drifts. Two methods are common, although each is still in the development stage. One is an "overcoring" method (developed in Sweden and South Africa) used for ranges up to about 100 feet out from the drift and employing a cylindrical instrument known as a borehole deformeter. A small hole is drilled into the rock and the deformeter inserted. Diameter changes of the borehole are measured and recorded by the deformeter as the geostress is relieved by overcoring (cutting a circular core around the small hole) with a six-inch bit. Measurements at several depths in at least three borings at different orientations furnish the data needed for computing the existing geostress. When measurement is desired only at the surface of the drift, the so-called French flat-jack method is preferred. In this, a slot is cut at the surface, and its closure is measured as the geostress is relieved by the slot. Next, a flat hydraulic jack is inserted in the rock. The jack pressure necessary to restore closure of the slot (to the condition before its cutting) is considered to equal the original geostress. As these methods require a long drift or shaft for access to the area of measurement, development is under way (particularly in the United States) to extend the range of depth to a few thousand feet. Such will aid in comparing geostress at alternate sites and hopefully avoid locations with high geostress, which has proved very troublesome in several past chamber projects.

Shear strength of a joint, fault, or other rock defect is a controlling factor in appraising strength of the rock mass in terms of its resistance to sliding along the defect. Although partly determinable in the laboratory, it is best investigated in the field by a direct shear test at the work site. While this test has long been used for soil and soft rock, its adaptation to hard rock is due largely to work performed in Portugal. Shear strength is important in all problems of sliding; at Morrow Point Dam, in Colorado, for example, a large rock wedge between two faults started to move into the underground powerhouse and was stabilized by large tendons anchored back in a drainage tunnel plus strut action provided by the concrete structure that supported the generator machinery.

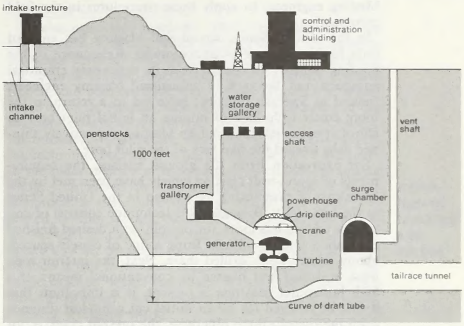


Figure 64: Churchill Falls underground powerhouse.
By courtesy of Churchill Falls Corp.

The modulus of deformation (that is, the stiffness of the rock) is significant in problems involving movement under stress and in sharing of load between rock and structure, as in a tunnel lining, embedded steel penstock, or foundation of a dam or heavy building. The simplest field test is the plate-jacking method, in which the rock in a test drift is loaded by hydraulic jacks acting on a plate two to three feet in diameter. Larger areas can be tested either by radially loading the internal surface of a test tunnel or by pressurizing a membrane-lined chamber.

Analysis methods in rock mechanics have helped in appraising stress conditions around openings—as at Churchill Falls—to identify and then correct zones of tension and stress concentration. Related work with rock block models is contributing to understanding the failure mechanism of the rock mass, notable work being under way in Austria, Yugoslavia, and the United States.

Chamber excavation and support. Excavation for rock chambers generally starts with a horizontal tunnel at the top of the area to be excavated and progresses down in steps. Rock is excavated by drilling and blasting, carried on simultaneously in several headings. This procedure may give way, however, as moles gain in their ability to cut hard rock economically and as a rock saw or other device is developed for squaring up the circular surface normally cut by the mole. High geostress can be a real problem (causing inward movement of the chamber walls) unless handled by a careful sequence of partial excavations designed to relieve it gradually.

Excavation methods for rock chambers

Many of the earlier underground hydroplants were roofed with a concrete arch, often designed for a major load, as in some Italian projects in weak rocks or where blast damage was considerable, as at a few projects in Scotland. Since about 1960, however, most have relied solely on rock bolts for support (sometimes supplemented with shotcrete). That such a light support has been widely successful can be attributed to careful investigation resulting in locations with strong rock, employment of techniques to relieve high geostress, and controlled blasting to preserve rock strength.

Sound-wall blasting. Sound-wall blasting is a technique, primarily developed in Sweden, that preserves the finished rock surfaces in sound condition by careful design of the blasting charges to fit the rock conditions. In underground work, Swedish practice has often produced remarkable results almost like rock sculpturing in which the excellent shaping and preservation of the rock surfaces often permit omitting concrete lining at savings greater than the extra cost of the engineered blasting. While Swedish success is due partly to the generally strong rock in that country, it is due even more to energetic research and development programs to develop (1) theoretical methods for blasting design plus field blast tests to determine pertinent rock properties, (2) special explosives for different rock conditions, and (3) institutes for the training of specialized

Swedish work on sound-wall blasting

blasting engineers to apply these procedures in the field construction.

In the United States, sound-wall blasting has enjoyed only indifferent success underground. Reluctance of the blasting industry to change from its customary empirical approach and the lack of specialized blasting engineers trained in Swedish practices have led to a return to the more costly technique of mining an initial pilot bore to afford stress relief, followed by blasting successively thinner slabs toward the free face of the pilot bore.

For excavation from the ground surface, the requirements of sound-wall blasting largely have been met by the technique of presplitting, developed in the United States in the late 1950s. Basically, this technique consists of creating a continuous crack (or presplit) at a desired finished excavation line by initially firing a line of closely spaced, lightly loaded holes drilled there. Next, the interior rock mass is drilled and blasted by conventional means. If a high horizontal geostress is present, it is important that it first be relieved (as by an initial cut a modest distance from the presplit line); otherwise, the presplit crack is not likely to occur in the direction desired. Stockton Dam, in Missouri, illustrates the benefit of presplitting. Here, vertical faces in dolomite up to 110 feet were successfully presplit and promptly rock-bolted; this permitted a major reduction in thickness of the concrete facing, resulting in a net saving of about $2.5 million.

SHAFTS

The mining industry has been the primary constructor of shafts, because at many locations these are essential for access to ore, for ventilation, and for material transport. Depths of several thousand feet are common. In public-works projects, such as sewer tunnels, shafts are usually only a few hundred feet deep and because of their high cost are avoided in the design stage wherever practical. Shallower shafts find many uses, however, for penstocks and access to underground hydroplants, for dropping aqueduct tunnels beneath rivers, for missile silos, and for oil and liquefied-gas storage. Being essentially vertical tunnels, shafts involve the same problems of different types of ground and water conditions but on an aggravated scale, since vertical transport makes the operation slower, more costly, and even more congested than with horizontal tunneling. Except when there is a high horizontal geostress in rock, the loading on a shaft support is generally less than for a tunnel. Inflowing water, however, is far more dangerous during construction and generally intolerable during operation. Hence, most shafts are concrete-lined and waterproofed, and the lining installation usually follows only a short distance behind excavation. The shape is usually circular, although, before current mechanized excavation methods, mining shafts were frequently rectangular. Shafts may be sunk from the surface (or drilled in smaller sizes), or, if an existing tunnel provides access, they may be raised from below.

Shaft sinking and drilling. Mining downward, generally from the surface, although occasionally from an underground chamber, is called shaft sinking. In soil, shallow shafts are frequently supported with interlocking steel sheetpiling held by ring beams (circular rib sets); or a concrete caisson may be built on the surface and sunk by excavating inside as weight is added by extending its walls. More recently, large-diameter shallow shafts have been constructed by the "slurry trench method," in which a circular trench is excavated while filled with a heavy liquid (usually bentonite slurry), which supports its walls until it is finally displaced by filling the trench with concrete. For greater depth in soil, another method involves freezing a ring of soil around the shaft. In this method, a ring of closely spaced freezing holes is drilled outside the shaft. A refrigerated brine is circulated in double-wall pipes in the holes to freeze the soil before starting the shaft excavation. It is then kept frozen until the shaft is completed and lined with concrete. This freezing method was developed in Germany and The Netherlands, where it was used successfully to sink shafts through nearly 2,000 feet of alluvial soil to reach coal beds in the underlying rock. It has also been applied under similar conditions in Britain, Poland, and Belgium. Occasionally, the freezing technique has been used in soft rock to solidify a deep aquifer (layer of water-bearing rock). Because of the long time required for drilling the freezing holes and for freezing the ground (18 to 24 months for some deep shafts), the freezing method has not been popular on public-works projects except as a last resort, although it has been used in New York City for shallow shafts through soil to gain access for deep-water tunnels.

Freezing method of shaft sinking

More efficient methods for sinking deep shafts in rock were developed in South African gold-mining operations, in which shafts 5,000 to 8,000 feet deep are common and are generally 20 to 30 feet in diameter. South African procedure has produced progress of about 30 feet per day by utilizing a sinking stage of multiple platforms, which permits concurrent excavation and concrete lining. Excavation is by drilling and blasting with muck loaded into large buckets, with larger shafts operating four buckets alternately in hoisting wells extending through the platforms. Grouting is carried a few hundred feet ahead to seal out water. Best progress is achieved when the rock is pregrouted from two or three holes drilled from the surface before the shaft is started. Since the shallower shafts on public-works projects cannot justify the investment in the large plant needed to operate a sinking stage, their progress in rock is much slower—in the range of 5 to 10 feet per day.

Occasionally, shafts have been sunk through soil by drilling methods. The technique was first used in British practice in 1930 and was subsequently further refined in The Netherlands and Germany. The procedure involves first advancing a pilot hole, then reaming in several stages of enlargement to final diameter, while the walls of the hole are supported by a heavy liquid (called drilling mud), with circulation of the mud serving to remove the cuttings. Then a double-wall steel casing is sunk by displacing the drilling mud, followed by injecting concrete outside the casing and within the annular space between its double walls. One use of this technique was in the 25-foot-diameter Statemine shaft in The Netherlands, 1,500 feet deep through soil that required about three and one-half years before completion in 1959. For the 1962 construction of some 200 missile shafts in Wyoming in soft rock (clay shale and friable sandstone), a giant auger proved effective for sinking these 65-foot-deep, 15-foot-diameter shafts, generally at the rate of two to three days per shaft. Perhaps the largest drilled shaft is one in the Soviet Union: 2,674 feet deep, which was enlarged in four stages of reaming to a final diameter of 28.7 feet, progressing at a reported rate of 15 feet per day.

More dramatic has been the adaptation in the United States of oil-well-drilling methods in a technique called big-hole drilling, used for constructing small shafts in the diameter range of three to six feet. Big-hole drilling was developed for deep emplacement in underground testing of nuclear devices, with more than 150 such big holes drilled in the 1960s up to 5,000 feet deep in Nevada in rocks ranging from soft tuff to granite. In big-hole drilling the hole is made in one pass only with an array of roller-bit cutters that are pressed against the rock by the weight of an assembly of lead-filled drill collars, sometimes totaling 300,000 pounds. The drill rig must be huge in size to handle such loads. The greatest impediment controlling progress has been the removal of drill cuttings, where an air lift is showing promise.

Big-hole drilling

Shaft raising. Handling cuttings is simplified when the shaft can be raised from an existing tunnel, since the cuttings then merely fall to the tunnel, where they are easily loaded into mine cars or trucks. This advantage has long been recognized in mining; where once an initial shaft has been sunk to provide access to and an opportunity for horizontal tunnels, most subsequent shafts are then raised from these tunnels, often by upward mining with men working from a cage hung from a cable through a small pilot hole drilled downward from above. In 1957 this procedure was improved by Swedish development of the raise climber, whose working cage climbs a rail fastened to the shaft wall and extends backward into the horizontal access tunnel into which the cage is retracted during

a blast. Simultaneously in the 1950s Germans began experimenting with several mechanized reamers, including a motor-cutter unit pulled upward by a cable in a previously down-drilled pilot hole. A more significant step toward mechanized shaft raising occurred in 1962 when American mole manufacturers developed a device called a raise borer, in which the cutting head is rotated and pulled upward by a drill shaft in a down-drilled pilot hole, with the power unit being located at top of the pilot hole. The capacity of this type of borer (or upward reamer) generally ranges from 3–8-foot diameters in lifts up to 1,000 feet with progress ranging up to 300 feet per day. Furthermore, available cutters when operating on raise borers can cut through rock often almost twice as hard as rock moles can deal with. For larger shafts, bigger-diameter reamers may be operated in an inverted position to ream downward, with the cuttings sluiced to the access tunnel below. A 12-foot-diameter, 1,600-foot-deep vent shaft was completed by this method in 1969 at the White Pine Copper Mine in Michigan. Starting from a 10-inch pilot hole, it was enlarged in three downreaming passes.

The introduction of a workable raise borer in the 1960s represented a breakthrough in shaft construction, cutting construction time to one-third and cost to less than one-half that for an upward-mined shaft. At the beginning of the 1970s, the procedure was being widely adopted for shaft raising, and some projects had been specifically designed to take advantage of this more efficient method. At a Northfield Mountain (Massachusetts) underground hydroplant (completed in 1971), the previously common large surge chamber was replaced by a series of horizontal tunnels at three levels, connected by vertical shafts. This layout permitted significant economy by the use of jumbos already available from other tunnels of the project and the use of a raise borer for starting the shafts. If very large shafts are involved, the raise borer is particularly useful in simplifying the so-called glory-hole method (Figure 65), in which the main shaft is sunk by blasting; the muck is then dumped in the central glory hole, previously constructed by a raise borer. The example is based on the construction of a 133-foot-diameter surge shaft above the Angeles penstock tunnel near Los Angeles. The glory-hole technique was also used in 1944 in constructing a series of 20 underground fuel-oil chambers in Hawaii, working from access tunnels driven initially at both top and bottom of the chambers and later used to house oil and vent piping. The advent of the raise borer should now make this and similar construction more economically attractive. Recently, some deep sewer projects have been redesigned to utilize the raise borer for shaft connections.

Glory-hole method

Adapted from A.D. Parker, *Planning and Estimating Underground Construction* (copyright 1970); used with permission of McGraw-Hill Book Co.

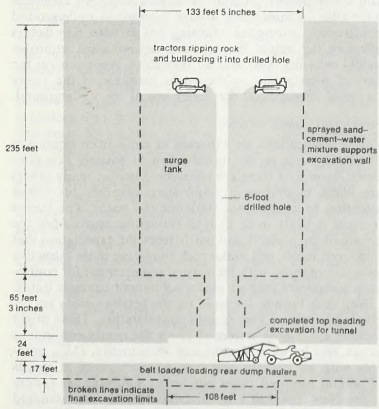


Figure 65: Shaft excavation by glory-hole method at the Angeles penstock tunnel near Los Angeles.

IMMERSED-TUBE TUNNELS

Development of method. The immersed-tube, or sunken-tube, method, used principally for underwater crossings, involves prefabricating long tube sections, floating them to the site, sinking each in a previously dredged trench, and then covering with backfill (Figure 66). While more correctly classified as a subaqueous adaptation of the dry-land cut-and-cover procedure often used for subways, the immersed-tube method warrants inclusion as a tunneling technique because it is becoming a preferred alternate to the older method of constructing a subaqueous tunnel under compressed air with a Greathead shield. A major advantage is that, once the new section has been connected, interior work is conducted in free air, thus avoiding the high cost and major risk of operating a large shield under high air pressure. Furthermore, the immersed-tube method is usable in water deeper than is possible with the shield method, which essentially is restricted to less than 100 feet of water by the maximum air pressure at which workers can safely work.

Advantages of immersed-tube method

From *Civil Engineering* (December 1966)

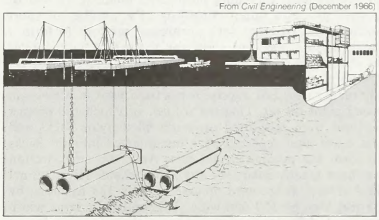


Figure 66: Immersed-tube tunnel-building procedure (see text).

The procedure was first developed by an American engineer, W.J. Wilgus, for the construction (1906–10) of the Detroit River twin-tube railroad tunnel between Detroit, Mich., and Windsor, Ont., where it was successfully used for the 2,665-foot river-crossing portion. A structural assembly of steel tubes was prefabricated in 262-foot-long sections with both ends temporarily bulkheaded or closed. Each section was then towed out and sunk in 60 to 80 feet of water, onto a grillage of I-beams in sand at the bottom of a trench previously dredged in the river-bottom clay. After being connected to the previous section by locking pins driven by a diver, the section was weighted down by surrounding it with concrete. Next, after removal of the temporary bulkheads at the just-completed connection, the newly placed section was pumped out, permitting completion of an interior concrete lining in free air. With subsequent refinements these basic principles still form the basis of the immersed-tube method.

After use on a four-tube New York City subway crossing under the Harlem River in 1912–14, the method was tried for a vehicular tunnel in the 1925–28 construction of the 3,545-foot-long, 37-foot-diameter Posey tunnel at Oakland in California. Because these and other experiences have indicated that the problems encountered in building large vehicular tunnels could be better handled by the immersed-tube method, it has been preferred for subaqueous vehicular tunnels since about 1940. While shield tunneling continued in a transition period (1940–50), subsequently nearly all of the world's large vehicular tunnels have been constructed by the immersed-tube method, including such notable examples as the Bankhead tunnel at Mobile, Ala.; two Chesapeake Bay tunnels; the Fraser River tunnel at Vancouver, B.C.; the Maas River tunnel in The Netherlands; Denmark's Limfjord tunnel; Sweden's Tingstad tunnel; and the Hong Kong Cross Harbor tunnel.

Modern practice. The world's longest and deepest application to date is the twin-tube subway crossing of San Francisco Bay, constructed between 1966 and 1971 with a length of 3.6 miles in a maximum water depth of 135 feet. The 330-foot-long, 48-foot-wide sections were constructed of steel plate and launched by shipbuilding procedures. Each section also had temporary end bulkheads and upper pockets for gravel ballast placed during sinking. After

The San Francisco subway tunnel

placement of the interior concrete lining at a fitting-out dock, each section was towed to the site and sunk in a trench previously dredged in the mud in the bottom of the bay. With diver guidance, the initial connection was accomplished by hydraulic-jack-powered couplers, similar to those that automatically join railroad cars. By relieving the water pressure within the short compartment between bulkheads at the new joint, the water pressure acting on the forward end of the new section provided a huge force that pushed it into intimate contact with the previously laid tube, compressing the rubber gaskets to provide a watertight seal. Following this, the temporary bulkheads were removed on each side of the new joint and interior concrete placed across the connection.

Most applications of the immersed-tube procedure outside the United States have been by a Danish engineer-constructor firm, Christiani and Nielsen, starting in 1938 with a three-tube highway crossing of the Maas River in Rotterdam. While following American technique in essence, European engineers have developed a number of innovations, including prestressed concrete in lieu of a steel structure (often consisting of a number of short sections tied together with prestressed tendons to form a single section 300 feet in length); the use of butyl rubber as the waterproofing membrane; and initial support on temporary piles while a sand fill is jetted beneath. An alternate to the last approach has been used in a Swedish experiment on the Tingstad tunnel, in which the precast sections were supported on water-filled nylon sacks and the water later replaced by grout injected into the sacks to form the permanent support. Also, the cross section has been greatly enlarged—the 1969 Schelde River tunnel in Antwerp, Belg., used precast sections 328 feet long by 33 feet high by 157 feet wide. This unusually large width accommodates two highway tubes of three lanes each, one two-track railroad tube, and one bicycle tube. Particularly unusual was a 1963 use of the immersed-tube technique in subway construction in Rotterdam. Trenches were dug or, in some cases, made out of abandoned canals and filled with water. The tube sections were then floated into position. This technique had been first tried in 1952 for a land approach to the immersed-tube Elizabeth tunnel in Norfolk, Va.; in low-elevation ground with the water table near the surface, it permits a considerable saving in bracing of the trench because keeping the trench filled eliminates the need for resisting external water pressure.

European innovations

Thus, the immersed-tube method has become a frequent choice for subaqueous crossings, although some locations pose problems of interference with intensive navigation traffic or the possibility of displacement by severe storms (one tube section of the Chesapeake Bay tunnel was moved out of its trench by a severe storm during construction). The method is being actively considered for many of the world's most difficult underwater crossings, including the long-discussed English Channel Project.

Future trends in underground construction

ENVIRONMENTAL AND ECONOMIC FACTORS

Improvement of surface environment. Unexpectedly rapid increases in urbanization throughout the world, especially since World War II, have brought many problems, including congestion, air pollution, loss of scarce surface area for vehicular ways, and major traffic disruption during their construction. Some cities relying principally on auto transport have even found that nearly two-thirds of their central land area is devoted to vehicular service (freeways, streets, and parking facilities), leaving only one-third of the surface space for productive or recreational use. During the past decade there has been a growing awareness that this situation could be alleviated by underground placement of a large number of facilities that do not need to be on the surface, such as rapid transit, parking, utilities, sewage and water-treatment plants, fluid storage, warehouses, and light manufacturing. The overriding deterrent, however, has been the greater cost underground—except in Sweden, where energetic research has reduced underground costs to nearly equal the surface alternates. Hence planners have rarely dared to propose underground construction except where the surface alternate was widely recognized as intolerable. Underground construction in urban areas has, thus, generally been limited to situations without a viable surface alternate; as a result, additional increases in surface construction have further aggravated the problem. At the same time, the low volume of underground construction has provided insufficient incentive for the development of innovative technology.

A different approach for the United States was crystallized from a 1966–68 study by the National Academy of Sciences and the National Academy of Engineering, which proposed cost reduction from government-stimulated technological research plus broader evaluation of social impacts. This would often show the underground alternate as the better investment for society. A reduction of at least one-third in cost and one-half in construction time over the next two decades was foreseen, and it was proposed that social and environmental costs be included in estimates as well as construction costs. In 1970 an international meeting of some 20 countries was held in Washington, D.C., under the Organisation for Economic Co-operation and Development (an assembly of NATO countries), to share views and develop recommendations on government policy in this area. The conference recommended that energetic stimulation of underground construction be adopted as national policy in each of the 20 countries represented and in effect visualized the underground as a largely undeveloped natural resource. This resource, it was pointed out, could be used to expand urban areas downward to help preserve the upper environment—for example, by tunnels for transport and interbasin water transfer, for recovery of minerals increasingly needed by the economy, and in developing currently unreachable resources under ocean areas adjoining the continents. Such international consensus suggests that this is indeed a powerful concept ready for acceptance.

New approach to underground construction

Scope of the tunneling market. While informed people foresee a great increase in underground construction, numerical estimates are crude at best, particularly since statistics have not been accumulated in the past for underground construction as a separate item either in the public-works or the mining sectors. The 1970 conference mentioned above included a survey suggesting an average annual volume in its 20 member countries of about $1 billion in public works for the 1960–69 decade ($3 billion including mining). Estimates made at that time of a doubling of volume over the next decade assumed the continuation of the current rate of technological improvement and recognized that the increase would be far greater if stimulated by government support in an energetic research and development program to reduce cost. All estimates were alike in forecasting a huge increase in underground construction during the following two decades. Key factors affecting the actual increase are technological improvements reducing costs and an increasing awareness on the part of society and public-works planners of the many potential applications for better use of the underground.

POTENTIAL APPLICATIONS

Future applications are expected to range from expansion of existing uses to the introduction of entirely new concepts. Several of these are considered below; many others are likely to emerge as innovative planners turn their attention to utilizing the underground space. The largest increase is likely to be in rock tunneling: partly from the nature of the projects and partly from the expectation that improved moles will make rock tunneling more attractive than soil tunnels, with their usual requirement for continuous temporary support plus a permanent concrete lining.

Deep rock tunnels for rapid transit between cities are beginning to receive very serious consideration. These might include a 425-mile system to cover the nearly continuous urban area between Boston and Washington, D.C., probably with an entirely new type of conveyance at speeds of several hundred miles per hour. A forerunner system is the New Tōkaidō Line in Japan, which uses standard railroad equipment at about 150 miles per hour. Highway tunnels are beginning to increase in number as well. Ur-

Intercity transit

ban highway tunnels conceivably may offer a convenient opportunity to reduce pollution by treating the exhaust air that has already been collected by the ventilating system essential for longer vehicular tunnels.

There is increasing recognition that many more interbasin water transfers will be needed, involving systems of tunnels and canals. Notable projects include the California Aqueduct, which transfers water from the northern mountains some 450 miles to the semiarid Los Angeles area; the Orange-Fish Project in South Africa, which includes a 50-mile tunnel; and studies for possible transfer of surplus Canadian water into the southwestern United States. Drainage can also be a problem, as in the old lake bed area occupied by Mexico City, where current expansion of the drainage system involves some 60 miles of tunnel.

Urban tunnels

Shallower tunnels for subways are bound to increase beyond those expansions undertaken in recent years in many cities, including San Francisco, Washington, D.C., Boston, Chicago, New York, London, Paris, Budapest, Munich, and Mexico City. Multiple use is likely to receive further consideration as communication agencies begin to show interest in adding space within the structures for the several types of utilities. Some merchants visualize mechanized movement of pedestrians between stores. One notable example is Montreal's extensive assembly of underground shopping malls, which interconnect most new downtown buildings as well as provide access to the subway and commuter railroads—a project that has relieved the streets from pedestrian traffic, particularly during severe weather. Another example involves utilization of space excavated above subway stations for parking facilities, as on the Toronto subway and more recently on the Paris Métro, where the space above one of the stations in the Champs-Élysées area provides seven levels of parking.

Subaqueous crossings are becoming more ambitious. The world's longest railroad tunnel, for example, currently under way in Japan, is the 34-mile Seikan undersea rock tunnel between the islands of Honshu and Hokkaido; the 14.4-mile pilot tunnel, completed in 1983 after 19 years of work, was utilized as a proving ground for several new types of moles. Of comparable scope is the more publicized projected English Channel tunnel for a rail connection between France and England, using special cars for auto transport. Studies have concentrated on two alternatives: twin mole-excavated tunnels in chalk plus a service tunnel or an immersed-tube structure providing comparable space. The immersed-tube procedure has also been considered for a number of other difficult crossings—*e.g.*, from Denmark to Sweden and from Sicily to Italy. Immersed tubes are likely to become more attractive with improvement in methods for trench dredging in deeper water and for grading the trench bottom to support the tube structure. The Japanese are experimenting with an underwater bulldozer, robot-manned and television-monitored. One innovative proposal for supplying additional water to southern California visualizes the immersed-tube method to construct a large pipeline for some 500 miles under the shallower ocean along the continental shelf. Subaqueous tunneling also is likely to be involved as procedures are developed for utilizing the vast continental-shelf areas of the world; concepts are already being studied for tunnels to service oil wells and for extensive undersea mining such as has been pioneered in Britain and eastern Canada.

Fluid storage

Both Norway and Sweden have reduced the direct costs of fluid storage by storing petroleum products in underground chambers, thus eliminating the maintenance cost for frequent repainting of steel tanks in a surface facility. Locating these chambers below the permanent water table (and below any existing wells) ensures that seepage will be toward the chambers rather than outward; thus, the oil is prevented from leaking out of the chamber, and the lining may be omitted. Further economies may result from orienting the chambers vertically to take advantage of the raise borer and glory-hole techniques, previously mentioned. There are a number of underground installations for the storage of highly compressed gas cooled to a liquid state; these may increase once improved types of lining have been developed. Although the method involves only limited tunneling for access, the United States Atomic Energy Commission has developed an ingenious method for disposal of nuclear waste by injecting it into fissured rock within a cement grout so that hardening of the grout reconverts the nuclear minerals into a stable rocklike state. Other disposal methods involve more tunneling, such as within salt, which has particularly good ability for shielding against radiation.

A good example of an imaginative concept is Chicago's Underflow Tunnel and Reservoir Plan, which is intended to alleviate both pollution and flooding. Like most older cities, Chicago has a combined sewer system that carries both storm runoff and sanitary sewage during wet weather but only sanitary sewage during dry weather. The city's huge growth has so overtaxed older portions of the system that severe storms cause flooding in low areas. While sewage treatment has essentially eliminated sewage pollution of Lake Michigan, making Chicago virtually the only major city on the Great Lakes continuing wide recreational usage of its lake beaches, the treatment plants generally are sized to handle only the dry-weather flow. Thus, overflow during major storms is discharged into streams draining away from the lake as a mixture of sanitary sewage diluted by storm water. Conventional solutions adopted in the past, such as adding a second pipe system to collect only the storm water, discharging it into the streams, or adding plant capacity to treat all combined flow during severe storms, have proved tremendously expensive. An early version of the plan is illustrated in Figure 67; a novel aspect of the plan is temporary storage

By courtesy of the Metropolitan Sanitary District of Greater Chicago

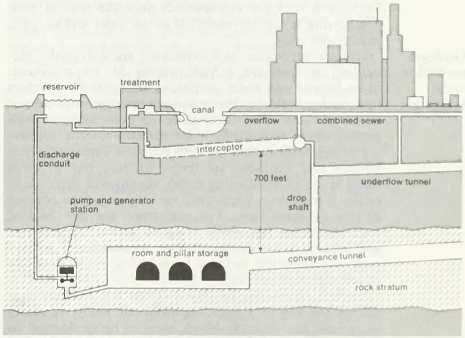


Figure 67: Potential water control plan for Chicago.

of excess water in large underground caverns, which after each storm could be pumped out for gradual treatment by the existing sewage plants. Inclusion of the surface reservoir shown in Figure 67 makes practical the use of the diluted sewage in a pumped storage hydroplant; in this type of facility the fluid is pumped up during off-peak-electric-power night periods, when steam power is cheaply available, and then allowed to flow back to generate peak power when demand exceeds economic capacity of the steam plants. A second multiple use is the opportunity to reduce present surface quarrying for crushed stone aggregate by using the dolomitic limestone mined from the deep tunnels and caverns.

The use of rock chambers for underground hydroplants seems certain to increase in most countries, particularly those in which until recently surface plants have been favoured because of their apparently lower cost. Scotland has been one of the first countries to recognize that extra construction cost can often be warranted to preserve the scenic environment, also recognized by choice of an underground location for recent U.S. pump-storage plants—Northfield Mt. in Massachusetts and Raccoon Mt. in Tennessee, plus others being planned. Sweden's use of the underground for plants treating sewage and water, for warehouses, and for light manufacturing is likely to find further application. The relatively small annual temper-

Temperature underground

ature range in the underground has made it a desirable environment for facilities requiring close atmospheric control. In the vicinity of Kansas City in Missouri, mined-out space in underground limestone quarries is being used effectively for laboratory space, for dehumidified storage of corrosion-sensitive equipment, and for refrigerated food storage, an application also favoured in Sweden.

Similar environment factors plus the probability of less disturbance during earthquakes have made the underground desirable for a number of scientific installations, including atomic accelerators, earthquake research, nuclear research, and space telescopes. Since earthquake risk is a big factor in locating nuclear power plants, the merits of an underground location are attracting interest.

IMPROVED TECHNOLOGY

Worldwide efforts are under way to accelerate improvements in the technology of underground construction and are likely to be stimulated as a result of the 1970 OECD International Conference recommending improvement as government policy. The endeavour involves specialists such as geologists, soil- and rock-mechanics engineers, publicworks designers, mining engineers, contractors, equipment and materials manufacturers, planners, and also lawyers, who aid in the search for more equitable contractual methods to share the risks of unknown geology and resulting extra costs. Many improvements and their early applications have been previously discussed; others are briefly mentioned here, including several that have not yet moved from the research stage to the pilot, or trial, stage. Projects in rock are emphasized, since the field of rock engineering is less developed than its older counterpart, soils engineering.

Geologic prediction and evaluation

Geologic prediction and evaluation are universally recognized as deserving a high priority for improvement. Since ground and water conditions are controlling factors in choosing both the design and construction method underground and seem destined to be even more so with greater use of moles, efforts are directed toward improving boring information (as with borehole cameras), faster borings (the Japanese are trying to bore one to three miles ahead of a tunneling mole), geophysical methods to estimate rock-mass properties, and techniques to observe pattern of water flows. For evaluation, the new field of rock mechanics is concentrating on measuring geostress and rock-mass properties, failure mechanics of jointed rock, and analytical methods for applying results to design of underground openings.

For rock excavation, improved cutters are generally considered the key for expanding economic ability of moles to include harder rock. Much effort is being devoted to improving current mechanical cutters, including technical advances based upon space metallurgy, geometry of cutter shape and arrangement, mechanics of cutting action, and research in presoftening rock. Concurrently, there is

New rock-cutting methods

an intensive search for entirely new rock-cutting methods (some nearing a pilot application), including high-pressure water jets, Russian water cannon (operated at high pressures), electron beam, and flame jet (often combined with abrasive powder). Other methods under research involve lasers and ultrasonics. Most of these have high power requirements and might increase ventilating needs from an already overtaxed system. Though some of these novel methods will eventually reach the stage of economic practicality, it is not possible to predict at present which ones will eventually succeed. Also needed is a means for testing rock in terms of mole drillability plus correlation with mole performance in different rocks, where promising work is under way at several locations.

A decided change in current materials-handling systems seems inevitable to keep up with fast-moving moles by matching the mole's rate of excavation and fragmentation sizing of the muck produced. Schemes now under study include long belt conveyors, high-speed rail with completely new types of equipment, and both hydraulic and pneumatic pipelines. Useful experience is being accumulated with pipeline transport of ore slurries, of coal, and even of such bulky material as canned goods.

For ground support, rock-mechanics engineers are working toward replacing past empirical methods with a more rational basis of design. One key factor is likely to be the tolerable deformation for mobilizing but not destroying the strength of the rock mass. There is wide agreement that progress will best be aided by field-test sections at prototype scale in selected ongoing projects. While several newer types of support have been discussed (rock bolts, shotcrete, and precast-concrete elements), developments are under way toward entirely new types, including lighter material plus yield-controllable types as a corollary to above tolerable deformation concept. For projects using concrete lining, major changes seem inevitable to keep pace with fast-moving moles, probably including some entirely new types of concretes. Current efforts include work with precast elements, plus research into stronger and faster set materials which use resins and other polymers in lieu of portland cement.

Underground support

Preservation of ground strength is beginning to win acceptance as vital for the safety of large rock chambers and also often a means of cost saving in tunnels. For preserving strength of the rock mass around tunnels, a mole-cut surface provides a solution. For large chambers, consideration is being given to cutting a peripheral slot with a wire saw of the type used to quarry monument stone. Where chambers are blasted, engineered sound-wall blasting has provided a solution in Sweden.

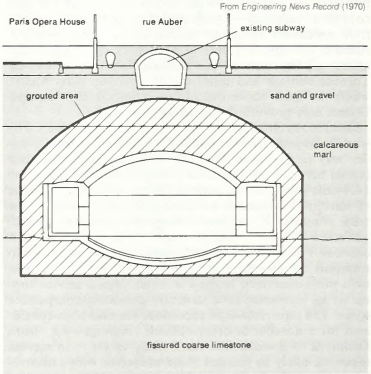


Figure 68: Ground strengthening at Auber Station of the Paris Métro.

Ground strengthening by precementation with chemical grouts is a technique notably developed in France and Britain through extensive research by specialized grouting firms. Figure 68 shows the world's outstanding application at the Auber Station of the Métro Express beneath the Place de L'Opéra traffic centre of Paris—a large chamber 130 feet wide by 60 feet high by 750 feet long in chalky marl below an existing subway, at a depth of 120 feet, about 60 feet below water table. This was completed in 1970 without interrupting surface traffic and without underpinning the many old masonry buildings above (including the historic National Opera Building), a truly courageous undertaking made possible by surrounding the chamber with a pregrouted zone to seal out water and to precement the overlying sand and gravel. Different types of chemical grout were successively injected (totaling about two billion cubic feet), working from crown and side drifts; then the chamber was mined and supported both top and bottom by prestressed arches of concrete elements. Similar procedure was also successful at the Étoile Station adjacent to the Arc de Triomphe. While this technique of ground strengthening by grout solidification requires highly skilled specialists, it is an instructive example of how a new technology is likely to make economically possible future projects previously considered beyond engineering ability.

(K.S.L.)

BIBLIOGRAPHY

Roads and highways. M.G. LAY, *Ways of the World: A History of the World's Roads and of the Vehicles That Used Them* (1992), provides an overview, while his *Handbook of Road Technology,* 2nd ed., 2 vol. (1990), includes technical information on planning, pavements, traffic, and transport. Additional historical works are HERMANN SCHREIBER, *The History of Roads: From Amber Route to Motorway* (1961; originally published in German, 1959), on the ancient roads of Asia, Europe, and Egypt and the Incan roads in South America; R.J. FORBES, "Roads to *c.* 1900," in CHARLES SINGER *et al.* (eds.), *A History of Technology,* vol. 4 (1958), pp. 520–547; W.J. READER, *Macadam: The McAdam Family and the Turnpike Roads, 1798–1861* (1980), treating the roads in 19th-century Great Britain; and UNITED STATES, FEDERAL HIGHWAY ADMINISTRATION, *America's Highways, 1776–1976* (1977), chronicling federal aid to transportation in the United States.

Aspects of road engineering and operation are detailed in *Highway Capacity Manual,* 2nd ed., rev. (1993), published by the Transportation Research Board (United States); AMERICAN ASSOCIATION OF STATE HIGHWAY AND TRANSPORTATION OFFICIALS, *AASHTO Guide for Design of Pavement Structures,* vol. 1, new ed. (1993) and vol. 2 (1986); and publications of the ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT, ROAD RESEARCH PROGRAMME: *Proceedings of the International Symposium on the Effects of Speed Limits on Traffic Accidents and Fuel Consumption* (1982), *Traffic Capacity of Major Routes* (1983), *Impacts of Heavy Freight Vehicles* (1983), and *Road Surface Characteristics: Their Interaction and Their Optimisation* (1984). PETER KINCAID, *The Rule of the Road* (1986), presents international traffic regulations . (M.G.L.)

Bridges. DAVID P. BILLINGTON, *The Tower and the Bridge: The New Art of Structural Engineering* (1983), traces the history of structures since the Industrial Revolution and emphasizes a series of individual engineers who were structural artists. DAVID J. BROWN, *Bridges* (1993), chronicles bridge building from its beginnings to future projects, with discussions of the world's most important bridges, some of which were failures. ERIC DELONY (ed.), *Landmark American Bridges* (1993), contains elegantly illustrated presentations of large and small bridges built in the United States from the late 18th century to the post-World War II era, with a time line encapsulating the history of bridges from 1570 to the present. C.M. WOODWARD, *A History of the St. Louis Bridge* (1881), on the Eads Bridge, is still the most complete work ever published on one bridge, a classic work for both the nontechnical reader and the engineer. DAVID MCCULLOUGH, *The Great Bridge* (1972, reissued 1982), narrates the political and human history of the building of the Brooklyn Bridge. DAVID P. BILLINGTON, *Robert Maillart's Bridges* (1979), gives substantial details of the major works of this bridge designer and builder.

FRITZ LEONHARDT, *Bridges: Aesthetics and Design* (1984), provides a substantial general discussion and includes a wide selection of striking photographs of old and new bridges. A finely illustrated nontechnical text by HANS WITTFOHT, *Building Bridges* (1984), centers on both the design and the construction of bridges. M.S. TROITSKY, *Planning and Design of Bridges* (1994), focuses on the selection of bridge type and preliminary design. WALTER PODOLNY, JR., and JOHN B. SCALZI, *Construction and Design of Cable-Stayed Bridges,* 2nd ed. (1986), treats the modern bridge form and provides substantial technical detail, although the nontechnical reader will understand much of the text. CHRISTIAN MENN, *Prestressed Concrete Bridges,* trans. and ed. by PAUL GAUVREAU (1990; originally published in German, 1986), a technical work primarily for engineers, provides historical coverage and deals elegantly with the many complex issues in bridge design. LOUIS G. SILANO (ed.), *Bridge Inspection and Rehabilitation* (1993), is a compendium of practical information.

Early works on bridge design and construction still widely available include HENRY GRATTAN TYRRELL, *History of Bridge Engineering* (1911), a fine, nontechnical sweep through history from ancient bridges to early reinforced concrete arches; WILBUR J. WATSON, *Bridge Architecture* (1927), exploring the collaboration between engineers and architects and focusing on early 20th-century designs, and *A Decade of Bridges, 1926–1936* (1937), with a narrower scope; CHARLES S. WHITNEY, *Bridges: A Study in Their Art, Science, and Evolution* (1929, reprinted as *Bridges: Their Art, Science, and Evolution,* 1983), a nontechnical work; and ELIZABETH B. MOCK (ELIZABETH B. KASSLER), *The Architecture of Bridges* (1949), the first major book on bridges to give a modern viewpoint. (D.P.B.)

Airports. NORMAN ASHFORD and PAUL H. WRIGHT, *Airport Engineering,* 3rd ed. (1992), comprehensively sets forth the planning, layout, and design of passenger and freight airports, including heliports and short take-off and landing (STOL) facilities. ROBERT HORONJEFF and FRANCIS X. MCKELVEY, *Planning and Design of Airports,* 4th ed. (1993), is a comprehensive civil engineering text on the planning, layout, and design of airports with strong emphasis on aspects such as aircraft pavements and drainage. INTERNATIONAL CIVIL AVIATION ORGANIZATION, *Aerodromes: International Standards and Recommended Practices* (1990), includes the internationally adopted design and operational standards for all airports engaged in international civil aviation.

CHRISTOPHER R. BLOW, *Airport Terminals* (1991), provides an architectural view of the functioning of airport passenger terminals with extensive coverage of design case studies. WALTER HART, *The Airport Passenger Terminal* (1985, reprinted 1991), describes the functions of passenger terminals and their design requirements. INTERNATIONAL AIR TRANSPORT ASSOCIATION, *Airport Terminals Reference Manual,* 7th ed. (1989), provides design and performance requirements of passenger and freight terminals as set out by the international airlines' trade association.

DENIS PHIPPS, *The Management of Aviation Security* (1991), describes the operational and design requirements of civil airports to conform to national and international regulations. NORMAN ASHFORD, H.P. MARTIN STANTON, and CLIFTON A. MOORE, *Airport Operations* (1984, reissued 1991), extensively discusses many aspects of airport operation and management, including administrative structure, security, safety, environmental impact, performance indices, and passenger and aircraft handling. NORMAN ASHFORD and CLIFTON A. MOORE, *Airport Finance* (1992), discusses the revenue and expenditure patterns of airport authorities, methods of financing, business planning, and project appraisal. RIGAS DOGANIS, *The Airport Business* (1992), examines the status of airport business in the early 1990s, performance indices, commercial opportunities, and privatization of airports. (N.J.A.)

Canals and inland waterways. Coverage of both Old and New World inland waterways is provided by CHARLES HADFIELD, *World Canals: Inland Navigation Past and Present* (1986), heavily illustrated. L.T.C. ROLT, *Navigable Waterways* (1969), provides a descriptive account of inland waterway networks.

J. PHILLIPS, *A General History of Inland Navigation, Foreign and Domestic,* 5th ed. (1805, reprinted as *Phillips' Inland Navigation,* 1970), contains a complete account of the canals built in England up to the end of the 18th century. RUTH DELANY, *Ireland's Royal Canal: 1789–1992* (1992), details the history of the canal from its beginnings through its years of decline and disuse to its late 20th-century restoration. OSBORNE MANCE (HARRY OSBORNE MANCE), *International River and Canal Transport* (1944), chronicles the history of the commissions controlling and regulating operations on the international waterways up to the end of World War II. ROGER CALVERT, *Inland Waterways of Europe,* 2nd ed. (1975), contains a nontechnical account of the principal European inland waterway systems, including interesting diagrams, maps, useful itineraries, and an international bibliography.

Discussions of the growth and importance of canals in the history of the United States are found in RONALD E. SHAW, *Erie Water West: A History of the Erie Canal, 1792–1854* (1966, reprinted 1990), and *Canals for a Nation: The Canal Era in the United States, 1790–1860* (1990); and RUSSELL BOURNE, *Floating West: The Erie and Other American Canals* (1992).

Two well-known international canals are treated in D.A. FARNIE, *East and West of Suez: The Suez Canal in History, 1854–1956* (1969), a comprehensive and detailed study of the history of the Suez Canal and of its impact on the foreign policies of the imperial powers in relation to the Middle East; and DAVID MCCULLOUGH, *The Path Between the Seas: The Creation of the Panama Canal, 1870–1914* (1977), a history of the canal's construction. (E.A.J.D./C.M./Ed.)

Dams. NORMAN SMITH, *A History of Dams* (1971), contains an outstanding detailed record of ancient dams. NICHOLAS J. SCHNITTER, *A History of Dams* (1994), chronicles the construction and use of dams from ancient civilization to the present. Two histories of individual dams are ANNE D. RASSWEILER, *The Generation of Power: The History of Dneprostroi* (1988), providing the political background to the dam's development; and ANDREW J. DUNAR and DENNIS MCBRIDE, *Building Hoover Dam* (1993), an oral history of those involved in the project.

ALFRED R. GOLZÉ (ed.), *Handbook of Dam Engineering* (1977), is a manual of design and construction. Earlier works are J. GUTHRIE BROWN (ed.), *Hydro-Electric Engineering Practice,* 3 vol., 2nd ed. (1964–70), a comprehensive textbook dealing principally with British and European practice; JAMES L. SHERARD *et al., Earth and Earth-Rock Dams* (1963), on the design and construction of foundations and embankments; and CALVIN VICTOR DAVIS and KENNETH E. SORENSON (eds.), *Handbook of Applied Hydraulics,* 3rd ed. (1969, reprinted 1984), a classic work on the basic principles of hydraulic engineering and the design of hydraulic structure. J. LAGINHA SERAFIM and R.W. CLOUGH (eds.), *Arch Dams* (1990), collects papers from an international workshop on this type of dam. (J.G.B./Ed.)

Harbours and sea works. WOLFGANG RUDOLPH, *Harbor and Town: A Maritime Cultural History* (1980; originally published in German, 1980), includes treatments of port anatomy and historical development. Books on engineering aspects include F.M. DU-PLAT-TAYLOR, *The Design, Construction, and Maintenance of Docks, Wharves & Piers,* 3rd ed., rev. and enlarged (1949); JOHN F. BRAHTZ (ed.), *Ocean Engineering: Goals, Environment, Technology* (1968); A.M. MUIR WOOD and C.A. FLEMING, *Coastal Hydraulics,* 2nd ed. (1981); HANS AGERSHOU, HELGE LUNDGREN, and TORBEN SORENSEN, *Planning and Design of Ports and Marine Terminals* (1983); GREGORY P. TSINKER, *Floating Ports: Design and Construction Practices* (1986); PER BRUUN, *Port Engineering,* 4th ed., 2 vol. (1989–90), covering topics such as harbour planning, breakwaters, fishing ports, sediment transport, and geomorphology; KENNETH M. CHILDS, JR. (ed.), *Ports '89* (1989), conference papers; and RICHARD SILVESTER and JOHN R.C. HSU, *Coastal Stabilization: Innovative Concepts* (1993). *The Dock and Harbour Authority* (monthly), is devoted to problems of dock and harbour operation and construction in Great Britain. (J.H.J./Ed.)

Lighthouses. KENNETH SUTTON JONES, *Pharos: The Lighthouse Yesterday, Today, and Tomorrow* (1985), a comprehensive and detailed treatment, includes a discussion of future trends. DAVID STEVENSON, *The World's Lighthouses before 1820* (1959), is a classic work by a member of the famous family of Scottish lighthouse engineers. PATRICK BEAVER, *A History of Lighthouses* (1971), provides historical descriptions of the construction of important lighthouses in various parts of the world, as well as the people involved.

DOUGLAS B. HAGUE and ROSEMARY CHRISTIE, *Lighthouses: Their Architecture, History, and Archaeology* (1975), offers a history of lighthouse building, construction, engineering, and architecture from earliest times throughout the world. ROBERT STEVENSON, *English Lighthouse Tours,* ed. by DAVID STEVENSON (1946), is an edited version of a detailed journal kept during a tour of English lighthouses. HILARY POLAND MEAD, *Trinity House* (1947), discusses the origins and history of the lighthouse authority for England and Wales. GEORGE R. PUTNAM, *Lighthouses and Lightships of the United States,* new and rev. ed. (1933), provides a historical survey of most of the major lighthouses and lightships of the United States through the early 20th century; it may be supplemented by FRANCIS ROSS HOLLAND, JR., *America's Lighthouses: Their Illustrated History Since 1716* (1972, reissued 1988). UNITED STATES, COAST GUARD, *Historically Famous Lighthouses* (1950), and *Coast Guard History,* new ed. (1982), provide brief insights into the work of the U.S. Coast Guard and some of its more famous lighthouses. (I.C.C.)

Environmental works. Broad coverage of water supply, waste disposal, pollution, and other public health and environmental sanitation topics may be found in the following texts: JERRY A. NATHANSON, *Basic Environmental Technology* (1986); GILBERT M. MASTERS, *Introduction to Environmental Engineering and Science* (1991); and JOSEPH A. SALVATO, *Environmental Engineering and Sanitation,* 4th ed. (1992), including administrative aspects.

NORMAN SMITH, *Man and Water* (1975), presents a comprehensive history of water technology, including irrigation, hydropower, and drinking water. MARK J. HAMMER, *Water and Wastewater Technology,* 2nd ed. (1986), an introductory engineering text, covers municipal water processing and distribution, wastewater collection and treatment, and sludge disposal. JAMES C. LAMB III, *Water Quality and Its Control* (1985), provides a comprehensive but nontechnical introduction to water resources, including the basics of municipal water and wastewater systems, aquatic ecology, and water quality controls. Comprehensive, in-depth technical descriptions of the major desalting methods are found in K.S. SPIEGLER and A.D.K. LAIRD, *Principles of Desalination,* 2nd ed., 2 vol. (1980). JAMES M. MONTGOMERY CONSULTING ENGINEERS, INC., *Water Treatment Principles and Design* (1985), is a technical presentation of the physical, chemical, and microbiological principles of water purification, including process control, facilities design, and operational issues. TERENCE J. MCGHEE, *Water Supply and Sewerage,* 6th ed. (1991), presents sanitary engineering principles with emphasis on the design of water and sewage works. WARREN VIESSMAN, JR., and MARK J. HAMMER, *Water Supply and Pollution Control,* 5th ed. (1993), an engineering textbook, emphasizes the application of scientific principles to solving problems related to the movement and treatment of water and sewage. WARREN VIESSMAN, JR., and CLAIRE WELTY, *Water Management: Technology and Institutions* (1985), presents the technical aspects of water resources systems as well as discussions of the related economic, political, and social problems. METCALF & EDDY, INC., *Wastewater Engineering: Treatment, Disposal, Reuse,* 3rd ed. rev. by GEORGE TCHOBANOGLOUS and FRANKLIN L. BURTON (1991), offers a detailed technical presentation of wastewater engineering principles and practices, in a form suitable as a reference for professional engineers and as a textbook for engineering students.

E.S. SAVAS, *The Organization and Efficiency of Solid Waste Collection* (1977), gives a thorough description of how cities organize, manage, and finance solid waste collection services. GEORGE TCHOBANOGLOUS, HILARY THEISEN, and ROLF ELIASSEN, *Solid Wastes: Engineering Principles and Management Issues* (1977), focuses on collection, transport, treatment, recovery, and disposal of residential and commercial solid wastes. C.L. MANTELL (ed.), *Solid Wastes: Origin, Collection, Processing, and Disposal* (1975), offers technical discussion and case studies and includes the topics of agricultural and food processing wastes, animal husbandry wastes, process industry wastes, and mineral and metallurgical wastes. DAVID GORDON WILSON (ed.), *Handbook of Solid Waste Management* (1977), is a comprehensive source of data on the properties, handling, processing, and disposal of municipal solid waste, including information on resource reclamation and energy recovery. AMALENDU BAGCHI, *Design, Construction, and Monitoring of Landfills,* 2nd ed. (1994), provides in-depth coverage of modern landfill design including site selection, leachate and gas control, operation, and economic analysis. LUIS F. DIAZ *et al., Composting and Recycling Municipal Solid Waste* (1993), thoroughly discusses municipal recycling facilities, waste composting operations, and integrated waste-management concepts.

J. WILLIAM HAUN, *Guide to the Management of Hazardous Waste* (1991), offers an introduction to the problems of hazardous waste production, identification, treatment, and disposal, and includes a discussion of laws and social issues. WILLIAM C. BLACKMAN, JR., *Basic Hazardous Waste Management* (1993), is an introduction to hazardous waste management methods and includes coverage of radioactive and infectious biomedical wastes. MICHAEL D. LAGREGA, PHILLIP L. BUCKINGHAM, and JEFFREY C. EVANS, *Hazardous Waste Management* (1994), a professional reference and graduate-level engineering textbook, provides a comprehensive overview of hazardous waste treatment, storage, disposal, site remediation, and management issues. (J.A.N.)

Tunneling and underground excavation. Ancient and modern tunnels in Europe, Japan, and the U.S. are described in the well-illustrated study by PATRICK BEAVER, *A History of Tunnels* (1972). T.M. MEGAW and J.V. BARTLETT, *Tunnels—Planning, Design, Construction,* 2 vol. (1981–82), is a substantial study. GRAHAM WEST, *Innovation and the Rise of the Tunnelling Industry* (1988), focuses on both hard- and soft-rock projects. AGRICOLA, *De Re Metallica,* trans. from Latin by HERBERT CLARK HOOVER and LOU HENRY HOOVER (1912, reprinted 1950), is a classic work on early mining in Europe. FREDERICK WALTER SIMMS, *Practical Tunneling,* 4th ed., rev. and expanded by D. KINNEAR CLARK (1896), on early public works, recounts the difficulties overcome by pioneers. C.A. PEQUIGNOT (ed.), *Tunnels and Tunnelling* (1963), on English practice, includes comprehensive tables comparing tunnels of the world. GÖSTA E. SANDSTRÖM, *Tunnels* (also published as *The History of Tunneling,* 1963), is a historical survey that summarizes Sweden's contributions to underground engineering. K.G. STAGG and O.C. ZIENKIEWICZ (eds.), *Rock Mechanics in Engineering Practice* (1968), an introductory work, contains 12 chapters, each written by a noted authority. HAROLD W. RICHARDSON and ROBERT S. MAYO, *Practical Tunnel Driving* (1941, reissued 1975), a history of U.S. practice to 1940, emphasizes tunnel equipment. ALBERT D. PARKER, *Planning and Estimating Underground Construction* (1970), also on U.S. practice, emphasizes construction engineering and estimating. Microtunneling methods are discussed in DIETRICH STEIN, KLEMENS MÖLLERS, and ROLF BIELECKI, *Microtunnelling: Installation and Renewal of Nonman-size Supply and Sewage Lines by the Trenchless Construction Method* (1989; originally published in German, 1988). Contemporary developments are detailed in the periodicals *ENR* (weekly), an engineering news record; and *Tunnels & Tunnelling* (monthly). (K.S.L./Ed.)

Publishing

Publishing is the activity that involves the selection, preparation, and distribution of written and visual matter. It has grown from small and ancient beginnings into a vast and complex industry responsible for the dissemination of all manner of cultural material; its impact upon civilization is impossible to calculate.

This article treats the history and development of book, newspaper, and magazine publishing in its technical and commercial aspects. The preparation and dissemination of information in these forms is followed from its beginnings in the ancient world to the modern period. For additional information on the preparation of early manuscripts, see WRITING. A more detailed examination of printing technology can be found in PRINTING, TYPOGRAPHY, AND PHOTOENGRAVING. The dissemination of published material via electronic media is treated in INFORMATION PROCESSING AND INFORMATION SYSTEMS. For a discussion of reference-book publishing, see ENCYCLOPAEDIAS AND DICTIONARIES.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* section 735, and the *Index.*

The article is divided into the following sections:

General considerations 415
Book publishing 416
  The origins of books 416
    Books on clay tablets
    The Egyptian papyrus roll
    Chinese books
  Books in classical antiquity 417
    Greek books
    Roman books
  Books in the early Christian Era 418
    The codex
    Vellum and parchment
    Christianity and the book
  The medieval book 419
    The monasteries
    The revival of the secular book trade
    Humanistic and vernacular books
  The age of early printing: 1450–1550 420
    Early printer-publishers
    Printed illustrations
    The book trade
    Controls over printing
  The flourishing book trade: 1550–1800 423
    Advances in Europe and America
    Spread of education and literacy
    Growth of libraries
    Decline of censorship
  Modern publishing: from the 19th century to the present 425
    The 19th century
    The 20th century
    Publishing practice
Newspaper publishing 431
  Origins and early evidences 431
    The Roman Empire
    China
    Medieval Europe
  The first newspapers 432
    Commercial newsletters in continental Europe
    Early newspapers in Britain and America
    Early newspapers in Japan
  Era of the Industrial Revolution 434
    Technological advances
    Foundations of modern journalism
    Growth of the newspaper business
  Era of the popular press 436
    The United States
    Great Britain
  The modern era 438
    Technological developments
    Financial developments
    The role of the press
Magazine publishing 440
  Beginnings in the 17th century 440
  Developments in the 18th century 440
    Great Britain
    Continental Europe
    America
  The 19th century and the start of mass circulation 441
    General periodicals
    Illustrated magazines
    Women's magazines
    Literary and scientific magazines
    Scholarly journals
  The 20th century 444
    The advertising revolution in popular magazines
    Publications outside Europe and the United States
    News and photo magazines
    Digests and pocket magazines
    Specialized magazines
    Scholarly, cultural, and literary magazines
Bibliography 448

General considerations

The history of publishing is characterized by a close interplay of technical innovation and social change, each promoting the other. Publishing as it is known today depends on a series of three major inventions—writing, paper, and printing—and one crucial social development—the spread of literacy. Before the invention of writing, perhaps by the Sumerians in the 4th millennium BC, information could be spread only by word of mouth, with all the accompanying limitations of place and time. Writing was originally regarded not as a means of disseminating information but as a way to fix religious formulations or to secure codes of law, genealogies, and other socially important matters, which had previously been committed to memory. Publishing could begin only after the monopoly of letters, often held by a priestly caste, had been broken, probably in connection with the development of the value of writing in commerce. Scripts of various kinds came to be used throughout most of the ancient world for proclamations, correspondence, transactions, and records; but book production was confined largely to religious centres of learning, as it would be again later in medieval Europe. Only in Hellenistic Greece, in Rome, and in China, where there were essentially nontheocratic societies, does there seem to have been any publishing in the modern sense—*i.e.,* a copying industry supplying a lay readership.

The invention and original function of writing

The invention of printing transformed the possibilities of the written word. Printing seems to have been first invented in China in the 6th century AD in the form of block printing. An earlier version may have been developed at the beginning of the 1st millennium BC, but, if so, it soon fell into disuse. The Chinese invented movable type in the 11th century AD but did not fully exploit it. Other Chinese inventions, including paper (AD 105), were passed on to Europe by the Arabs but not, it seems, printing. The reason may well lie in Arab insistence on hand copying of the Qur'ān (Arabic printing of the Qur'ān does not appear to have been officially sanctioned until 1825). The invention of printing in Europe is usually attributed to Johannes Gutenberg in Germany about 1440–50, although block printing had been carried out from about 1400. Gutenberg's achievement was not a single invention but a whole new craft involving movable metal type, ink, paper, and press. In less than 50 years it had been carried through most of Europe, largely by German printers.

Gutenberg's achievement

Printing in Europe is inseparable from the Renaissance and Reformation. It grew from the climate and needs of the first, and it fought in the battles of the second. It has been at the heart of intellectual development for the past 500 years. Although printing was thought of at first merely as a means of avoiding copying errors, its possibilities for mass-producing written matter soon became evident. In 1498, for instance, 18,000 letters of indulgence were printed at Barcelona. The market for books was still small, but literacy had spread beyond the clergy and had reached the emerging middle classes. The church, the state, universities, reformers, and radicals were all quick to use the press. Not surprisingly, every kind of attempt was made to control and regulate such a "dangerous" new mode of communication. Freedom of the press was pursued and attacked for the next three centuries; but by the end of the 18th century a large measure of freedom had been won in western Europe and North America, and a wide range of printed matter was in circulation. The mechanization of printing in the 19th century and its further development in the 20th, which went hand in hand with increasing literacy and rising standards of education, finally brought the printed word to its powerful position as a means of influencing minds and, hence, societies.

The functions peculiar to the publisher—*i.e.,* selecting, editing, and designing the material; arranging its production and distribution; and bearing the financial risk or the responsibility for the whole operation—often merged in the past with those of the author, the printer, or the bookseller. With increasing specialization, however, publishing became, certainly by the 19th century, an increasingly distinct occupation. Most modern Western publishers purchase printing services in the open market, solicit manuscripts from authors, and distribute their wares to purchasers through shops, mail order, or direct sales.

Published matter falls into two main categories, periodical and nonperiodical; *i.e.,* publications that appear at more or less regular intervals and are members of a series and those that appear on single occasions (except for reissues of essentially the same material).

Of the nonperiodical publications, books constitute by far the largest class; they are also, in one form or another, the oldest of all types of publication and go back to the earliest civilizations. In giving permanence to man's thoughts and records of his achievements, they answer a deep human need. Not every published book is of lasting value; but a society's books, taken as a whole and winnowed out by the passing years, can be said to be its main cultural storehouse. Conquerors or usurpers wishing to destroy a people's heritage have often burned its books, as did Shih Huang-ti in China in 213 BC, the Spaniards in Mexico in 1520, and the Nazis in the 1930s.

There is no wholly satisfactory definition of a book, as the word covers a variety of publications (for example, some publications that appear periodically, such as *The World Almanac and Book of Facts,* may be considered books). For statistical purposes, however, the United Nations Educational, Social and Cultural Organization defines a book as "a non-periodical printed publication of at least 49 pages excluding covers."

Periodical publications may be further divided into two main classes, newspapers and magazines. Though the boundary between them is not sharp—there are magazines devoted to news, and many newspapers have magazine features—their differences of format, tempo, and function are sufficiently marked: the newspaper (daily or weekly) usually has large, loose pages, a high degree of immediacy, and miscellaneous contents; whereas the magazine (weekly, monthly, bimonthly, or quarterly) has smaller pages, is usually fastened together and sometimes bound, and is less urgent in tone and more focused in content. Both sprang up after the invention of printing, but both have shown a phenomenal rate of growth to meet the demand for quick information and regular entertainment. Newspapers have long been the most widely read published matter; the democratizing process of the 19th and 20th centuries would be unthinkable without them. Magazines, close behind newspapers both historically and in terms of readership, rapidly branched out from their learned origins into "periodicals of amusement." Today there is probably not a single interest, frivolous or serious, of man, woman, or child, that is not catered to by a magazine.

There are, of course, many other types of publications besides books, newspapers, and magazines. In many cases the same principles of publishing apply, and it is only the nature of the product and the technicalities of its manufacture that are different; such is the case with electronic books or online magazines. There is also the business of map and atlas publishing. Another field is music publishing, which produces a great variety of material, from complete symphonic scores to sheet music of the latest popular hit. A further range of activities might be grouped under the term "utility publishing"; *i.e.,* the issuing of calendars, diaries, timetables, guide books, and all manner of informational or directional material, not to mention postcards and greeting cards. A great deal of occasional publishing, of pamphlets and booklets, is done by organizations to further particular aims or to spread particular views; *e.g.,* by religious groups, activist organizations, philanthropic societies, and political parties. This kind of publishing is sometimes subsidized.

Book publishing

The form, content, and provisions for making and distributing books have varied widely during their long history, but in general it may be said that a book is designed to serve as an instrument of communication. The Babylonian clay tablet, the Egyptian papyrus roll, the medieval vellum codex, the printed paper volume, the microfilm, and various other combinations have served as books. The great variety in form is matched by an equal variety in content. Both Shakespeare's collected plays, first published in 1623, and the most ill-conceived and trivial tract published in that or any other year were designed as instruments of communication.

The book is also characterized by its use of writing or some other system of visual symbols (such as pictures or musical notation) to convey a meaning. As a sophisticated medium of communication, it requires mastery of the hard-won skills of reading and writing. Another distinguishing feature is publication for tangible circulation. A temple column with a message carved on it is not a book. Signs and placards that are easy enough to transport are made to attract the eyes of passers-by from a fixed location and thus are not usually considered books. Private documents not intended for circulation also are not considered to be books.

A book, for the purpose of this discussion, is a written (or printed) message of considerable length, meant for public circulation and recorded on materials that are light yet durable enough to afford comparatively easy portability. Its primary purpose is to carry a message between people, depending on the twin faculties of portability and permanence. As such, the book transcends time and space to announce, to expound, and to preserve and transmit knowledge. Books have attended the preservation and dissemination of knowledge in every literate society. The following account, keeping mainly within the scope of civilization as it developed in western Europe and North America, considers the book as it appeared at different times in history, the characteristic content and survival of copies and texts, and the means of production and distribution.

THE ORIGINS OF BOOKS

How soon after the invention of writing men began to make books is uncertain because the books themselves have not survived. The oldest surviving examples of writing are on clay or stone. The more fragile materials used for writing at various times have generally perished. The earliest known books are the clay tablets of Mesopotamia and the papyrus rolls of Egypt. There are examples of both dating from the early 3rd millennium BC.

Books on clay tablets. The ancient Sumerians, Babylonians, Assyrians, and Hittites wrote on tablets made from water-cleaned clay. Although these writing bricks varied in shape and dimension, a common form was a thin quadrilateral tile about five inches long. While the clay was still

wet, the writer used a stylus to inscribe it with cuneiform characters. By writing on every surface in small characters, he could copy a substantial text on a single tablet. For longer texts he used several tablets, linking them together by numbers and catchwords as is done in modern books.

Book production on clay tablets probably continued for 2,000 years. The nature and volume of the surviving records from Mesopotamia and Asia Minor indicate a heavy emphasis on the preservative function of writing and the book. Either dried in the sun or baked in a kiln, clay tablets were almost indestructible. The latter process was used for texts of special value, legal codes, royal annals, and epics to ensure greater preservation. Buried for thousands of years in the mounds of forgotten cities, they have been removed intact in modern archaeological excavations. The number of clay tablets recovered approaches 500,000, but new finds continually add to the total. The largest surviving category consists of private commercial documents and government archives. Of the remainder, many are duplications of texts.

Clay tablets are usually associated with cuneiform writing, a script that takes its modern name from the wedge-shaped (from Latin *cuneus,* "wedge") marks made by the stylus in clay. When the Aramaic language and alphabet arose in the 6th century BC, the clay tablet book declined because clay was less suited than papyrus to the Aramaic characters.

The Egyptian papyrus roll. The papyrus roll of ancient Egypt is more nearly the direct ancestor of the modern book than is the clay tablet. Papyrus as a writing material resembles paper. It was made from a reedy plant of the same name that flourishes in the Nile Valley. Strips of papyrus pith laid at right angles on top of each other and pasted together made cream-coloured papery sheets. Although the sheets varied in size, ordinary ones measured about five to six inches wide. The sheets were pasted together to make a long roll. To make a book, the scribe copied a text on the side of the sheets where the strips of pith ran horizontally, and the finished product was rolled up with the text inside.

The use of papyrus affected the style of writing just as clay tablets had done. Scribes wrote on it with a reed pen or brush and inks of different colours. The result could be very decorative, especially when done in the monumental hieroglyphic style of writing, a style best adapted to stone inscriptions. The Egyptians created two cursive hands, the hieratic (priestly) and the demotic (a simplified form of hieratic suited to popular use), which were better adapted to papyrus.

Compared with tablets, papyrus is fragile, yet an example is extant from 2500 BC; and stone inscriptions that are even older portray scribes with rolls. This amazing survival is partly the result of the dry climate of Egypt, in which some papyrus rolls survived unprotected for centuries while buried in the desert sands. The practice of certain Egyptian funerary customs also contributed to the preservation of many Egyptian books. Obsessed by a concern with life after death, they wrote magical formulas on coffins and on the walls of tombs to guide the dead safely to the gates of the Egyptian underworld. When the space thus provided became insufficient, they entombed papyrus rolls containing the texts. These mortuary texts are now described collectively as the *Book of the Dead,* although the Egyptians never standardized a uniform collection. Such books, when overlooked by grave robbers, survived in good condition in the tomb. Besides mortuary texts, Egyptian texts included scientific writings and a large number of myths, stories, and tales.

Book of the Dead

Quotations from ancient writings show that scribes were highly regarded in ancient Egypt. They were the priests and government officials employed in the temples, pyramid complexes, and the courts of the pharaohs. The Greek historian Herodotus reported that Egyptian embalmers did a thriving business in copies of the *Book of the Dead.*

Chinese books. The Chinese, though not so early as the Sumerians and the Egyptians, were the third people to produce books on an extensive scale. Although few surviving examples antedate the Christian Era, literary and archaeological evidence indicates that the Chinese had writing and probably books at least as early as 1300 BC. Those primitive books were made of wood or bamboo strips bound together with cords. Many such books were burned in 213 BC by the Ch'in emperor Shih Huang-ti, who feared the strength of the tradition they embodied. The fragility of materials and the damp climate resulted in the loss of other ancient copies. Some books escaped, however, and these, together with whatever books may have been produced in the intervening period, constituted a large enough body for a Chinese national bibliography to appear in the 1st century BC. This was prepared by a corps of specialists in medicine, military science, philosophy, poetry, divination, and astronomy. A classified list of works on tablets and on silk, it mentioned 677 books. With such a tradition, the survival of Chinese texts was assured by continuous copying and was not dependent on the capacity of a lone example to withstand the wear of the centuries.

BOOKS IN CLASSICAL ANTIQUITY

Greek books. The Greeks adopted the papyrus roll and passed it on to the Romans. Although both Greeks and Romans used other writing materials (waxed wooden tablets, for example), the Greek and Roman words for book show identification with the Egyptian model. Greek *biblos* ("book") can be compared with *byblos* ("papyrus"), while the Latin *volumen* ("book") signified a roll. It has been suggested that papyrus was continuously in use in Greece from the 6th century BC, and evidence has been cited to indicate its use as early as 900 BC. Objects called books are mentioned by ancient Greek writers as having been in use in the 5th century BC. The oldest extant Greek rolls, however, date from the 4th century BC.

The 30,000 extant Greek papyri permit a generalized description of the Greek book. Rolled up, it stood about nine or 10 inches high and was an inch or an inch and a half in diameter. When the book was unrolled it displayed a text written in the Greek alphabet in columns about three inches wide separated by inch-wide margins. In spite of the Greek proficiency in decorative arts, few surviving books are illustrated. Such illustrations as have survived were of the practical sort found in later scientific books.

Practicality was a mark of the Greek book. The alphabet, although not invented by the Greeks, was adapted and stabilized by them as an instrument of verbal communication rather than of decorative purpose. Unlike the monumental Egyptian survivals in a decorative hand that sometimes exceeded 100 feet in length, Greek rolls seldom exceeded 35 feet in length and featured little embellishment. Such a roll was about as large as could be conveniently held in the hands to read, and it was big enough to contain a book of Thucydides or one of the longer New Testament Gospels. The average Greek book was shorter. Two books (here denoting a subdivision of a text) of Homer written in a later small hand fitted a 35-foot roll.

During the golden age of Athens in the 5th century BC, books were known and used but were lightly regarded as avenues of learning. Great tragedies and comedies, speeches, poems, histories, and lectures were produced, but all evidence indicates that the preferred method of publication at that time was oral. The actor, the orator, the rhapsodist, and the lecturer were supreme.

Given the interests and the scope of inquiry of Periclean Greeks, it is noteworthy that they had books and read them at all. Greek readers were general readers. Though it should not be assumed that all who lived in Athens could read, those who could included more than the narrow circle of scribes and scholars who were trained from youth to reverence books and to make a career of the difficult arts of reading and writing. The Greek alphabet reduced this difficulty, and the nonspecialized content of Greek books made them practical instruments of communication to a general public.

With the coming of Alexander the Great, the outlook of the Greeks was broadened into a universal attitude that was reflected in their use of books. As the Alexandrian kingdoms spread throughout the East, the Greeks were forced to extend their interest to alien peoples and the records of the past. Consequently, the range of matters

worth discussing became too extensive for oral transmission and for the solitary speaker. In the important Hellenistic cities, most notably at Pergamum and Alexandria, centres of learning grew up; these aimed at a world synthesis of knowledge. (A noteworthy example of this synthesizing work was the Septuagint, which was a translation of the Hebrew Scriptures into Greek.) Libraries were a distinguishing feature of these centres. The Museum and the Serapeum at Alexandria were reputed at various times to have from 200,000 to 700,000 rolls. The Ptolemies at Alexandria pursued a vigorous collecting policy in an attempt to acquire good copies of all important texts; and scholars were constantly at work on textual scholarship and the writing of new books. The book superseded the oral presentation as a primary means of publication. Greek writers even refer to the market in books and to prices paid for them. The discovery of surviving papyri in the rubbish heaps of provincial towns indicates that the trade was widely diffused. The large libraries maintained scriptoria in which extensive copying was done. However, survivals are scanty and there is no group of extant examples that bears such close resemblance to each other as to indicate that they were the product of the same scribe or scriptorium. Some surviving rolls bear the mark of professional work; others are amateurish.

The volume of surviving Greek texts is so slender that it arouses speculation about the nature of the large book collections of Alexandria. There are various explanations. First, the Alexandrians were doing textual criticism and required many copies of the same text to carry on the work. Second, the record indicates that the volume of Greek literature was much larger than what has survived, a majority of the texts having been lost. Literary and bibliographical references made by ancient writers and bibliographers indicate, for example, that the dramatists Aeschylus, Sophocles, Euripides, and Aristophanes wrote among them about 330 plays; those surviving number 46. Nearly all of Greek lyric poetry has been lost. Only one-fourth of the texts by Stobaeus, an anthologist of the 5th century AD, survived to modern times.

Survival of works

The survival of Greek texts depended on copying by succeeding generations. No manuscript in the hand of either a Greek or Roman author is extant, and the earliest extant copies of most works date from centuries after the composition. In such circumstances, the greatest factor in survival was the widespread and continuing popularity of a work. The centres of textual criticism fostered the preservation of some texts by establishing a canon of writings to be taught in the schools. This practice proved to be more important for a work's survival than the establishment of the great libraries, because the library collections were destroyed, while the widespread copying of books for use by students ensured that they were physically dispersed over a large area, thus rendering an author's work less vulnerable to local disasters. Finally, the universal interest and application of the content was an important factor that led to the survival of some nonliterary texts through translation into Arabic, Latin, and other foreign languages.

Roman books. Rome was the channel through which the Greek book was introduced to the people of western Europe. When the Romans conquered Greece they carried home Greek libraries to serve as a foundation for similar libraries in Rome. Roman libraries had separate collections of Greek and Latin books; but except for the substitution of the Latin language for Greek, a Roman papyrus roll closely resembled a Greek one in content, and there was much imitation.

The Romans developed a book trade on a fairly large scale. From the time of the 1st-century-BC orator Cicero there is evidence of large scriptoria turning out copies of books for sale. On several occasions Cicero referred to bookshops; the 1st-century-AD poet Martial complained about professional copyists who became careless in their speed; and the 1st-century-AD naturalist Pliny the Elder described the extensive trade in papyrus. The trade decrees of the emperor Diocletian set regulations for determining a price for the copying of books.

Book ownership was widespread among Romans of the upper class. Private libraries were common and were considered the necessary badge of distinction for anyone who aspired to high position or social importance. On the other hand, books were also within reach of less prosperous people because the use of slave labour to multiply copies kept prices relatively low. From a comparative study of prices, it has been concluded that books were cheap enough for people with only moderate incomes to buy them. As many as 30 copies of a work might be made simultaneously by a reader dictating to slave copyists. In many ways these enterprises were prototypes for modern publishing houses. Roman publishers selected the manuscripts to be reproduced; advanced money to authors for rights to the manuscripts, thus assuming the risks of publication; chose the format, size, and price of each edition; and developed profitable markets for their merchandise.

BOOKS IN THE EARLY CHRISTIAN ERA

The codex. The substitution of the codex for the roll was a revolutionary change in the form of the book. Instead of having leaves fastened together to extend in a long strip, the codex was constructed from folded leaves bound together on one side—either the right or the left, depending on the direction of writing. (Some variant forms were bound at the top of the leaves.) The codex enjoyed several advantages over the roll. A compact pile of pages could be opened instantly to any point in the text, eliminating the cumbersome unrolling and rerolling, and facilitating the binding of many more leaves in a single book. In addition, the codex made feasible writing on both sides of the leaf; this was not practical for the roll. Because of its compactness, its ease of opening, and its use of both sides of the leaf, the codex could conveniently contain longer texts. The difference can be illustrated with copies of the Bible. While the Gospel of Matthew reached the capacity of the roll, a common codex included the four Gospels and Acts bound together; and complete Bibles were not unknown.

The folded note tablets used by the Greeks and the Romans may have suggested the codex form, but its development to the point of eventual supremacy was related to changes in the world of learning and in the materials for making books. The change in the scholarly outlook came from the rise of Christianity; the new material was vellum or parchment.

Vellum and parchment. Vellum and parchment are materials prepared from the skins of animals. Strictly speaking, vellum is a finer quality of parchment prepared from calf skins, but the terms have been used interchangeably since the Middle Ages. The forerunner of parchment as a writing material was leather. Egyptian sources refer to documents written on leather as early as 2450 BC, and a fragmentary Egyptian leather roll of the 24th century BC survives; but leather was rarely used because papyrus was plentiful. The Hebrews also used leather for books. The spectacular discovery of the Dead Sea Scrolls in the 1940s turned up collections of both leather and papyrus rolls that had been stored in earthen jars in caves along the Dead Sea for centuries. These liturgical and biblical books, produced by a Jewish ascetic sect, were written between the mid-2nd century BC and AD 68.

Leather as a writing material

Parchment is a greatly refined form of leather. The skins of various animals—cattle, sheep, and goats being most common—are washed and divested of hair or wool. Then the skin is stretched tight on a frame, scraped thin to remove further traces of hair and flesh, whitened with chalk, and smoothed with pumice. Tradition has it that parchment was invented as the result of book-collecting rivalry between Ptolemy V of Egypt and Eumenes II of Pergamum about 190 BC. Fearing the library at Pergamum might outstrip the collections at Alexandria, Ptolemy placed an embargo on papyrus to prevent his rival from making any more books, whereupon Eumenes made parchment. The fact that both the Greek and Latin words for parchment mean "stuff from Pergamum" offers some support for the tradition.

Although parchment was used to produce book rolls, and although many early codices were made from papyrus, the new writing material facilitated the success of the codex. A sheet of parchment could be cut in a size larger than a sheet of papyrus; it was flexible and durable, and it

could better receive writing on both sides. These qualities were important. In making a parchment or vellum codex, a large sheet was folded to form a folio of two leaves, a quaternion (quarto) of four, or even an octavo of eight. Gatherings were made from a number of these folded sheets, which were then stitched together to form a book. Because papyrus was more brittle and could not be made in large enough sheets, the folio collected in quires (*i.e.*, loose sheets) was the limit of its usefulness. At the same time, because of the vertical alignment of the fibres on one side, papyrus was not well adapted for writing on both sides in a horizontal script.

For 400 years the roll and the codex existed side by side. There are contemporary references to the codex book dating from the 1st century BC; actual survivals date from the 2nd century AD, however. In the 4th century AD vellum or parchment as a material and the codex as a form became dominant, although there are later examples of rolls, and papyrus was occasionally used for official documents until the 10th century. There were similarities between the two forms; an example of the influence of the roll on the codex can be seen in the use of multiple columns on the pages of early codices, much like the columnar writing on the rolls.

Christianity and the book. In books surviving from the first four centuries AD, codices more often contained Christian writings, whereas pagan works were usually written on rolls. Several points in the Christian use of books contributed to a preference for vellum and the codex. First, Christianity was rooted in Judaism, which for centuries had revered sacred writings. The Christians retained the Jewish Scriptures and added some writings of their own, collected in a New Testament. There was strong motivation for preserving these unchanging words on the most durable materials, and vellum was more durable than papyrus. Second, in referring to their sacred writings the Christians made comparative studies of sources. The writings were related, and students liked to refer from one source to another. This reference entailed having a comparatively large volume of writings available and increased the attractiveness of the easy turning of pages possible with a codex. In this respect it is noteworthy that Roman legal scholarship, which also required a comparison of sources, likewise showed an early preference for the codex. A third point was the expressed intention of early Christians to shun pagan literature by using an entirely different form of book. Conversely the clinging of the pagan authors to an outmoded form may be ascribed in part to a conservative resistance to the Christian ideas.

The social potential of books was illustrated by the Christian emphasis on their dissemination. Christianity, which aimed at universality, produced a stream of books, whereas the literary remains of pagan religions are scarce. The process of introducing the universal religion throughout the Roman Empire extended over three centuries, covered thousands of miles, and embraced peoples of the most varied backgrounds and individuals of the greatest differences in rank. The worldwide outlook thus led to a greater dependence on books. Biblical texts and translations, commentaries, polemical tracts, and pamphlets were important in the circumstances, not only to record belief but also to disseminate and explain it.

By the 4th century, the same time that the vellum codex had superseded the papyrus roll, the Christian book had replaced the pagan book in every form. Little of importance was written in the classical tradition after AD 100. The greatest writers of the following three centuries were Christian scholars such as Origen, Pamphilus of Caesarea, Tertullian, St. Augustine, and St. Jerome. Of all Christian books, however, the most numerous survivals are New Testament codices and apocryphal New Testament writings.

THE MEDIEVAL BOOK

The monasteries. The dissolution of the western Roman Empire during the 5th century, and the consequent dominance of marauding barbarians, threatened the existence of books. It was the church that withstood the assaults and remained as a stable agency to provide the security and interest in tradition without which books can be neither disseminated nor wholly enjoyed. Books found refuge in monasteries. The 6th-century Rule of St. Benedict enjoined monks to read books at certain times. The surrounding social chaos placed upon monasteries the responsibility for making books and creating libraries in order to implement the injunction. A more specific model was set by the historian and grammarian Cassiodorus, who, after serving the Ostrogothic kings in high positions, retired from public life in about 540 to found a monastery and establish a scriptorium at Vivarium. The scriptorium was the centre of his interest there. He supervised the copying of books and wrote a guide to learning, the *Institutions of Divine and Human Readings*. He also composed works that presented certain writers as models, discussed rules for editing, and suggested procedures for establishing a scriptorium and a library.

Scriptoria

Following the early examples, monastic houses throughout the Middle Ages characteristically had libraries and scriptoria where monks copied books to add to their collections. Arrangements for this activity varied from place to place. Occasionally the scriptorium was a single large room. Sometimes the copying was done in carrels, individual cells built in the cloister or library. Fittings for the scriptoria were spare; they lacked heat and artificial light. Work was undertaken only during the daylight hours, because fear of fires that might result from artificial light prevented working after dark. The labour (if contemporary complaints can be believed) was hard, for it was often said, "Two fingers hold the pen, but the whole body toils." The scribe sat at a desk copying in silence a text that was spread before him. The monks did not follow the practice of the Roman commercial scriptorium where a reader dictated a book while several scribes made simultaneous copies of it. Instead, after the scribe's work was finished it was proofread and titles and notes were inserted. The book might then be given to an illuminator, who supplied any needed illustrations or decorative devices. Finally, the book was bound. This procedure closely resembles that of modern book production, except that in the scriptoria each step in the preparation of a manuscript was repeated for each copy of a work. Book production was slowed to a trickle, and a monastic library with as many as 600 volumes was considered fairly large.

Bettmann Archives



A medieval monk copying from a text, in a scriptorium.

The medieval book was a codex written on vellum or parchment, although by the 15th century paper manuscripts were normal. Many medieval manuscripts attained a high perfection of colour and form and are renowned for their beauty. Such examples as the Book of Kells from Ireland, the Lindisfarne Gospels from England, and the many brilliant "books of hours" made in France are world-renowned as examples of art. The customary book was less splendid, however. Written in a neat book hand that developed into the models from which printing

types were designed, the manuscript books of the Middle Ages were the models for the first printed books.

Because the monastic book trade was largely internal, the contents of books are evident from the monastic library catalogs. Generally the catalogs grouped the books in three divisions. First came the Bible and commentaries. Writings of the Church Fathers and contemporary theologians followed. Finally there was a smaller section of worldly books—including at various places some classics, mathematics, medicine, astronomy, law, and historical and philosophical writings. Scriptoria flourished throughout Europe. Books in the Greek language were found only in Byzantine monasteries; in western Europe books were written in Latin. Only with the onset of humanistic scholarship in the 14th century, and the rise of important vernacular writers at about the same time, did books in Greek and various vernacular languages assume any prominence in the catalogs of western European monasteries.

Books in vernacular languages

The revival of the secular book trade. For six centuries after Cassiodorus, references to book production outside monasteries are few and hard to interpret. A definite expansion in book production came with the rise of the universities in the 12th century and a revived interest in ancient Greek writings, although these were studied mainly in Latin translation. The universities were located in cities and generated a demand for books. University stationers were established to supply the demand; these were controlled by the universities, which framed regulations about the content and size of books and set prices for sale and for rental. The University of Vercelli in Piedmont, Italy, framed such a regulation in 1228, and many similar acts are recorded for other universities. To satisfy the growing demand, the university stationer, unlike the monastic scriptoria, produced multiple copies of works.

There can be no doubt that books were readily exposed for sale in the 14th century. This is evident in *Philobiblon,* a book finished in 1345 describing the book-collecting activities of Richard de Bury, bishop of Durham. The book relates how the bishop established good relations with stationers and booksellers in England, France, Germany, and Italy by sending advance payments. Evidence from the same century indicates that the stationers were organized in craft guilds in the same way that other trades were organized. A London record of 1357 granted exemption from jury service to writers of text hand (a compressed, angular hand used for the main text of a book). In 1403 the Stationers' Company of London appealed to the city for the right to have their own ordinances.

Humanistic and vernacular books. The manuscript books of the 14th and 15th centuries were affected by the rise of humanism and the increased use of the vernacular languages. The emergence of humanism has long stood as a notable example of the capacity of the book to preserve knowledge through centuries of disinterest and neglect. In the first half of the 14th century the intellectually curious began seeking out texts of classical authors. Many texts were found in monastery libraries, and soon considerable enthusiasm for the style of writing and pagan contents of the classical works developed. Library collections throughout western Europe were searched with the aim of recovering and purifying the classical texts. The restored texts, often with humanistic commentaries, became prized books that were collected by whoever could afford them. The Biblioteca Medicea-Laurenziana Library in Florence, the modern Biblioteca Apostolica Vaticana, and important collections in the Bibliothèque Nationale date from this time period. By 1450 most of the Latin classics had been recovered, and the humanists turned their attention to Greece, even before the fall of Constantinople in 1453 caused the exodus of so many books and scholars from the Eastern capital.

Concurrently with the revived interest in classical literature and language came the production of vernacular books. A vernacular literature had long been growing; and anonymous medieval authors had composed poems and stories of first importance before the 14th century, but their transmission had been largely oral. In the 14th and 15th centuries vernacular books appeared. The anonymous classics were put in writing, and new books by such creative geniuses as Dante, Petrarch, Boccaccio, Chaucer, and Villon appeared.

The expanded literary production found a much larger audience capable of participating in the use and enjoyment of books. Lay princes as well as churchmen promoted learning and were among the patrons of humanism, although the practicing humanists themselves were for the most part ecclesiastics. An increasing number of books were written in the vernacular, and there is evidence that tradesmen and artisans in the cities were learning to read and write. It was partly to them that John Wycliffe directed his English translation of the Bible.

During the 15th century the manuscript book came to resemble its successor, the printed book, in scope. In the wake of the humanists, the content of books expanded to embrace a large sphere of human activity. New authors wrote in the language of the people. Increasing numbers of people enjoyed the advantage of literacy. Books were recognized as objects in trade, and their production and sale were handled by guilds in the same way as other articles of commerce. Paper, which had come to Europe from China by way of Arab traders, was replacing vellum as the material for books. Creation of the printing press wanted only ingenuity and patience.

THE AGE OF EARLY PRINTING: 1450–1550

Before the invention of printing, the number of manuscript books in Europe could be counted in thousands. By 1500, after only 50 years of printing, there were more than 9,000,000 books. These figures indicate the impact of the press, the rapidity with which it spread, the need for an artificial script, and the vulnerability of written culture up to that time.

The incunabula

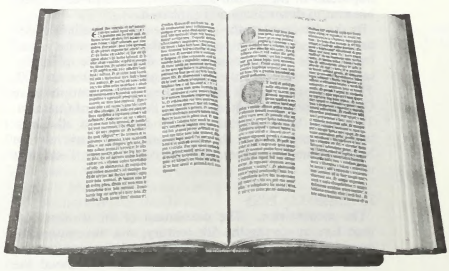
The printed books of this initial period, up to 1500, are known as incunabula; *i.e.,* "swaddling clothes" or "cradle," from a Latin phrase used in 1639 to describe the beginnings of typography. The dividing line, however, is artificial. The initial period of printing, a restless, highly competitive free-for-all, runs well into the 16th century. Printing began to settle down, to become regulated from within and controlled from without, only after about 1550. In this first 100 years, the printer dominated the book trade. The printer was often his own typefounder, editor, publisher, and bookseller; only papermaking and, usually, bookbinding were outside his province.

Early printer-publishers. *Germany.* Printing has been called the great German contribution to civilization; in its early days it was known as the German art. After its invention (about 1440–50) by a goldsmith of Mainz, Johannes Gutenberg, it was disseminated with missionary zeal—and a keen commercial sense—largely by Germans and largely along the trade routes of German merchants.

Gutenberg Bible

Gutenberg himself is usually credited with what is known as the 42-line Bible (1455); the 36-line Bible; and a popular encyclopaedia called the *Catholicon* (1460); however, he lost control of his assets in collection proceedings brought against him by his business partner in 1455. Gutenberg's partner, Johann Fust, and his employee, Peter Schöffer (later Fust's son-in-law), continued the business together

Rare Books and Manuscripts Division, The New York Public Library; Astor, Lenox and Tilden Foundations



The Gutenberg 42-line Bible, printed in Mainz, Ger., in 1455.

after 1455; but Mainz itself never became a major centre of the book trade. It was soon challenged by Strassburg (Strasbourg) where, in 1460–61, Johann Mentelin, with an eye for the lay market, brought out a Bible compressed into fewer pages and followed this with the first printed Bible in German or any other vernacular. A few years later, Cologne had its first press (1464) and became an important centre of printing in the northwest. Cologne's early production was almost entirely in Latin because of the heavy bias of its university toward orthodox Thomist theology. In the south, printing quickly spread to the other great trading centres, Basel (1466), Nürnberg (1470), and Augsburg (1472). Basel became famous for the scholarly editions of Johann Amerbach and Johann Froben, who had the benefit of distinguished advisers, including the Dutch humanist scholar Desiderius Erasmus. In Augsburg, the first press was set up alongside the renowned scriptorium of the Abbey of SS. Ulrich and Afra; and the tradition of the illuminated manuscript was carried over into equally sumptuous editions of illustrated printed books. At Nürnberg, which soon took the lead in the book trade, Anton Koberger operated on a large, international scale. At his peak, he ran 24 presses and had links with Basel, Strassburg, Lyon, Paris, and many other cities. He could be called the first great businessman publisher and the first publisher to rise socially—to membership in the town council. By 1500 there were presses in some 60 German towns, including Lübeck (1475), the head of the Hanseatic League. From there, printing spread to Denmark, Sweden, Rostock, Danzig, and Russia, though the first printer who went to Russia was apparently murdered before he could achieve anything. Printing first began in Russia in 1552, with the help of a printer from Copenhagen.

Italy. It may be said that book printing, after its birth in medieval Germany, was carried to maturity in humanistic Italy. The printing press reached Italy very early (1462–63), via the Benedictine monastery of Subiaco, near Rome, which had strong German connections and a famous scriptorium. Two German printers, Konrad Sweynheim and Arnold Pannartz, who had settled there, soon moved to Rome (1467), where the church encouraged the production of inexpensive books. In Italy as in Germany, however, it was the great commercial towns that became centres of printing and publishing. By 1500, Venice had no fewer than 150 presses; and two Venetian printers exercised a decisive influence on the form of the book: Nicolas Jenson, an outstanding typographer who perfected the roman typeface in 1470, and Aldus Manutius, the greatest printer-publisher of his time. Aldus began printing in 1490 with a series of Greek texts. He then hit on the idea of bringing out inexpensive "pocket editions" for the new readers produced by the humanist movement. Beginning in 1501 and continuing with six titles a year for the next five years, he issued a series of Latin texts that were models of scholarship and elegance. To keep down the cost, Aldus printed editions of 1,000, instead of the more usual 250; and to fill the page economically, he used an italic type designed for him by Francesco Griffo. The Aldine editions were widely copied, by pirating (*i.e.*, without permission from the publisher or payment to him) and other methods, and their dolphin and anchor was one of the first instances of a publisher's device (roughly equivalent to the modern logo).

First pocket editions

France. The way in which printing came to France is of special interest because it shows a publisher (rather than a printer-publisher) in command from the start. In Paris in 1470, the rector and librarian of the Sorbonne invited three German printers to set up a press on university premises. The scholars chose the books and supervised the printing, even to specifying the type. Their preference for roman type greatly helped the eventual defeat of blackletter, or Gothic, type. Among the early French printers were Jean Dupré, a businessman publisher of *éditions de luxe* ("luxury editions"), who set up in 1481, and Antoine Vérard, who began printing in 1485. Vérard was the first to print a Book of Hours, a book containing the prayers or offices appointed to be said at canonical hours, and his work set a standard of elegance for French book production. After 1500, when the full force of the Renaissance began to be felt in France, a brilliant group of scholarly printers, including Josse Bade, Geoffroy Tory, and the Estienne (Stephanus) family, who published without a break for five generations (1502–1674), carried France into the lead in European book production and consolidated the Aldine type of book—compact, inexpensive, and printed in roman and italic types. The golden age of French typography is usually placed in the reign of Francis I (1515–47), one of the few monarchs ever to take a keen personal interest in printing. He was the patron and friend of Robert Estienne. In 1538 he ordered Estienne to give a copy of every Greek book he printed to the royal library, thus founding the first copyright library. In 1539 he laid down a code for printers, which included a prohibition on the use of any device that could be confused with another. Outside Paris, the only significant centre of printing in France was Lyon. While Paris was under the watchful eye of the predominantly Roman Catholic theologians at the Sorbonne, Lyon was able to publish humanist and Protestant works more freely. Among its foremost printers were Johann Trechsel and his sons, Melchior and Caspar; Sebastian Greyff, or Gryphius; and a fine typographer, Robert Granjon. By about 1600, however, religious pressure and the competition of Paris had put an end to printing in Lyon. Thereafter, the French book trade was based entirely in Paris.

Other continental printers. Other parts of Europe established presses quickly; *e.g.*, Utrecht (1470), Budapest (1473), and Cracow (1474), in each case through Germans. In Spain the German connection is particularly evident. The first Spanish press was set up in 1473 at Valencia, where the German trading company of Ravensburg had an important base. Though Madrid became dominant after 1566, publishing flourished in the early period at Barcelona, Burgos, Zaragoza, Seville, and the university towns of Salamanca and Alcalá de Henares. Spain quickly evolved its own distinctive style of book, full of dignity and printed largely in black-letter types. The most remarkable production of the period was the magnificent Complutensian Polyglot Bible (which presented the text in several languages in adjacent columns), sponsored by Cardinal Francisco Jiménez de Cisneros "to revive the hitherto dormant study of the scriptures," which it effectively did. It was printed at Alcalá de Henares, in Hebrew, Chaldee, Syriac, Greek, and Latin, by Arnaldo Guillermo de Brocar, the first great Spanish printer. Editorial work was begun in 1502, the six volumes were printed in 1514–17, and the book finally was issued in 1521 or 1522. In Lisbon, the first printed book was a Pentateuch (the first five books of the Bible) produced in 1489 by Eliezer Toledano; he was reinforced in 1495 by two printers summoned by the Queen of Portugal. From Spain, printing crossed the Atlantic during this early period. In 1539 Juan Cromberger of Seville, whose father, Jacob, had set up a press there in 1502, secured the privilege for printing in Mexico and sent over one of his men, Juan Pablos. In that year, Pablos published the first printed book in the New World, *Doctrina christiana en la lengua mexicana e castellana* ("Christian Doctrine in the Mexican and Castilian Language").

Complutensian Polyglot Bible

England. Compared with the Continent, England in the early days of printing was somewhat backward. Printing only reached England in 1476, and in 1500 there were still only five printers working in England, all in London and all foreigners. Type seems to have been largely imported from the Continent until about 1567, and paper until about 1589 (except for a brief spell during 1495–98). In an Act of 1484 to restrict aliens engaging in trade in England, Richard III deliberately exempted all aliens connected with the book trade in order to encourage its domestic development. In the following year, Henry VII appointed a foreigner, Peter Actors of Savoy, as royal stationer, with complete freedom to import books. For about 40 years, England was a profitable field for continental printers and their agents. This necessary free trade was brought to an end and native stationers protected under Henry VIII, whose acts of 1523, 1529, and 1534 imposed regulations on foreign craftsmen and finally prohibited the free importation of books. It has been estimated that up

to 1535 two-thirds of those employed in the book trade in England were foreigners.

William Caxton and his press

It is thus all the more remarkable that the man who introduced printing to England was a native, William Caxton. After learning to print at Cologne (1471–72), Caxton set up a press at Bruges (about 1474), where he had long been established in business. His first book, *The Recuyell of the Historyes of Troye,* was his own translation from the French, and its production was probably the main reason why this semiretired merchant gentleman took to printing at the age of 50. He then returned to England through the encouragement of Edward IV and continued to receive royal patronage under Richard III and Henry VII. Caxton is important not so much as a printer (he was not a very good one) but because from the first he published in English instead of Latin and so helped to shape the language at a time when it was still in flux. Of the 90-odd books he printed, 74 were in English, of which 22 were his own translations. Some, such as the *Ordre of Chyvalry* and the *Fayttes of Armes,* were for the pleasure of his royal patrons; but his range was wide and included *Dictes and Sayenges of the Philosophers* (1477; his first book in England); two editions of Chaucer's *Canterbury Tales* (the second undertaken because a better manuscript came to hand); *The Fables of Aesop* (in his own translation from the French); Sir Thomas Malory's *Kyng Arthur;* and his largest work, *The Golden Legend,* a compilation of such ecclesiastical lore as lives of the saints, homilies, and commentaries on church services, a considerable editorial labour apart from the printing.

Caxton's press was carried on after his death by his assistant, Wynkyn de Worde of Alsace. In the absence of court connections and also because he was a shrewd businessman, he relied less on the production of expensive books for the rich and more on a wide variety of religious books, grammars and other schoolbooks, and collections of popular tales. He published more than 700 titles, mostly small volumes for the ordinary citizen, and continued Caxton's standardizing of the language, a solid contribution to the native book trade. The best of the early printers was Richard Pynson of Normandy, who began printing in 1492 and became printer to the king in 1508. Pynson, the first to use roman type in England (1509), published the first English book on arithmetic (1522). After his early liturgies and some fine illustrated books, he concentrated mainly on legal works. In 1521 he published Henry VIII's answer to Luther in defense of the papacy, for which the King received the title of *fidei defensor* ("defender of the faith") from the Pope.

Printed illustrations. Although 15th-century printers characteristically were content to exploit the existing book format, their use of printed illustrations in fact produced a new means of expression. Printers used woodcuts to print illustrations by the relief process and experimented with intaglio in copper engravings. Woodcut pictures were produced before metal types, and it was a simple development to make woodcuts in appropriate dimensions for use with type to print illustrated books. Albrecht Pfister of Bamberg was printing books illustrated with woodcuts about 1461. Copper engravings, which were better able to produce fine lines, were especially suitable for the reproduction of maps; among the few incunabula illustrated with engravings is a Ptolemy *Geographia* printed at Rome by Arnoldus Buckinck in 1478. But because engravings required a different press and introduced a separate process into printing, and because experiments with woodcut illustrations were so satisfactory, there was no extensive use of engravings before 1550.

Once a picture was prepared for printing, it could be repeated an indefinite number of times with little loss in detail, accuracy, form, or original vigour. When great artists such as Albrecht Dürer designed woodcuts the result was books of high aesthetic value that could be produced in great numbers. *Hypnerotomachia Poliphili,* printed by Aldus Manutius in 1499, is a monument to the early perfection of the woodcut and to book illustration in general. Equally as important as the reproduction of great art was the opportunity that printed illustrations offered for the faithful reproduction of pictures and diagrams in scientific books. The dawning scientific scholarship profited from the development of printed illustration; it is significant that studies in both anatomy, with its need for precise illustration of the human body, and cartography greatly expanded after development of printed illustrations.

The book trade. The book trade during this early period showed enormous vitality and variety. Competition was fierce and unscrupulous. A printer of Parma in 1473, apologizing for careless work, explained that others were bringing out the same text, and so he had to rush it through the press "more quickly than asparagus could be cooked." Though most of the early firms were small printer-publishers, many different arrangements were made and at least one businessman, Johann Rynmann of Augsburg, published nearly 200 books but printed none of them. Publishing companies, which both financed and guided the printing enterprise, were also tried, as at Milan in 1472 and at Perugia in 1475. Publishers were not slow to promote their books. The medieval scribes had placed their names, the date when they finished their labours, and perhaps a prayer or a note on the book, at the end of their codices. From this grew the printer's colophon, or tailpiece, which gave the title of the book, the date and place of printing, the name and house device of the printer, and a bit of self-advertisement. By about 1480, the information of the colophon began to appear at the front of the book as a title page, along with the title itself and the name of the author. Advertisements for books, in the form of handbills or broadsheets, are known from about 1466 onward, including one of Caxton's of 1477, ending with a polite request not to tear it down, *Supplico stet cedula* ("Please let the poster stand"). Publisher's lists and catalogs occur almost as early. Distribution of books along the trade routes, with their courier services, appears to have been highly effective. In 1467, for instance, a bookseller in Riga on the Baltic coast had a stock of books issued by Schöffer in Mainz on the Rhine. Another effective channel for the distribution of books was the regular trade fairs, especially those at Frankfurt and at Stourbridge in England. Besides the stationers, who may sometimes have functioned as wholesalers, there were also retailers known as "book-carriers."

Publishing's effect on languages

Early publishing had a profound effect on national languages and literatures—it began at once to create, standardize, and preserve them. Caxton, in the preface to his translation of the *Aeneid,* after telling a story of confused dialects, ended up "Lo! what should a man in these days now write, eggs or eyren?" By choosing words "understood of common people" and by printing all he could of English literature, he steered the English language along its main line of development. The early printing of great vernacular works, such as those of Dante, Petrarch, and Boccaccio in Italy, or a vernacular Bible, such as that of Luther in Germany, gave many languages their standard modern form. The French language owes much to the early printer-publisher Robert Estienne, who is known not only for his typographical innovations of the 1530s but also for his dictionaries. His work in the latter field caused him to be known as the father of French lexicography. Up to 1500, about three-quarters of all printing was in Latin, but thereafter that proportion steadily declined as books appeared in the vernacular and reached an ever-widening public.

Controls over printing. The church at first had every reason to welcome printing. Bibles (preferably in Latin), missals, breviaries, and general ecclesiastical literature poured from the early presses of Europe; and the first best-seller in print was a devotional work by Thomas à Kempis, *De imitatione Christi* (*Imitation of Christ*), which went through 99 editions between 1471 and 1500. Such sales were matched, however, between 1500 and 1520 by the works of the humanist Erasmus, and, after 1517, by those of the "heretic" Martin Luther. The church had always exercised censorship over written matter, especially through the universities in the late Middle Ages. As the works of the reformers swelled in volume and tone, this censorship became increasingly harsh. The Inquisition was restored, and it was decreed in 1543 that no book might be printed or sold without permission from the church.

The Index of Forbidden Books

Lists of banned books were drawn up, and the first general *Index Librorum Prohibitorum* (Index of Forbidden Books) was issued in 1559. Dutch printers in particular suffered under the Inquisition and a number went to the stake for publishing Protestant books. To avoid such a fate, some resorted to the fake imprint, putting a fictitious printer or place of publication on the title page, or omitting that information.

Censorship also began to be exercised in varying degrees by individual rulers, especially in England, where church and state had been united under Henry VIII after his defection from Rome. The Tudors, with little right under common law, arrogated to themselves authority to control the press. After about 1525, endless proclamations were issued against heretical or seditious books. The most important was that of 1538 against "naughty printed books," which made it necessary to secure a license from the Privy Council or other royal nominees for the printing or distribution of any book in English.

The Stationers' Company

In this attempt at control, an increasingly prominent part came to be played by the Stationers' Company. Since its formation in 1403 from the old fraternities of scriveners, limners, bookbinders, and stationers, it had sought to protect its members and regulate competition. Its first application for a royal charter in 1542 seems to have gone unheeded; but in 1557, an important date in the English book trade, the interests of the crown (then the Roman Catholic Mary Tudor), which wanted a ready instrument of control, coincided with those of the company (under a Roman Catholic first Master), and it was granted a charter that gave it a virtual monopoly. Thereafter, only those who were members of the company or who otherwise had special privileges or patents might print matter for sale in the kingdom. Under the system of royal privileges begun by Henry VIII, a printer was sometimes given the sole right to print and sell a particular book or class of books for a specified number of years, to enable him to recoup his outlay. This type of regulation now came into the hands of the Stationers' Company. After licensing by the authorities, all books had to be entered in the company's register, on payment of a small fee. The first stationer to enter a book acquired a right to the title or "copy" of it, which could then be transferred as might any other property. As the beginning of a system of copyright, this procedure was an admirable development; but the grip that the company obtained and its self-interested subservience to authority were to stunt the free growth of the English book trade for the next 100 years.

THE FLOURISHING BOOK TRADE: 1550–1800

From the mid-16th through the 18th century, there were virtually no technical changes in the methods of book production, but the organization of the trade moved gradually toward its modern form. The key functions of publishing, selecting the material to be printed and bearing the financial risk of its production, shifted from the printer to the bookseller and from him to the publisher in his own right; the author, too, at last came into his own. The battle with the censor became increasingly fierce before any measure of freedom of the press was allowed. Literacy grew steadily and the book trade expanded, both within and beyond national boundaries.

Advances in Europe and America. *The Netherlands.* After 1550, the lead in book publishing passed for a time to the Netherlands. The business founded at Antwerp in 1549 by Christophe Plantin, a Frenchman by birth, came to dominate the Roman Catholic south of the country, both in quantity and in quality. Its finest production was probably the eight-volume polyglot Bible (1569–72), the *Biblia regia* ("Royal Bible"). The firm was carried on for generations by the descendants of Plantin's son-in-law, Joannes Moretus (Jan Moerentorf). In the Protestant north, the house of Elzevir occupied a similar position. After its founding by Louis Elzevir, who issued his first book in 1593, its publishing endeavours were extended by succeeding generations to The Hague, Utrecht, and Amsterdam, with varying fortunes. A duodecimo (small-format) series of classical Latin texts that the Elzevirs began issuing in 1629 more than matched the earlier Aldine editions in excellence at a reasonable standard price. The Dutch, as great seafarers, were preeminent publishers of atlases, a word that was first used when the maps of Gerardus Mercator were published by his son, Rumold, in 1595. The high skill of Dutch engravers also went into their emblem books (books of symbolic pictures with accompanying verse), for which there was a considerable demand between 1580 and 1650.

The *Biblia regia*

France. In France, as the monarchy reasserted its authority after the wars of religion, publishing, which was already heavily concentrated in Paris, became increasingly centralized. In 1620, Louis XIII set up a private press in the Louvre, the Imprimerie Royale, which the Cardinal de Richelieu turned into a state establishment in 1640. This national press established and continued to maintain a standard of excellence for book production in France.

Louis XIII also tried to regulate the trade in books. By an ordinance of 1618, a body called the Chambre des Syndicats was established. It was organized along lines similar to the Stationers' Company in England, but because it contained two royal nominees, its control was even more absolute. The power of censorship, though it remained for a time with the Sorbonne, also passed eventually to officials of the crown. Under these conditions, publishers were inclined to exercise caution; as in other strictly regulated areas, more controversial works first appeared outside the country (often in Holland or Geneva) or under a false imprint. But French books fully upheld the influence of French taste in Europe. The vernacular made strong inroads (even as the language of scholarship) when Descartes published his *Discours de la méthode* (*Discourse on Method*) in French in 1637. A remarkable publishing feat of the 18th century was the 70-volume collected edition of Voltaire's works (1784–89) produced at Kehl, in Baden, by Pierre-Augustin Caron de Beaumarchais, the author of *The Barber of Seville* and *The Marriage of Figaro*. Beaumarchais bought the printing equipment (especially for the purpose) from the widow of the great English typographer John Baskerville.

Publishing the works of Voltaire

Germany. After the Reformation, the intellectual life of Germany was predominantly Protestant and the book trade almost entirely so. Through its book fairs, Frankfurt had become the centre of German publishing and even a kind of European clearinghouse. In 1579, however, the fair came under the supervision of the imperial censorship commission (Frankfurt being a free imperial city), and this action gradually killed it. After about 1650, though Frankfurt continued to be important for the production of type and illustrated books, the centre of the trade shifted decisively to Leipzig. There, an enlightened government and a celebrated university favoured cultural life and patronized book publishing. Two Leipzig firms dating from the 17th century survive to the present day: that founded by Johann Friedrich Gleditsch in 1694, which was taken over by the firm of F.A. Brockhaus in 1830, and that founded by Moritz Georg Weidmann in 1682. A Weidmann partner, Philipp Erasmus Reich, was known in the 18th century as "the prince of the German book trade." He could be said to have invented the net price principle (see below *Price regulation*) and the idea of a booksellers' association (1765), which in 1825 became the Börsenverein der Deutschen Buchhändler, a unique organization of publishers, wholesalers, and retailers. Toward the end of the 18th century, three publishers were outstanding—Georg Joachim Göschen in Leipzig; Johann Friedrich Cotta in Tübingen and Stuttgart; and Johann Friedrich Unger in Berlin, all of whom had a share in publishing Schiller and Goethe. Unger also published the magnificent translation of Shakespeare by August von Schlegel (8 vol., 1797–1810).

England. In the golden age of Elizabeth I, publishing in England was probably at its most turbulent. Through her Injunctions of 1559, Elizabeth confirmed the charter of the Stationers' Company and the system of licensing by the crown or its nominees, which now included church dignitaries. Controls were tightened in 1586 by a decree of the Star Chamber, which confined printing to London, except for one press each in the universities of Oxford and Cambridge. The Stationers' Company was given powers

to inspect printing offices and to seize and destroy offending material or presses, which it zealously did, as much in defense of its monopoly as in support of the crown. But despite stern measures, the great religious question, in which Elizabeth steered a precarious course between Papists and Puritans, continued to be fought out with secret presses on both sides.

Within the legitimate trade, the booksellers had begun to get the upper hand. The incorporation of the Stationers' Company, like that of other London companies, was in itself an indication of the ascendancy of the trader over the craftsman. During the reign of Elizabeth, as part of a developing system of monopolies, the former short-term privileges for publishing certain works or classes of works (called "copies") were granted, for a consideration, as life patents with rights of reversion, such as those enjoyed by Richard Tottel for law books or John Day for alphabet books and catechisms. The printers had already been driven by high costs to make arrangements with the booksellers, to their own disadvantage. Gradually, the very best copies came into the hands of a rich few, who ruled the company and who, in the words of a report of 1582, "keepe no printing howse, neither beare any charge of letter, or other furniture but onlie paye for the workmanship." In 1577 an abortive revolt was led by John Wolfe, who maintained his right to print whatever he pleased. Wolfe was twice imprisoned, but he was finally bought off by admission to the Stationers' Company. In 1584 to still the discontent, some of the rich patentees surrendered a number of copies to the company for the benefit of its poorer members. These were supplemented in 1603, when King James I withdrew some patents from individuals and sold them to the company, again for "the poore of the same."

In this way the Stationers' Company itself became a publishing organization; and having tasted the advantages, it bought up more and more copies on its own account. These came to be divided into "stocks," the English Stock, Bible Stock, Irish Stock, Latin Stock, and Ballad Stock, with shares allocated among its members. By 1640, through leasing the patents at its discretion, the company controlled most of the printing offices in London. The benefit to the poor stationers was somewhat marginal and the monopoly and lack of foreign stimulus caused England to lag behind the Continent in standards of production.

For all that, the privileged men were sometimes good publishers; a few even supported authors during their labours. Some landmarks of the period were John Lyly's *Euphues*, published by Gabriel Cawood (1578); Sir Thomas North's translation of Plutarch's *Lives*, so important for Shakespeare, by Thomas Vautroullier (1579); Edmund Spenser's *Faerie Queene*, by William Ponsonbie (1589–96), and the Authorised (or King James) Version of the Bible (1611), which was completed in a room at Stationers' Hall and printed at the expense of Robert Barker, the king's printer.

Publication of drama was left, along with much of the poetry and the popular literature, to publishers who were not members of the Stationers' Company and to the outright pirates, who scrambled for what they could get and but for whom much would never have been printed. To join this fringe, the would-be publisher had only to get hold of a manuscript, by fair means or foul, enter it as his copy (or dispense with the formality), and have it printed.

Printing of Shakespeare's works

Just such a man was Thomas Thorpe, the publisher of Shakespeare's sonnets (1609); the mysterious "Mr. W.H." in the dedication is thought by some to be the person who procured him his copy. The first Shakespeare play to be published (*Titus Andronicus*, 1594) was printed by a notorious pirate, John Danter, who also brought out, anonymously, a defective *Romeo and Juliet* (1597), largely from shorthand notes made during performance. Eighteen of the plays appeared in "good" and "bad" quartos before the great First Folio in 1623. A typical imprint of the time, of the "good" second quarto of *Hamlet* (1604), reads: "Printed by I.R. for N.L. and are to be sold at his shoppe under Saint Dunston's Church in Fleetstreet"; *i.e.*, printed by James Roberts for Nicholas Ling. For the First Folio, a large undertaking of more than 900 pages, a syndicate of five was formed, headed by Edward Blount and William Jaggard; the Folio was printed, none too well, by William's son, Isaac.

Attempts to control the publishing business continued through most of the 17th century. In 1637 the Star Chamber issued its most drastic decree, which confirmed previous enactments, laid down detailed licensing procedures, reduced the total number of printers to 23, and prescribed severe penalties for offenses. Four years later, however, the Star Chamber itself was swept away by Parliament, and in the ensuing uncertainty the book trade had a taste of freedom. This new situation quickly alarmed not only the Stationers' Company, which saw its privileges vanishing, but also Parliament, which proved to be as reactionary as the royalists. In 1643 it passed an ordinance restoring both licensing and the powers of the company. It was this act that prompted John Milton to write his *Areopagitica*, a noble and powerful plea for freedom of the press, which vigorously argued against every claim of justification for censorship. After the Restoration, the Licensing Act of 1662 was ruthlessly enforced until after the Great Plague of 1664–65, when its rigours were mitigated; it lapsed in 1679. James II revived licensing in 1685, but Parliament refused to renew it in 1694. Thereafter, restraint, harassment, and persecution continued, but by other means, under a broad interpretation of the meaning of libel. With the end of licensing and the gradual breakdown of the whole guild system, the Stationers' Company declined in importance; but it remained useful in connection with copyright.

In the latter part of the 17th century, publishing expanded rapidly, partly through the rise of the periodical press (see below *Magazine publishing*), with its growing body of writers and readers. Successful books became highly profitable, and the author's right to a proper share was more widely recognized. The poet John Dryden is said to have received a total of £1,200 for his *Virgil* (1697), at a time when a shopkeeper might receive £50 a year and a labourer £15. Patronage continued, with all its political implications; but dedications became increasingly cut-and-dried, costing five guineas for a poem, perhaps, or 20 for a play; royalty was naturally expected to pay more. By the 1750s it was virtually at an end; "We have done with patronage," said Dr. Johnson. In its place came the public at large, to whom Henry Fielding dedicated his satirical piece for the theatre, *The Historical Register for the Year 1736*, on its publication in the following year. In the expanding literary market, the enterprising publisher tried to collect all the most promising authors to write for him. Through his personal inclinations, his sense of public taste, and his readiness to risk novelty, he began to play a part of his own in the course of literary development. As this side of the business absorbed more and more of his energies, the final separation of publisher and bookseller came about, though never so decisively as that between bookseller and printer.

Payment to authors

In Britain this transition was marked—and fostered—by the passing of the Copyright Act of 1709, the first of its kind in any country. It was "An Act for the encouragement of Learning, by vesting the copies of printed books in the authors or purchasers of such copies during the times therein mentioned." For books printed before the act, the time was 21 years, "and no longer" (from April 10, 1710, when the act came into force). For works not yet published, the copyright was 14 years, "and no longer," though if the author was still living at the end of that time period, the copyright returned to him for a further period of 14 years. Penalties were also laid down, and registration at Stationers' Hall was made a condition for their enforcement.

Copyright Act of 1709

The Copyright Act of 1709, like all subsequent measures, tried to strike a balance between the needs of those who make a living from books—writers, printers, and publishers—and the interests of the reading public, which are far from identical; it tried, in other words, to limit privilege as well as piracy. The terms it set were amended when they came to be regarded as too short; but in setting any term at all, and in focusing attention on the author as prime producer, it was revolutionary.

The fathers of modern publishing in Britain, which may

be said to date from this time, were Jacob Tonson, who acquired the copyright of Milton's *Paradise Lost* and published works by Dryden, Joseph Addison, Sir Richard Steele, and Alexander Pope, among others; and Barnaby Bernard Lintot, who also published Pope, paying him some £5,300 in all for his verse translation of the *Iliad.* Charles Rivington began publishing in 1711, and Longmans, Green & Co. was begun in 1724 by Thomas Longman when he bought the business of William Taylor, the publisher of Daniel Defoe's *Robinson Crusoe.* At mid-century the best-known figure in the trade was Robert Dodsley, the footman-poet who was befriended by Pope. Among "his" authors were Pope himself, Oliver Goldsmith, Laurence Sterne, and Samuel Johnson. He is credited with suggesting the idea of the *Dictionary* to Dr. Johnson, and his name heads the list of "gentlemen partners" who financed it. Such cooperative associations were popular as a means of financing longer works. They were known as congers and developed into a system of shares in individual books, which could be bought and sold at will.

America. During the 18th century, the book trade in the American colonies began to flourish. Printing had begun there in 1639, when the first printers, Stephen Day (also spelled Daye) and his two sons, left Cambridge, Eng., for Cambridge, Mass. After printing *The Oath of a Free-Man* (1638) and *An Almanack for the Year of Our Lord 1639,* the Days produced their first book, *The Whole Booke of Psalmes,* in 1640. In the early years of the Colonies, Cambridge, Mass., had the sole privilege of printing, but the monopoly was broken in 1674, when Marmaduke Johnson, who had come over to print an Indian Bible (1663), moved his press to Boston. Gradually others followed—Philadelphia had a press in 1685; New York City, in 1693. It was difficult for the colonial printer, as for any small printer, to produce large works because of a shortage of type; but patronage by the government helped to give his products a dignified style. Almanacs, primers, and law books were the staples of book production; works of theology formed the leading category. Until 1769 American printers bought their presses from England, but thereafter they acquired their equipment and supplies, including ink and paper, domestically. Books were sold in various ways—by subscription, by the printer himself, by hawkers, and through shopkeepers. Though Massachusetts passed a law against hawkers in 1713, it carefully excluded book peddlers, who had a valuable function in rural areas. The first bookseller seems to have been Hezekiah Usher of Boston, who added books to his general merchandise in about 1647.

Spread of education and literacy. The great increase in available reading matter after about 1650 both resulted from and promoted the spread of education to the middle classes, especially to women. The wider readership is reflected among the middle classes by the rich development of the prose novel in the 18th century and, among the less well-to-do, by the large sales of almanacs and chapbooks. The almanacs, such as Benjamin Franklin's *Poor Richard's Almanack* (Philadelphia, 1732–64), usually consisted of miscellaneous information and homiletic matter (collections of religious and moral sayings), while the chapbooks, consisting of a few pages cheaply produced, contained a popular story or ballad illustrated by a crude woodcut; a well-known example is *The famous and remarkable History of Sir Richard Whittington, three times Lord Mayor of London* (1656).

National copyright libraries

Growth of libraries. Growth in the book trade led naturally to growth in libraries. Some of the oldest collections of books developed into national "copyright libraries," of immense value for bibliographical purposes. Sir Thomas Bodley opened his famous library at Oxford in 1602, and in 1610 the Stationers' Company undertook to give it a copy of every book printed in England. Later, Acts of Parliament required the delivery of copies of every book to a varying number of libraries, the most important being the library of the British Museum, founded in 1759. This idea of a definitive collection was adopted elsewhere; *e.g.,* in the United States, where the Librarian of the Library of Congress (founded in 1800) was appointed copyright officer in 1870.

In the 18th century a characteristic development was the commercial lending library, and in the 19th the free public library. Despite the fears of publishers and booksellers that the availability of books in library collections would discourage people from purchasing copies for their own use, circulating libraries have promoted rather than diminished the sale of books, besides being a steady market in themselves.

Decline of censorship. From the 18th century censorship in most Western countries diminished. It was abolished in Sweden in 1766, in Denmark in 1770, and in Germany in 1848. The clearest statement, to which lip service, at least, is now almost universally paid, came from the French National Assembly in 1789: "The free communication of thought and opinion is one of the most precious rights of man; every citizen may therefore speak, write and print freely." In the United States, no formal censorship has ever been established; control over printed matter has always been exercised through the courts under the law of libel. This was also the case in Britain after the lapsing of the Licensing Act in 1694; but two important steps had yet to be taken: in 1766, Parliament put an end to general warrants (*i.e.,* for the arrest of unnamed persons and for the seizure of unspecified papers); and in 1792, Charles James Fox's Libel Act finally gave the jury the right to decide the issue, which had previously depended mainly on the judge. Subsequent efforts to suppress printed matter have centred on questions of libel, obscenity, or national security.

MODERN PUBLISHING: FROM THE 19TH CENTURY TO THE PRESENT

The 19th century. In the 19th century a whole new era in publishing began. A series of technical developments, in the book trade as in other industries, dramatically raised output and lowered costs. Stereotyping, the iron press, the application of steam power, mechanical typecasting and typesetting, new methods of reproducing illustrations—these inventions, developed through the century and often resisted by the printer, amounted to a revolution in book production. Paper, made by hand up to 1800, formed more than 20 percent of the cost of a book in 1740; by 1910 it had fallen to a little more than 7 percent. Bindings, too, became less expensive. After 1820 cloth cases began to be used in place of leather, and increasingly the publisher issued his books already bound. Previously, he had done so only with less expensive books; the bindings of others had been left to the bookseller or private buyer. In Europe and America, expansion and competition were the essence of the century, and the book trade had a full share of both. While the population of Europe doubled, that of the United States increased fifteenfold. Improved means of communication led to wider distribution, and a thirst for self-improvement and entertainment greatly expanded readership, leading to a rapid growth in every category of book from the scholarly to the juvenile. The interplay of technical innovation and social change was never closer. As the development of the railways encouraged people to travel, a demand arose for reading material to lessen the tedium of the long journeys. The only victim in the book trade was design, part of the price that was paid almost universally in the first phase of machine production.

Revolutionary inventions

Publishing was now well established, with its characteristic blend of commerce and idealism. Their tendency to specialize made French and German publishers more vulnerable to change than their British colleagues, who aimed as a rule at greater flexibility. Literary and intellectual currents were flowing strongly and the number of new books rose by leaps and bounds. Rough figures for Britain indicate 100 new titles per year up to about 1750, rising to 600 by 1825, and to 6,000 before the end of the century. Equally characteristic was the appearance of popular series at low prices, "literature for the millions," as Archibald Constable was the first to call it. The forerunner was the publisher John Bell's *The Poets of Great Britain* (rivaling Dr. Johnson's), which appeared in 1777–83, in 109 volumes at six shillings each, when even a slim volume usually cost a guinea or more.

The low-priced popular series in England

By the 1850s the application of the new techniques of

mass production had brought down the price of an inexpensive reprint to one shilling, as in the Railway Library of novels (George Routledge, 1,300 vol., 1848–98), for instance, or in the three series of classics issued by H.G. Bohn in 1846, 1850, and 1853. Later reprints were cheaper still. Least expensive was Cassell's National Library (209 vol., 1886–90), bound in paper for threepence and in cloth for sixpence—that is, one-twelfth the price of the Bell set. On the Continent, two German series were outstanding. The Tauchnitz Collection of British and American Authors (1841–1939) became known to thousands of travelers. Tauchnitz voluntarily paid royalties and forbade the sale of his editions in Britain. Even more successful was Reclams Universal-Bibliothek, begun in 1867. An important factor in this series, as in others later, was the release of works through the expiration of copyright.

Book piracy. In the United States, publishing gradually became centralized in a few cities—Philadelphia, Boston, and New York City. Although American literature put down strong roots during the 19th century, piracy from Britain rose to great heights. There was sharp competition to be the first to secure proofs of any important new book. Publishers waiting at the dockside for new British books could produce an American edition almost within hours, as they did in 1823 with Sir Walter Scott's *Peveril of the Peak.* In the absence of international copyright agreements, the British author usually received nothing, but there were honourable exceptions; Harper Brothers, for instance, paid considerable royalties to Charles Dickens and Thomas Macaulay, among others. There was also at least one famous case of piracy in reverse. When Harriet Beecher Stowe's antislavery novel *Uncle Tom's Cabin* came out in the United States in 1852, 1,500,000 copies rapidly appeared in England, some editions selling for sixpence. Though it can be argued by some people that piracy is not only inevitable but possibly even desirable for the sake of cultural diffusion in some circumstances, the availability of inexpensive foreign books, if prolonged as it almost certainly was in the United States, can damage the prospects for home-produced literature. Though there were some household names, such as Washington Irving, James Fenimore Cooper, Ralph Waldo Emerson, and Henry Wadsworth Longfellow, American writers in general had a lean time; and the strong development of the magazine short story and the lecture tour in the United States has been attributed in part to their difficulties. Toward the end of the century American publishing was further enriched by translations of many foreign works, as a result of the flood of immigrants into New York City.

Price regulation. While 19th-century publishing was competitive and individualistic, its growing volume pointed increasingly to the need for greater organization. A major problem, once booksellers had become distinct from publishers, was suicidal price-cutting in the retail trade. Though price regulation ran counter to accepted notions of free competition and met with fierce opposition, in the general interest of the industry it was inevitable. Like copyright, it helped to provide a firm structure within which fair prices could be calculated. The net price principle, first raised in the previous century by the German publisher Reich, was adopted in Germany in 1887 through the work of the Börsenverein, the trade organization founded in 1825. Under this principle, the publisher allows a trade discount to the bookseller only on condition that the book is sold to the public at not less than its "net published price" as fixed by the publisher. In England, a first attempt to introduce the net price principle by the booksellers in the 1850s was condemned to failure by the Free Traders; but toward the end of the century some publishers, led by Alexander Macmillan, began to replace the variable discounts by fixed prices. To press for the new system, the Associated Booksellers of Great Britain and Ireland was formed in 1895, and the Publishers Association was created in 1896. These two organizations then worked out the Net Book Agreement (1901), primarily through the efforts of Frederick (later Sir Frederick) Macmillan. The principle has since been generally adopted, although only to a limited extent in the United States. At roughly the same time, the founding of the Society of Authors (1884) in England and the Authors' League (1912) in the United States helped to standardize fair dealing over contracts and the payment of royalties to authors.

The net price principle

Trade catalogs. The trade also became better organized in the provision of comprehensive catalogs of current books. These began as early as the twice-yearly book fairs at Frankfurt (first catalog 1564) and Leipzig (first catalog 1594). So great was the value of the Frankfurt catalog that an English edition was published in 1617–28. Eventually, all such semiprivate ventures, as *A Catalogue of all the Books Printed in the United States* (1804) or English catalogs deriving from *The Publishers' Circular* (1837) or *Whitaker's* (1874), became national lists, such as the *Bibliographie de la France* (from 1811), the U.S. *Cumulative Book List* (from 1898), the *Deutsche National-bibliographie* (from 1931), and the *British National Bibliography* (from 1950).

Development of copyright law. Copyright, too, underwent considerable development. By the end of the century, most countries had some provision, and various terms of protection were tried, running from publication or from the date of the author's death. The United States first enacted legislation in 1790, France in 1793, and Germany in 1839. Moves toward an international code began in 1828 in Denmark. They took the form of reciprocal treaty arrangements between individual countries by which foreign authors received the same protection as did native authors. Britain joined the movement in several arrangements between 1844 and 1886. In 1885 a uniform international system of copyright was initiated by the Berne Convention. The customary term of protection is the author's lifetime plus 50 years. Most countries subscribed to the Convention, but not the United States or Russia. The United States continued to protect its domestic printing industry up to 1955, when it joined the Universal Copyright Convention (Unesco 1952). While the Berne Convention prescribed a minimum level of protection, the Universal Convention was based on the concept of "national treatment"—each member country treating works by citizens of other member countries as it would those of its own citizens. Thus the United States was able to enter into an international agreement without the necessity of immediately revising its own copyright law. Since the Universal Convention contained a provision that the Convention would not be applicable between any two countries that belonged to the Berne Union, it served primarily as a treaty between the United States and the countries that recognized international copyright. The Soviet Union became a party to the Berne Convention in 1973.

International copyright code

The 20th century. In the 20th century, the effects of state education in the more advanced countries became increasingly apparent. Standards of living rose, and, as in earlier times, these two conditions brought increased use and publication of books. During the late 1890s and early 1900s, many new publishing houses were founded. In the industrialized countries, though wages were rising, a small business could be staffed economically, and printing costs were such that it was economically feasible to print as few as 1,000 copies of a new book. It was thus comparatively easy to make a start, especially because the long-term credit that printers were prepared to grant made a minimum of capital necessary.

Book publishing grew to a substantial industry, consisting mostly of small units in the Western world but also embracing a number of large concerns, many of which were public corporations employing staffs of 1,000 or more. Specialization became frequent, particularly in educational books, as the needs of the new school populations were realized. Some companies, such as Macmillan, in both its British and American houses, had begun to issue schoolbooks almost by chance; then, as their sales grew most profitably, they developed separate departments for school and college textbooks. Others, such as The American Book Company and Methuen in London, had begun specifically with educational books in mind. For more than one leading London firm, India, despite its high illiteracy rate, began to grow strongly as a market and to repay the care and expense involved in setting up separate Indian branches.

Textbook publishing

The first literary agents. A new factor at this time, which was to change the financial climate for fiction publishers in particular, was the advent of the literary agent. The first agent began business in 1875, and between 1900 and 1914 many more appeared. Reasonable though it was that authors who were unable themselves to handle their business with publishers satisfactorily should employ a professional to bargain for them, the higher rates of royalty and larger advance payments thus secured cut seriously into a publisher's profit. The increased cost made it considerably more difficult to finance the most speculative part of the business, the encouragement of new talent. The system of literary agents began in Britain but spread rapidly to the United States and also to the Continent, though in the latter it did not assume so great an importance. Keenly resented at first, the literary agent, by pressing for higher payment to authors he represented, may have been indirectly responsible for the greater selling efforts that some publishers began to make early in the century.

Sales methods. The discreet sales methods of the 19th century, whereby the sales representative merely showed his samples and the publisher took small spaces in newspapers for the bare announcement of title and author of his new books, were replaced by more forceful techniques. In this effort American publishers took a prominent part. Less hampered by inhibitions over the more blatant forms of salesmanship than their European colleagues, publishing houses in New York City began to take large advertisements, make extravagant claims for the qualities of their books, and thus build up bigger sales for new books than was customary in other countries. The existence of a prosperous middle class with fast-growing incomes was one factor; the vast spread of the population across the continent was another. These factors, combined with the development of the railroad, led to the successful development of mail-order advertising and selling. The sale of books, such as works of reference, by subscription was another technique that rapidly developed and grew into a business worth millions of dollars in the United States and elsewhere. It involved securing an undertaking to buy on installments over many months an already published set of books; it could also be used to secure advance orders for an expensive work, probably in several volumes, that the publisher was planning to issue, as was sometimes done in the 18th century. Continental countries also exploited the method, and considerable use was made of the door-to-door canvasser.

Installment selling of books

Effects of World War I. The coming of war in 1914 naturally had a disrupting, though not wholly destructive, effect upon book publishing in European countries. Shortage of paper necessitated rationing to two-thirds of prewar consumption in the case of Britain, while from hundreds of thousands of those in the armed forces came a tremendous demand for light reading. Although at one time the cost of paper rose to eight times its prewar level, sales of books increased sharply. The extra quantities could be supplied only at the expense of quality, and the standards of paper and binding were appalling. It would have been disastrous for a publisher to be left with large stocks of these books since paper supplies quickly returned to normal after the war, and the poorly produced books became unsalable. Of continental countries, Germany suffered the worst shortages, though the principal publishers were able to stay in business; in many respects a worse ordeal awaited them in the postwar inflation. In Britain, there was reluctance to recognize books as of any special importance to the national effort; virtually no direct use of them was made by the government, and it was not until the last four months of the war that a small proportion of publishers' staffs were granted any relief from compulsory national service.

An immediate aftereffect of the war in Europe was a sharp reduction in the purchasing power of the middle class. Whereas before, in most European countries, a proportion of the educated and professional classes bought new books regularly, high taxation, inflation, and trade depression in the postwar years cut down on spare money. Those publishers who continued to cater only to that public found it increasingly difficult to trade profitably, and many went out of business or were absorbed into larger firms. In the United States, on the other hand, boom conditions in the postwar years produced a still more prosperous and enlarged middle class ready to absorb an increasing supply of books. The number of publishing houses grew; and more American authors, such as Sinclair Lewis and Ernest Hemingway, found a world market. British and continental publishers turned more readily than before to New York City in search of fresh talent. Universities also increased in number more rapidly in the United States than elsewhere, producing a larger demand for college textbooks. Publishing them became an immensely important part of the business for many U.S. firms, which in some cases depended upon their profitable college departments to finance other parts of their operation, such as the fiction side.

Loss of purchasing power by the middle class

The book club. A new development of vast potential at this time was the book club, an association of members who undertook to purchase, usually each month, a book selected for them by a committee, the advantage being that the book in question was supplied at a lower price than that at which it could be bought in a bookshop. The scheme, of which an early forerunner was the Swiss Co-operative Movement in about 1900, had obvious attractions for the part of the reading public that had no direct access to a bookseller. The pioneer Book-of-the-Month Club in America (1926) developed a membership that ran into hundreds of thousands, followed by The Literary Guild, its great rival, and specialized book clubs that covered a variety of special reader interests. These clubs were strongly opposed at first by both publishers and booksellers, who disliked the additional emphasis placed upon the potential best-seller, but they came to supply a genuine need. They also helped to offset the enormous amount of book borrowing from libraries. From the 1950s onward, however, their popularity was somewhat affected by the availability of inexpensive paperbound books sold in thousands of outlets outside the regular book channels.

Design standards. As noted above, machine production had lowered standards of design. The English designer William Morris and his Kelmscott Press, however, had begun to work for better typography and book design in the 1890s; and his example had led to the establishment of other private presses, such as The Doves Press and the Ashendene Press, which produced editions (usually limited) of exceptional beauty, printed on handmade paper. Though aimed essentially at the collector and issued at high prices, such books began to influence the more discerning publisher; and by the 1920s a few firms, such as Alfred Knopf in New York City, Chatto and Windus and Jonathan Cape in London, and the Insel Verlag in Leipzig, were seen to be far ahead of their competitors in their standards of design. With careful planning, skillful selection of typeface, and provision of layouts to guide the printer, more and more publishers managed to achieve typographically handsome books at a commercial price. These efforts were part of the Design in Industry movement, which sought to demonstrate that mass production need not preclude beauty. It should be noted, however, that responsibility for design was passing from the printer to the publisher; as the former, with the growth of his business, became more the industrialist and less the craftsman, the latter realized that he must himself take charge of this aspect of the book.

The Great Depression. The great trade slump that began in October 1929 brought a swift decline in the prosperity of American publishing. By 1931 British publishers could no longer depend upon selling a high proportion of their books to the United States, either in the form of physical copies or by way of a contract conceding the U.S. rights. Though the book trade of Europe proved a little more resilient than some other industries, it passed through a difficult period. Sales declined, profits were negligible, and there were many bankruptcies. Attempts were made to find new outlets for books and fresh ways to attract the public to them. In London an annual Book Exhibition was run by *The Sunday Times* from 1933 to 1938; and *The New York Times* tried a similar venture in its city. The Germans continued to hold their annual

Book Fair in Leipzig, but this was primarily a trade function. Some British newspapers, striving for higher circulation, approached publishers to supply them with huge numbers of their popular books, specially printed, to be given away or sold very cheaply in exchange for coupons from the papers. Booksellers resented the practice, but for hard-pressed publishers it was financially attractive. In the rather desperate climate of the times, some publishers also spent inordinate amounts on newspaper advertising. Reprint book clubs proliferated too, again to the benefit of the few publishers and authors fortunate enough to secure a choice. In 1932 a valuable innovation that stimulated sales was the Book Token, a form of gift certificate. The invention of an English publisher, Harold Raymond, the Book Token could be exchanged for a book of specified value at any participating bookshop. It was at first opposed by many booksellers; but it went on to become a major factor in Christmas sales, and the system was adopted in other countries and by other trades.

Appearance of paperbacks

Even in the depressed conditions, publishers still dreamed of tapping a wider readership. This began to become a reality in 1935, when Allen Lane launched his pioneer Penguin series of paperbacks. It was a risky operation, involving speculatively high initial printings to keep down the unit cost. But, despite the strongly held belief that paperbacks would not appeal outside the Continent, where they had sold freely, and the resistance of booksellers, who feared a sharp reduction in their receipts, the new series quickly caught on. They represented remarkable value at the original price of sixpence, equivalent to the cost of a small item in a variety store. Though printed on cheap paper, the books employed good typography—far superior to that of any earlier attempts at paperbacks—and the original cover design was attractive in the bold simplicity of its orange and white stripes. A U.S. agency was arranged shortly before World War II and was later taken over by Victor Weybright, who subsequently established the highly successful New American Library for the mass promotion of paperbacks in the world market.

Nazi persecution of the Jews in the immediate prewar years and the impact of the war itself caused a wave of emigration, from Germany and Austria in particular, which brought fresh publishing talent to both Britain and the United States as well as to other countries, including Australia. Some of the striking developments in the production of art books, with beautiful coloured illustrations, were a direct result of this movement, which bore its fullest fruit after the war.

Effects of World War II. The war that in 1939 European publishers had feared would utterly destroy their business proved in many respects less terrible in its effects on books than had been imagined. While the destruction of buildings, plants, and vast stocks of books, most notably in London and later in Leipzig, brought publishing to a standstill for individual firms, the activity as a whole continued. As in 1914 but to an even greater extent, the demand for reading matter for both instruction and entertainment grew enormously. The nature of the war, with its long periods of waiting alternating with intense bouts of frenzied activity, both induced the need and provided the opportunity for reading. As a result, book sales in the "free" countries rose to fresh heights. The occupied countries of Europe endured censorship and a tight control of materials; but most publishers survived and were swift to renew contacts with colleagues in London and New York City immediately after the war.

Wartime paper shortages

In the United States, though they were subject to some shortages and inconvenience, publishers were comparatively untouched by the war, and their business expanded rapidly. In Britain, however, because of the acute pressure on shipping, the importation of esparto grass, an essential ingredient for good book papers, was strictly limited, and a publisher's paper ration was reduced to 37½ percent of his prewar annual consumption. By closer setting of type and the use of much thinner paper, the ration was stretched to produce the maximum number of copies, but the final appearance of British books inevitably suffered, and they began to compare very unfavourably with those produced in the United States.

In countries that suffered severe paper shortage there was, of course, a sharp reduction in the number of new books and in the size of editions; consequently, with the increase in demand, the available books were rapidly sold out. The result was an enormous, if illusory, increase of profitability for publishers; and despite heavy wartime taxation they found themselves in far better shape financially than ever before. Instead of holding large and often very slow-selling stocks with insufficient cash resources, publishers had little stock but ample cash. There was, too, the marginal advantage that those new authors who were able to secure publication in the war years could be virtually certain that their books would be quickly sold out. In these artificial conditions, many publishers were more prepared to risk the work of an untried author. Against this, however, was the very serious shortage of standard works of every kind, including classics and educational and reference books; at one time the cry went up that "Shakespeare is out of print!" While a small extra tonnage of paper was released in Britain in 1942 for the reprinting of books that were considered "nationally important" in wartime, no one could possibly pretend that there was not a real book famine in most European countries. After the war it took about five years for paper to become reasonably plentiful again. Despite the disruption brought by the war, however, interest in books had increased enormously, and sales were furthered by the total disappearance or severe rationing in most of the warring countries of so many consumer articles that normally compete with books. Contrary to the fears of many publishers, a new reading public was emerging, and it was not lost in the postwar world.

The postwar period. After the end of the war, there was an awkward year or so of reorganization and anticlimax, when many wartime publications suddenly became unsalable; but then publishing, in almost every country, once more expanded rapidly. People who had been cut off entirely from the rest of the world displayed an immense hunger for the books that had appeared during the previous six years. Much new business developed in the sale of the actual books and in translation rights. Such conditions continued at a higher level than they had attained in the 1930s, and they were to be further stimulated with the rise of the Frankfurt Book Fair. Social change came to many countries, bringing a broader spread of purchasing power and above all wider educational opportunity for much of the population. The change was to set book publishing upon a bolder and more adventurous course, turning it from a minor industry into one of sufficient growth and profitability to attract professional investors.

Revival of German publishing

A feature of the early postwar years was the remarkable phoenixlike rise of the German book trade, literally from the ashes of the Allied air raids, which had destroyed the principal cities with their publishing offices and printing works. Because Leipzig was in the Soviet-controlled zone of Germany, however, the centre of the trade moved to Frankfurt for the first time since about 1650. As part of its drive to become the commercial capital of West Germany, Frankfurt developed its exhibition facilities rapidly. Thus, the book trade fair had ideal conditions in which to thrive. Before 1939 it had been largely a domestic affair at which German publishers displayed their new works to booksellers, with only a small number of foreign publishers participating and those almost entirely continental; but it steadily grew to be the greatest meeting place for publishers from throughout the world.

In the nations that formed the Soviet bloc following World War II, publishing was subjected to a state control similar to that initiated in Soviet Russia in 1917. Very few of the famous publishing houses of Poland and Czechoslovakia survived, and the houses that did survive came under the ownership and control of the state. The normal pattern was for all books on a particular group of subjects to be issued from one publishing house. Thus in Hungary, for example, the principal houses dealt with science, political history, agriculture, music, belles lettres, or military or technical subjects. The organization in Romania was similar; but in East Germany it was significant that many of the prewar firms remained, though all were subject to government control.

Besides the economic and social changes that favoured publishing after 1945, an outburst of knowledge, particularly in science and technology, produced many new subjects, many of them highly specialized, all of which called for new books. The many new universities and colleges of technology that sprang up throughout the world formed a strong market for the thousands of college books that came to make up such a large part of many a publisher's list. At the same time, there was a major advance in printing, a break away from the traditional letterpress system dependent upon lead type. Photocomposition (composing of printed matter by photographic means rather than by hand), coupled with offset printing technique, obviated much of the handwork of the earlier methods, improved working speeds, and prevented costs from rising as steeply as they would otherwise have done. The trend was toward giant machines for mass production, giving a favourable price for cases in which 100,000 or more copies were needed. Such giant machines became essential for the printing of paperbacks, but the problem remained of printing economically those "short runs" of 3,000 or so in which the works of new authors, from whom many of the important books of the future must come, are normally tried out.

Rise of photocomposition

The paperback revolution. By the early 1950s the paperback revolution was well under way. Growing from the prewar Penguins and spreading to many other firms, paperbacks began to proliferate into well-printed, inexpensive books on every conceivable subject, including a wide range of first-class literature. Generally known as pocket books on the Continent, they swept the world, converting book borrowers into buyers and creating new book readers on a scale never known before. Their use has been particularly widespread in the developing countries, notably those of Africa. The new paperbacks had remarkable ubiquity, being found not only in bookshops but also in drugstores, street kiosks, and newsstands in railway stations, airports, and hotel lobbies. The low price of the paperback, which moved books for the first time into the area of impulse buying, is due essentially to the large number printed, seldom fewer than 30,000 and frequently far more, and not, as is often supposed, to the use of paper instead of a hard cover for the binding.

By far the greater number of paperbacks have been reprints of books that have had some success in their original clothbound form. Normally the paperback publisher makes an offer to buy the paperback rights from the publisher of the hardcover edition, and the paperback royalties are shared between the author and the hardcover publisher. While many of the big paperback houses have produced a certain number of new, hitherto unpublished books, the paperback operation is dependent in the main upon books originating with the conventional publishers. It is a fallacy therefore to suppose that, for all their seeming dominance, the paperback is likely to oust the hardcover book.

Scholarly paperbacks

Another type of paperback, selling in smaller numbers, has sprung from the enormous growth in the number of university students throughout the world. This is the reissue of works of scholarship, science, religion, literature, and art. Many had been out of print for years, and they had often been issued originally in small editions of no more than 2,000 copies by university presses or other specialized publishers. This great extension of the market began in the United States in the 1950s, with prices ranging from 65 cents to $1.95, at that time unusually high levels for paperbacks; the idea soon spread to Britain and the Continent. This operation has usually remained in the hands of the original publishers of the books, who have developed their own series of "university paperback books." It became customary for many new academic books to be issued simultaneously in both cloth (hardcover) editions and as paperbacks, the usual price of the latter being a little more than half that of the cloth edition.

University and government presses. The increase in the number of universities was accompanied by an increase in the number of university presses. The purpose of these presses is to serve the needs of scholarship—*i.e.*, to publish specialized material that a purely commercial firm would find impracticable to handle. Their freedom from the more acute profit-making pressures, often a result of direct subsidies, coupled with their assured, if limited, market, enables many to reach high standards of production and commercial viability. Some of the older establishments, such as the Oxford University Press, are, of course, large, profitable organizations with worldwide connections and a long list of more general publications.

Another type of publishing house not usually in direct competition with ordinary firms is the state printing office, which is responsible in many countries for issuing public and official material. In England, Her Majesty's Stationery Office, which was originally created in 1786 to coordinate office supplies for government departments, has come to issue a wide range of excellent books and pamphlets in connection with museums, galleries, and the advisory function of ministries, besides official papers. In the United States, the Government Printing Office in Washington, D.C., was established by Congress in 1860 for similar purposes; it too has steadily widened its field of operations. China has developed a similar organization to issue its publications.

Publishing practice. Every publishing house has manufacturing, marketing, and accounts departments, but the heart of the business lies in the editorial function. This has changed in its mode of operation through the years and still varies from one country to another and between firms but not in essentials. The editor—who is sometimes called the sponsor and who is often a director—selects the books to be published, deals with the author, and is responsible for the critical reading of the typescript (and its revision if necessary) and for seeing the book through the press, in consultation with the manufacturing and marketing departments. So vital can this role be that a particular editor's presence in a firm or transfer to another can be a major factor in attracting authors. Besides the editor, there is also an editorial department, which is responsible for the detailed preparation of the typescript before it is printed. This receives more attention today than in the past. Facts, figures, and references are checked, and inelegancies of style are polished where necessary. Careful attention by a skilled editor at this stage can contribute greatly to the quality of many books.

Role of the editor

Educational publishing. A particular branch of editorial work that has grown to be of cardinal importance since World War II concerns the conception, planning, and publication of the hundreds of books needed for educational programs at every level. Throughout the world editors specializing in school books visit teachers and lecturers to promote the writing of the required texts. The educational editor must concentrate almost wholly upon the commissioning of books to fit a particular syllabus in a school or university. Rarely, if ever, does the editor receive an unsolicited typescript that can be accepted at once. The editor must seek material by regular visits, either personally or by an assistant, to schools or colleges to find the teachers who have the makings of authorship. Outlines or drafts of texts are evaluated by editors who develop the central themes into a usable form. Much time must then be spent on revision and production before the book is completed.

In the United States the boards of education in some of the larger states review the available textbooks and approve a selection for use in their school districts. This selection process is called adoption, and publishers compete to have their books adopted for use because of the large volume of sales that are thus guaranteed. The schoolbook that is widely adopted may sell for a generation and reward author and publisher on a scale beyond the dreams of those concerned only with general books. Equally, nothing can fail so completely as the schoolbook that gets no adoptions.

Forms of copyright. Book publishing depends fundamentally on copyright, which is the sole right to copy or to produce a work, conceded to the publisher by the author through a mutual agreement. Without this element of monopoly, it would be impossible for a publisher to trade. It is also the guarantee for an author that he has legal rights to prevent the use of his material without fair com-

pensation. On the expiration of copyright, anyone is free to publish the work in question without payment to the author or his heirs. Copyright at one time was simple and indivisible; many alternative forms of text reproduction have developed, however. Their exploitation is governed by individual clauses in the agreement. These subsidiary rights may be briefly summarized. American rights for a British book and British rights for a book of American origin can prove to be exceptionally profitable. Though a book normally has its greatest sale in its country of origin, there are cases in which it does even better abroad. The richness of the American market gives it a particular attraction for publishers and authors of almost every other country. Translation rights have become a valuable source of additional revenue, particularly since the establishment of the Berne Convention.

Subsidiary rights

All the signatory countries agreed to copyright protection for the unpublished works of nationals of other member countries and for all the work first published in the Convention countries. While many books may earn no more than a few hundred dollars from the rights of translation in a single country, some world best-sellers, by authors of international stature, have a demand in almost every country, new or old, for a translation, and the aggregate earnings are then immense. Paperback rights for the more salable books, whether fiction or fact, are customarily offered to one of the major paperback houses, which flourish in most larger countries. For a best-seller there can be keen competition between the paperback houses, and advances well into seven figures may be offered to the original publisher, who normally controls the reprint rights. The original publisher also stipulates the earliest date at which the paperback may appear; as a rule, this is not less than 12 months after first publication. Rights for serial publication may be sold in several divisions: first serial rights, for which the best price can be obtained from a large-circulation newspaper or magazine in the capital city, may allow the publication of a number of installments appearing several weeks ahead of the issue of the book, or the serialization may "straddle" the appearance of the book, some installments before, the rest after. Second serial rights, for which much less is paid, can still yield useful sums: after first serialization has taken place, lesser papers in other parts of the country, or in other countries where the same language is spoken, can use the book. Digest rights, and their allied condensed book rights, represent another lucrative subsidiary use for books of wide general appeal.

Book club rights are also among those the publisher can exploit; the fees received from the clubs are also shared with the author. Broadcast and television rights in books interest a publisher primarily for the possibility of bringing a book and its author to the attention of a large segment of the public, rather than for the amounts paid. As a rule, there must be direct quotation from the text if a broadcasting company is to pay anything to the publisher. A television interview with the author, including sight of a copy of the book, is of great publicity value, and the author may even receive a fee for the appearance, but this is not part of the book's earnings. If the author can show a film relating to the book, it would be paid for at the appropriate rates for television use. In radio broadcasting, the reading of a book as a serial is one most remunerative possibility; the other is its full dramatization as a serial. The latter is, of course, still more valuable on television. Such use of new books has become more frequent; in the past this treatment was more often accorded to works of classic status. Dramatic and film rights can have importance for fiction, biography, and other general books, but only a small fraction of 1 percent of those published can be exploited by these means. From the publisher's standpoint, it is reasonable to share in the proceeds from the sale of these rights, for they result from the publisher's efforts.

Broadcast and television rights

Remote copying

The last group of subsidiary rights, rights for mechanical reproduction by film micrography, xerography, tape or disc recording, or any other technique of sight or sound, are of increasing concern to publishers. Dry-copying machines, easily operated, are to be found throughout the world in public, university, and school libraries, and while ordinarily only single copies can legally be made solely for the purposes of private study, it is a simple matter, though illegal, to run off a number of copies of long extracts, which then make it unnecessary to buy more than one copy of the book. Similarly, microfilm enables a single copy to satisfy many users and reduces the number of copies of the book that must be kept available in a library. Wherever material originates in the form of a book, however, the publisher must retain an interest in all forms of reproduction as part of his resources for promoting experimental and imaginative work.

Publisher's agreement. A publisher's agreement with an author normally specifies that in consideration of certain payments the former shall, during the legal term of copyright, have the exclusive right to produce or reproduce the said work in any material legible form throughout the world. In many cases, however, this agreement is modified to exclude some of the subsidiary rights named above, depending on the bargaining power of the author or his agent. After clauses specifying the extent of the rights conferred, the basic clause of a royalty agreement is that which states the rate of royalty to be paid. A typical wording is as follows: "On all copies of the said work sold on the normal terms a royalty of ten percent shall be paid on the published price rising to twelve percent after the sale of 5,000 copies and to fifteen percent after 10,000 copies." Other clauses provide for somewhat lower royalty rates on export sales and on cheap editions, on which the publisher's margin of profit is considerably less. Provision is also made for division between author and publisher of any payments received for such subsidiary rights as are included in the agreement. A publisher can fairly claim a share in them if they arise from the fact of book publication. Proofreading is another important matter covered by the agreement, the author being responsible for this. If the cost of making his corrections exceeds a stated figure he must pay for the excess. Lastly, in the majority of publishing agreements there is an option clause under which the author undertakes to give the publisher the first offer of his "next literary work suitable for publication in book form," usually with the addition that if, after a stipulated time, no terms shall have been agreed on for its publication the author is free to submit it elsewhere. The exact form of the legal instrument varies in detail; it is possibly drawn up in the greatest detail by U.S. firms because of the complexities of their system of selling: *e.g.*, by mail order, subscription, and similar means, in which the publisher must incur abnormal costs in order to secure the business. The vital condition for this publisher–author relationship, in the past often conducted with complete informality, is that there must be a legal document, a contract, setting out the rights and obligations of the two parties.

Literary agents and scouts. Literary agents have become increasingly important and prominent as publishing has grown more complex. A high proportion of the more successful authors of novels and general books now employ literary agents to place their books with publishers and to handle negotiations with them, the author being charged a commission of 10 percent. Besides negotiating and drawing up the contract with the firm, the good agency is equipped to handle the many subsidiary rights. Because an important element in the agent's value to an author is his capacity to extract better terms than the author would himself, it is not surprising that publishers have resented the agent's intrusion into the personal, and often very friendly, relationships between themselves and their authors. There can be no doubt, however, that agents do perform a valuable service in relieving an author of the considerable amount of routine work that his literary affairs may involve. Advice on possible new books to be written and occasionally, for the author of exceptional promise, an advance on anticipated earnings are also part of the assistance that the agent may offer. It must be emphasized, however, that agents are interested mainly in general books; they are seldom equipped to handle specialized and technical works.

Another publishing auxiliary who became significant in the 1950s and 1960s is the literary scout. Though a few had been employed earlier, mainly by U.S. publishers, who

Literary scouts

had their "lookouts" in one or two European cities, the practice is now more widespread. Many European publishers employ residents in London, Paris, and New York City to alert them at once to any promising new book, either written or just published. The scout, who may be connected with a newspaper or literary agency, is usually paid some modest amount as a retainer, probably with a commission of 1 or 2 percent on the published price of the books he recommends, in effect a small royalty on sales. On occasion a valuable find can be quite lucrative to the scout; frequently everything depends upon the speed with which a copy of the work can be got into the hands of his principal.

Selling and promotion. The publisher's techniques for book promotion have become increasingly sophisticated in all advanced countries. The typical traveler or book salesman is likely to hold a college degree, certainly in the United States; he receives a careful briefing from the home office, with elaborate samples and sales aids, and perhaps a car provided, or partly provided, by the firm. The itinerary for calls on bookshops (or in the case of the educational representative, schools and colleges) is prescribed by a supervisor, who usually checks the resulting orders against a quota. A well-run publishing house issues two or three seasonal announcement lists with details of its forthcoming books, as well as an annual catalog of its present and past books still in print, which are sent to the principal booksellers and librarians. For many books, a prospectus may be issued, both for the use of booksellers and for direct mailing by the publisher. The distribution of review copies to the press is the last item in the normal program. These three steps, traveling, catalogs, and reviews, are the vital elements in the machinery of book distribution, which it is virtually impossible to accomplish without the professional work of a publisher. The capacity of some authors to produce a quite presentable book with the help of a printer still leaves them far from their objective unless they can find a publisher to undertake its distribution.

Book advertising

Newspaper and periodical advertising is the publisher's principal means of reaching the public, and standards here have also risen considerably since World War II. Originally handled entirely by the publisher's own staff, it is now not uncommon for the larger houses, especially in the United States and in some European countries, to employ advertising agencies to prepare the copy and the general details of the campaign for any important book. While few authors consider that their books are advertised adequately and most publishers are highly doubtful whether press advertising does in fact sell books, the amounts spent in relation to sales revenue are much higher than for most other commodities, seldom less than 5 percent for new books. Without their receipts from publishers' advertising, some periodicals would find it impossible to devote so much space to book reviews, which are in themselves a most valuable aid to sales. The news value of many new books also enables them to secure free publicity through references in the general, as distinct from the literary, pages of a newspaper. A publisher with imagination, or the firm's press officer if there is one, can often suggest aspects of a book susceptible to such treatment. Broadcasting and television services, too, can sometimes be interested in books and their authors, and the resultant publicity may then be extremely effective.

Over the whole field of sales promotion, as publishing houses have grown in size and profitability, there has been a marked tendency for the more commercial methods of general business to be applied to books, which are aggressively promoted to retailers and the public in the same manner as are many other commodities. Though this may increase sales, at least in the short term, it may be doubted whether it is in the interests of the public and to the long-term advantage of good publishing. (P.U./G.U./Ed.)

Newspaper publishing

"A community needs news," said the British author Dame Rebecca West, "for the same reason that a man needs eyes. It has to see where it is going." For William Randolph Hearst, one of America's most important newspaper publishers, news was "what someone wants to stop you printing; all the rest is ads." Both idealistic and mercenary motives have contributed to the development of modern newspapers, which continue to attract millions of regular readers throughout the world despite stern competition from radio and television. Modern electronics, which has put a television set in almost every home in the Western world, has also revolutionized the newspaper publishing process, allowing many more newspapers to be born. An increasing number of these new newspapers are given away free, their production costs being borne entirely by the revenue from advertisements, which are of much greater importance than they were in Hearst's day.

Newspapers can be published daily or weekly, in the morning or in the afternoon; they may be published for the few hundred inhabitants of a small town, for a whole country, or even for an international market. A newspaper differs from other types of publication by its immediacy, characteristic headlines, and coverage of a miscellany of topical issues and events. According to the Royal Commission on the Press in Great Britain, to qualify as news "an event must first be interesting to the public, and the public for this purpose means for each paper the people who read that paper." But the importance of newspapers stretches far beyond a passing human interest in events. In the 19th century the first independent newspapers contributed significantly to the spread of literacy and of the concepts of human rights and democratic freedoms. Newspapers continue to shape opinions in the "global village" of the late 20th century, where international preoccupations are frequently of concern to the individual, and where individual tragedies are often played out on an international stage. Since it is commonly held that individuals have a right to know enough about what is happening to be able to participate in public life, the newspaper journalist is deemed to have a duty to inform. Whenever this public right to know comes under attack, a heavy responsibility falls on the journalist.

ORIGINS AND EARLY EVIDENCES

The daily newspaper is essentially the product of an industrialized society. In its independent form, the newspaper is usually integral to the development of democracy. The newspaper thus defined was fairly late in emerging, since it depended on a certain basic freedom of speech and relatively widespread literacy.

The Roman Empire. The urge to inform the public of official developments and pronouncements has been a characteristic of most autocratic rulers. This urge was fulfilled in ancient Rome by the *Acta Diurna* ("Daily Events"), a daily gazette dating from 59 BC and attributed in origin to Julius Caesar. Handwritten copies of this early journal were posted in prominent places in Rome and in the provinces with the clear intention of feeding the populace official information. The *Acta Diurna* was not, however, restricted to proclamations, edicts, or even to political decisions taken in the Roman Senate, the actions of which were reported separately in the *Acta Senatus* (literally "Proceedings of the Senate"). The typical *Acta Diurna* might contain news of gladiatorial contests, astrological omens, notable marriages, births and deaths, public appointments, and trials and executions. Such reading matter complemented the usual fare of military news and plebiscite results also given in the *Acta Diurna* and presaged the future popularity of such newspaper fillers as horoscopes, the obituary column, and the sports pages.

China. If the *Acta Diurna* was the forerunner of the modern newspaper in terms of content, it was, nevertheless, a government publication: the authorities decided what qualified as news for public consumption. The same applied to the regular *pao*, or reports of court affairs, circulated among the educated civil servants of Peking for more than a thousand years (AD 618–1911). The *pao* changed in format and title under the various dynasties, and technological change brought a shift from hand copying to printing from wooden type in the 17th century, but the durability of the *pao* was a testament to the stability of the civil servant class.

Medieval Europe. In Europe, the impetus for regular publication of news was lacking for several centuries after the breakup of the Roman Empire. The increased output of books and pamphlets made possible by the development of the printing press in the 16th century did not include any newspapers, properly defined. The nearest form was the newssheet, which was not printed but handwritten by official scribes and read aloud by town criers. News was also contained in the newsbook, or news pamphlet, which flourished in the 16th century as a means of disseminating information on particular topics of interest. One such pamphlet, printed in England by Richard Fawkes, and dated September 1513, was a description of the Battle of Flodden Field. Titled *The Trew Encountre,* this four-leaved pamphlet gave an eyewitness account of the battle together with a list of the English heroes involved. By the final decade of the 15th century, publication of newsbooks was running at more than 20 a year in England alone, matching a regular supply on the Continent. Authors and printers escaped official censorship or penalties by remaining anonymous or cultivating a certain obscurity, for it took a long time before the pamphlets came to the attention of the authorities. In any case the topics most frequently chosen for coverage—scandals, feats of heroism, or marvelous occurrences—were mainly nonpolitical and could not be regarded as a threat to the powerful. Governments in various countries were already in the vanguard of news publishing for propaganda purposes. The Venetian republic set a precedent by charging an admission fee of one *gazeta* (approximately three-fourths of a penny) to public readings of the latest news concerning the war with Turkey (1563), thus recognizing a commercial demand for news, even on the part of the illiterate. The term gazette was to become common among later newspapers sold commercially. Another popular title was to be Mercury (the messenger of the gods). The *Mercurius Gallobelgicus* (1588–1638) was among the earliest of a number of periodical summaries of the news that began to appear in Europe in the late 16th century. Newspaper names like *Mercury, Herald,* and *Express* have always been popular, suggesting the immediacy or freshness of the reading matter. Other names, such as *Observer, Guardian, Standard,* and *Argus* (a vigilant watcher), stress the social role played by newspapers in a democratic society.

Newspaper names

THE FIRST NEWSPAPERS

Newspaper development can be seen in three phases: first, the sporadic forerunners, gradually moving toward regular publication; second, more or less regular journals but liable to suppression and subject to censorship and licensing; and, third, a phase in which direct censorship is abandoned but attempts at control continue through taxation, bribery, and prosecution. Thereafter, some degree of independence has followed.

Commercial newsletters in continental Europe. The newsletter had been accepted as a conventional form of correspondence between officials or friends in Roman times, and in the late Middle Ages newsletters between the important trading families began to cross frontiers regularly. One family, the Fuggers, were owners of an important financial house in the German city of Augsburg; their regular newsletters were well-known even to outsiders. Traders' newsletters contained commercial information on the availability and prices of various goods and services, but they could include political news, just as the financial editor of today must consider the broader sweep of events likely to influence economic transactions. The commercial newsletter thus became the first vehicle of "serious" news, with its attempt at regular, frequent publication and concern with topical events generally.

The newsletter usually accorded primacy as a definite newspaper is the *Relation* of Strasbourg, first printed in 1609 by Johann Carolus. A close rival is the *Avisa Relation oder Zeitung* (*Zeitung* is the German word for "newspaper"), founded in the same year by Heinrich Julius, duke of Brunswick-Wolfenbuttel. In 1605, in the Low Countries, Abraham Verhoeven of Antwerp had begun publication of the *Nieuwe Tijdingen,* although the earliest surviving copy is dated 1621. In any case, this historical rivalry is evidence of a fairly sudden demand for newspapers at the start of the 17th century, and the continuous publication of the *Nieuwe Tijdingen* indicates that this demand soon became well-established. Although these publications were emerging throughout western Europe, it was the Dutch, with their advantageous geographical and trading position, who pioneered the international coverage of news through their "corantos," or "current news." The *Courante uyt Italien, Duytsland, & C.,* began to appear weekly or twice-weekly in 1618.

Similar rudimentary newspapers soon appeared in other European countries: Switzerland (1610), the Habsburg domains in central Europe (1620), England (1621), France (1631), Denmark (1634), Italy (1636), Sweden (1645), and Poland (1661). English and French translations of Dutch corantos were also available. But signs of official intolerance emerged fairly soon, and censorship stifled newspaper development in the late 17th century and into the 18th century in continental Europe. In Paris in 1631, the *Nouvelles Ordinaires de Divers Endroits,* a publishing venture by the booksellers Louis Vendosme and Jean Martin, was immediately replaced by an officially authorized publication, *La Gazette,* published under the name of Théophraste Renaudot but with influential backing by Cardinal de Richelieu. The new publication was to continue (as *La Gazette de France*) until 1917, casting the shadow of authority over nonofficial newspapers throughout its life. The first French daily—*Le Journal de Paris*—was not started until 1777; and although the Revolution of 1789 brought a temporary upsurge in newspaper publishing, with 350 papers being issued in Paris alone, the return to monarchy brought another clampdown. Napoleon I had his own official organ—*Le Moniteur Universel,* first published by Charles-Joseph Panckoucke (one of a family of booksellers and writers) in 1789 and lasting until 1869—and there were only three other French newspapers.

Censorship

In Germany, early newsletter development was soon hampered by the Thirty Years' War (1618–48), with its restrictions on trade, shortage of paper, and strict censorship. Even in peacetime censorship and parochialism inhibited the German press. Among the important regional newspapers were the *Augsburger Zeitung* (1689), the *Vossische Zeitung* in Berlin (1705), and the *Hamburgische Correspondent* (1714). In Austria the *Wiener Zeitung* was started in 1703 and is considered to be the oldest surviving daily newspaper in the world. The oldest continuously published weekly paper is the official Swedish gazette, the *Post-och inrikes tidningar,* begun in 1645. Sweden is also notable for having introduced the first law (in 1766) guaranteeing freedom of the press, but the concept of an independent press barely existed in most of Europe until the middle of the 19th century, and until then publishers were constantly subject to state authority.

Early newspapers in Britain and America. *Britain.* The British press made its debut—an inauspicious one—in the early 17th century. News coverage was restricted to foreign affairs for a long time, and even the first so-called English newspaper was a translation by Nathaniel Butter, a printer, of a Dutch coranto called *Corante, or newes from Italy, Germany, Hungarie, Spaine and France,* dated Sept. 24, 1621. Together with two London stationers, Nicholas Bourne and Thomas Archer, Butter published a stream of corantos and avisos, including a numbered and dated series of *Weekley Newes,* beginning in 1622. But a number of difficulties confronted a prospective publisher: a license to publish was needed; regular censorship of reporting was in operation from the earliest days; and foreign news no longer appeared because of a Star Chamber decree (in force from 1632 to 1638) completely banning the publication of accounts of the Thirty Years' War.

Between the abolition of the Star Chamber in 1641 and the establishment of the Commonwealth in 1649 publishers enjoyed a short spell of freedom from strict official control. Publication of domestic news began to appear more regularly, shedding the original book form. News and headlines increasingly replaced the old title page. The Civil Wars (1642–51) acted as a stimulus to reporters and publishers, and 300 distinct news publications were brought out between 1640 and 1660, although

many of these were only occasional reports from the battle front, such as *Truths from York* or *News from Hull.* The names of some contemporary publications, like the *Intelligencers, Scouts, Spys,* and *Posts,* reflected the bellicosity of the times, but the *Mercurys* still abounded, including the propaganda papers *Mercurius Academicus* (Royalist) and *Mercurius Britannicus* (Parliamentarian). The Parliamentarian victory brought strict control of the press from 1649 to 1658, and the restored monarchy was even more absolute, with the press being restricted to just two official papers. During the period of the Licensing Act (1662–94), an official surveyor of the press was given the sole privilege of publishing newspapers. The Revolution of 1688 produced a return to more permissive publishing laws and the first provincial presses were set up, starting with the *Worcester Post Man* (1690) and, in Scotland, the first *Edinburgh Gazette* (1699), although the British press was to remain principally a national one, centred on Fleet Street in London. Appearing briefly was *Lloyd's News* (1696), issuing from Edward Lloyd's coffeehouse, which had become a centre of marine insurance. The subsequent *Lloyd's List and Shipping Gazette* (from 1734), with its combination of general and shipping news, exemplified both the importance of the City of London's financial activities to the newspapers and the importance of a reliable and regular financial press to business.

Daily publication

In the early years of the 18th century the British newspaper was approaching its first stage of maturity. After 1691, improvements in the postal system made daily publication practical, the first attempt at doing so being the single-sheet *Daily Courant* (1702–35), which consisted largely of extracts from foreign corantos. A more radical departure was the triweekly *Review* (1704–13), produced by Daniel Defoe, in which the writer's opinion on current political topics was given, introducing the editorial, or leading article. Defoe had been imprisoned, in 1702, for his pamphlet "The Shortest Way with Dissenters," but many eminent British writers were being attracted to the newspapers. Henry Muddiman had gained eminence as the "journalist" who edited the *London Gazette* (from 1666). John Milton had edited the *Mercurius Politicus* under Oliver Cromwell, and Sir Richard Steele and Joseph Addison, *The Spectator* (published daily 1711–12). *The Spectator* and *The Tatler* (triweekly, 1709–11, also written by Steele) are commemorated in the modern magazines of the same name (see below *Magazine publishing*), but their incorporation of social and artistic news and comment influenced the content of the contemporary newspaper permanently. Sales of the popular *Spectator* sometimes ran as high as 3,000 copies, and already this circulation level was enough to attract advertising. An excise duty on advertisements was introduced by the Stamp Act (1712), along with other so-called taxes on knowledge aimed at curbing the nascent power of the press. The rate of duty, at one penny on a whole sheet (four sides of print), was the same as the cover price of *The Spectator,* and this effective doubling of the price killed it, along with many other newspapers. But the newspaper had already become a permanent part of the social and literary life in London, and not even higher duties could prevent the proliferation of newspaper titles throughout the century.

Typical of the new breed of English papers was *The Daily Advertiser* (1730–1807), which offered advertising space along with news of a political, commercial, and social nature. An important gap in the political pages was filled from 1771, when the right to publish proceedings in Parliament had been granted. This right was not won lightly, for illicit accounts of debates in the House had appeared in the monthly *Political State of Great Britain* (1711–40) and every effort had been made to stop them. But campaigners such as the political reformer John Wilkes (with the *North Briton,* 1762) eventually won out. Politicians of both Whig and Tory sympathies ran their own often scurrilous newspapers or simply bribed journalists with occasional handouts and annual stipends, but later in the century there emerged a more sophisticated reader who demanded, and received, an independent viewpoint. Eminent newspapers of the time included the *Morning Post* (1772), *The Times* (from 1788, but started as the *Daily Universal Register* in 1785), and *The Observer* (1791), each of which is still published (although the *Morning Post* was later merged with the *Daily Telegraph*). Censorship continued in the guise of frequent libel prosecutions, and as late as 1810 the radical political essayist William Cobbett was imprisoned and fined for denouncing flogging in the army, but the principle of a free press, at least in peacetime conditions, had been firmly established.

North America. In North America, publication of newspapers was deterred during colonial times by the long arm of the British law, but after independence the United States could boast one of the world's least restrictive sets of laws on publication. A first attempt at publishing, albeit abortive, was made in Boston by a radical from London, Benjamin Harris, in 1690. His *Publick Occurrences, Both Foreign and Domestick,* intended as a monthly series, was immediately stopped by the Governor of Massachusetts. It was clear that free speech and a nonofficial press were not to be tolerated in the colonies. Boston was also the site of the first official newspaper, the *Boston Newsletter* (1704), with which the authorities replaced the proclamations, pamphlets, and newsletters previously used to convey news from London. In 1719 the original title was replaced by the *Boston Gazette,* printed by Benjamin Franklin's elder brother, James, who soon produced the first independent American newspaper, the *New-England Courant* of 1721. William Bradford founded the first New York City newspaper, the *New-York Gazette,* in 1725, and his son Andrew was the first newspaper proprietor in Philadelphia. Further expansion of the colonies created 37 different titles by the outbreak of the War of Independence.

First North American newspaper

Colonial editors were aware of their responsibilities in creating a historical record of what was to be the new nation, and they cooperated in passing news to one another. In the absence of municipal offices, the printing office and newspaper headquarters often became a vital centre of community life. But frontier tensions led to passionate arguments, and newspapers became closely involved with political change. The Boston Tea Party (1773) itself is said to have been planned in a backroom of the *Boston Gazette,* already nicknamed "Monday's Dung Barge" by Loyalists. After independence the burning issues created with the new republic were aired in many new papers, most of which took up highly partisan stances. Thomas Jefferson and the first Republicans (later Democrats) were supported by the *Philadelphia Aurora* (1790), while Alexander Hamilton and the Federalists benefited from the support of the *Gazette of the United States* (1789–1818). Many city papers moved from weekly to daily publication, the first of these being the *Pennsylvania Evening Post* in 1783. The *Pennsylvania Packet* changed its name to *Pennsylvania Packet and Daily Advertiser* when it became a daily in 1784, indicating a new source of revenue for newspapers; and this was confirmed by the *New-York Daily Advertiser* (1785), the first to be published as a daily from the beginning.

The First Amendment to the U.S. Constitution specifically guaranteed "the freedom of speech or of the press." The right to criticize the government had been established as early as 1735, however, when John Peter Zenger, the publisher of the *New-York Weekly Journal,* was acquitted of criminal libel. After the temporary Alien and Sedition Acts (1798–1801), which included censorship clauses, were repealed, newspapers in the United States returned to polemics and public campaigns and set off on a course that was to help shape the modern character of the popular newspaper worldwide.

Early newspapers in Japan. A long tradition of news publication existed in Japan in the form of *yomiuri* ("sell and read," as the papers were sold by reading them aloud) or *kawara-ban* ("tile-block printing," the method of production). The *kawara-ban* broadsheets appeared continuously throughout the Tokugawa period (1603–1867), reporting popular festivals, personal scandals—notably the double suicides fashionable during the Genroku era (1608–1709)—natural disasters, and important events, such as the siege of Osaka Castle in 1615. Although much reporting concerned fairly innocuous occurrences, most writers preferred to remain anonymous for fear of the

punishments that could be imposed by the shogunate officials for unauthorized public discussion of political and social problems.

ERA OF THE INDUSTRIAL REVOLUTION

By 1800 educated citizens of most countries in Europe and in the United States could expect some access to independent news coverage and political comment, even if it was only to be found in clandestinely published newssheets. The basic formulas for serious newspapers and commercially successful, if sensational, popular newspapers had been worked out by shrewd writers and editors—members of the new profession of journalism. These formulas were to be elaborated throughout the 19th century, and by the end of the century the modern pattern of newspaper ownership and production had already been set in the United States and Britain, with newspapers passing from the realm of literature to that of big business.

Technological advances. New technology influenced newspapers both directly, through the revolution in printing techniques, and indirectly, through the rapid developments in transport and communications. In printing technology, necessity determined invention when the demand for newspapers exceeded the few thousand weekly copies required of the most popular titles. In 1814, the steam-driven "double-press" was introduced at *The Times* in London, allowing an output of 5,000 copies per hour. The higher output was a contributory factor in the rise of *The Times'* circulation from 5,000 to 50,000 by the middle of the century. The hand-operated wooden press used for books, newspapers, and single sheets alike was further pushed into obsolescence by the invention of mechanical lead type, the Fourdrinier machine (which produced cheap cellulose paper in rolls), curved printing plates, automatic ink-feeds, and, in 1865, the cylindrical rotary press.

Linotype

The main breakthrough, however, did not take place until the end of the century, with the introduction of automatic typesetting on Ottmar Mergenthaler's Linotype machine. Until then, each line of words to be printed had to be lined up and justified (made to fill exactly the allotted space between margins) by hand. After printing, the letters were replaced in alphabetical order by hand for reuse. The new machines were operated by a keyboard which selected a matrix for the correct letter from a channel in the magazine; the line of text was automatically justified (made to fill the line exactly by adjusting the space between words); the line of lead type was cast; and the matrices were automatically returned to the correct channels, thus saving the need for the lengthy process of manual distribution. The first Linotype machines were introduced at the *New York Tribune* in 1886 and in Britain at the *Newcastle Chronicle* in 1889. By 1895 every publisher in Fleet Street (then the centre of London newspaper publishing) was using the new machines. Linotype keyboard operators could set copy six times faster than the hand compositor. Electricity, introduced in 1884, was also a spur to the printing industry, as were machines that could not only print but could also cut, fold, and bind together newspapers of any size.

Newspaper production was also transformed by the speeding up of communication, which allowed news to be gathered instantly from distant cities via the telephone or even from foreign countries through the seabed cables laid between Dover, Eng., and Calais, Fr., in 1851 and across the Atlantic in 1866. In 1815, when the mounted courier was the chief means of getting news, it ordinarily took four days before news of an event as near as Brussels could be reported in London. The railway and other improvements in communications, such as the telegraph, revolutionized the newsman's conception of time and space. The railway networks not only moved reporters rapidly to and from their destinations but also helped to distribute newspapers, thus making them a more urgent and attractive commodity. Rapid and widespread delivery, especially in Britain and France, gave the larger newspapers based in capital cities a national status.

Foundations of modern journalism. The creation of new industrial occupations in society as a whole was reported by a new set of newspapermen who had far more specific jobs than their 18th-century predecessors. Earlier journalists might write, edit, and print each copy of the paper by themselves. With the expansion of newspapers, full-time reporters, whose job was to go and get the news, were recruited, and they replaced many occasional correspondents, although there was always room for the stringer, a part-time reporter based in a small town or a remote region. William Howard Russell, a reporter for the London *Times* during the Crimean War (1853–56), became famous as one of the first war correspondents, and his writings inspired Florence Nightingale to take up her mission to the Crimea. More than 150 war correspondents reported on the U.S. Civil War (1861–65). The reporter could become as celebrated as the soldier, and vigilant reporting could perhaps prevent some of the atrocities perpetrated in wartime. In peacetime the fearless on-the-spot reporter hoping to "scoop" rival papers for a big story also became a folk hero, and his byline (the name or nom de plume published with the article) could become better known than that of the editor.

The reporter

The expense of employing a large team of reporters, some of whom could be out of the office for months, proved impossible for smaller papers, thus paving the way for the news agency. The French businessman Charles Havas had begun this development in 1835 by turning a translation company into an agency offering the French press translated items from the chief European papers. His carrier-pigeon service between London, Paris, and Brussels followed, turning the company into an international concern that sold news items and that, eventually, also dealt in advertising space. Paul Julius Reuter, a former Havas employee, was among the first to exploit the new telegraphic cable lines in Germany, but his real success came in London, where he set up shop in 1851 as a supplier of overseas commercial information. Expansion soon led to the creation of the Reuters service of foreign telegrams to the press, an organization that grew with the spread of the British Empire to cover a large part of the world. In the United States, meanwhile, a very different type of agency—the newspaper cooperative—had arisen. Six New York City papers were the founding members; they suspended their traditional rivalries to share the cost of reporting the war with Mexico (1846–48) by establishing the New York Associated Press agency. Between 1870 and 1934, a series of agency treaties divided the world into exclusive territories for each major agency, but thereafter freedom of international operation was reinstated. The press agencies ensured a continuous supply of international "spot news"—*i.e.*, the bare facts about events as they occur—and raised standards of objective news reporting. For their feature pages, American newspaper editors came to rely on the feature syndicates, which supplied ready-to-use material from medical columns and book reviews to astrological forecasts and crossword puzzles.

News agencies

Growth of the newspaper business. Advances in newspaper production matched a quickening in the pace of life for the millions of people who read newspapers in the late 19th century. The railways, which transported newspapers rapidly from town to town, contributed to the breakdown of rural isolation, while the steamship and the telegraph brought nations closer together. Mass-produced newspapers with a broad appeal became available for the newly literate or semiliterate industrial worker. Circulations of some popular papers were climbing toward 1,000,000 by the end of the century, and newspaper publishing and advertising had become big business.

The United States. The movement toward a popular and politically independent press was spearheaded in the United States, where many potential readers were refugees from European political and religious persecution. The teeming immigrant population of New York City was the seedbed for several of the newspapers that were to shape the character of modern journalism. In 1835 the *New York Herald* was founded as the first American newspaper to proclaim and to maintain complete political independence. Its publisher, James Gordon Bennett, announced that the *Herald* would endeavour to record news, "with comments suitable, just, independent, fearless and good-tempered," while supporting no political party. The popularity of the *Herald*, with its exciting amalgam of news and

views presented in brief, easily digestible articles, was soon represented by a print run of more than 30,000 copies. The New Yorker's appetite for news was a substantial one, and in 1841 Horace Greeley introduced the *New York Tribune.* Whereas Bennett was an entertainer, Greeley was a campaigner, the first of the many idealists and crusaders who were to occupy American newspaper offices. "Go West, young man!" was a phrase coined in the *Tribune,* which also reflected its proprietor's fierce opposition to slavery and which influenced opinion well beyond the bounds of New York City. In the rough-and-ready frontier territories of the Midwest, crude sensationalism was a characteristic of the new popular press under editors such as Wilbur F. Storey of the *Chicago Times* (founded 1854), while painstaking investigation and exposure of political corruption was used by Rockhill Nelson of the *Kansas City Star* (1880) as new evidence of the independence of the press. In the South newspapers helped in rebuilding civic consciousness after the desolation of the Civil War through the efforts of men like Henry W. Grady at the *Atlanta Constitution* (after 1880) in Georgia and Henry Watterson at the *Louisville Courier-Journal* in Kentucky (after 1868).

The character of a newspaper could change radically under a new owner or editor. In New York City, an individual stamp was put on the influential *Evening Post* by its scholarly editor, Parke Godwin. The *New York Sun* had started life in 1833 as the first of the inexpensive popular papers known as the "penny press," with its founder, Benjamin H. Day, successfully exploiting a vein of demand for inconsequential "human-interest" stories. Later, under Charles A. Dana (after 1868), the *Sun* rose in style and prominence. The *New York Times* (1851), long in the shadow of the more vigorous *Herald* and *Tribune,* struck an important and lasting blow for the independence of the press by exposing an attempted bribe of the *Times'* editor by Tammany Hall politician William Marcy ("Boss") Tweed; the reported $5,000,000 sum offered and rejected was an ample indication of the growing power of the press.

Great Britain. In Europe, Britain alone could boast the presence of an independent press in the first half of the 19th century. The London *Times* demonstrated the value of journalistic objectivity and the need to criticize governments if hard-won rights were to be preserved. Under the consistent management of John Walter II and John Walter III, son and grandson of the founder, and with enlightened editorial control from outstanding journalists, such as Thomas Barnes and John Thaddeus Delane, *The Times* became a model for most serious British newspapers. In 1819 its reporting of the Peterloo Massacre by government troops at a political rally in Manchester was uncompromising; it campaigned for Parliamentary reform (achieved in 1832) and exposed the horrors of the Crimean War. From a technical standpoint *The Times* led the way in the introduction of advanced printing machinery and provided a fast and reliable news service as early as the Napoleonic Wars.

The Times

In 1836 the Stamp Tax was reduced to one penny, and in 1855 it was abolished entirely. This gradual relaxation of an impost on newspapers produced higher circulations for existing newspapers and encouraged the publication of new titles. Many were cheap, lurid crime sheets that disappeared as fast as they emerged. One exception was the sensational Sunday paper, the *News of the World* (1843), which attracted more readers than any other Sunday paper in Britain for more than a century. More characteristic of the age was the *Daily Telegraph* (1855), a penny paper, but one that competed directly with *The Times* by covering serious news stories and including thoughtful editorial comment on four sides of print, but at a quarter of the price of the fourpenny *Times.*

Continental Europe. During much of the 19th century, fear of popular insurgence led the European monarchies to keep a watchful eye on the newspaper presses. At the same time, prior to the unification of the modern states of Germany and Italy, newspapers covering national affairs were of limited interest. The first signs of a popular press appeared with the founding in Paris of *La Presse* (1836) by Émile de Girardin, who might be called one of the first press barons. He introduced new features and serials to raise circulation as high as 20,000 and thus to enable him to lower the price of his newspapers. A prominent contemporary of Girardin was Louis-Désiré Véron, who founded the *Revue de Paris* (1829) and revived the liberal daily *Le Constitutionnel* (1835). Aspiring French authors could gain publicity for their literary talents in these papers, especially when the Tanguy Law (1850) made it compulsory for them to sign the articles they wrote. But this literary slant to French newspapers, which persists to some degree in the modern era, could not disguise their paucity of hard news.

The French press

Disunity and political censorship continued to restrict the German press, although one independent daily, the *Allgemeine Zeitung* (Tübingen, 1798), managed to achieve wide influence. Farther north in Sweden, despite the freedom of speech granted to the press in 1766, the first notable newspaper (the *Aftonbladet,* founded by Lars Johan Hierta) was not begun until 1830.

Toward the middle of the century, censorship was abolished or relaxed in many other countries, including Switzerland (1848) and Denmark (1849). The new freedoms, together with the spread of literacy, gave birth to important newspapers, many of which still survive, including *Le Figaro* (Paris, 1854, daily from 1866), *Frankfurter Zeitung* (1856), *Le Peuple* (Brussels, date unknown), and the *Corriere della Sera* (Milan, 1876). In Spain and Portugal, censorship continued to prevent the development of true journalistic independence; any periods of comparative freedom were quickly followed by the reimposition of controls. In Russia strict censorship remained in force under the tsars, apart from a single decade (1855–65) of tolerance under Alexander II, when many new papers appeared. But limitations on publication were reimposed when it was found that greater freedom allowed radical ideas to be voiced, and the Russian press, like that in much of Europe, was forced to concentrate on literary rather than journalistic achievements.

Japan. The arrival of the U.S. naval officer Commodore Perry in Japan in 1853 was announced to the public in *kawara-bans,* which continued to be published for some years, though they began to be superseded by English-language newspapers. The first of these, the *Nagasaki Shipping List and Advertiser* (1861), was followed in the next five years by numerous periodicals, mainly translations produced by the shogunate Office for Reviewing Barbarian Papers. The office translated items from newspapers of China, Hong Kong, and the United States, as did Joseph Heko, a naturalized U.S. citizen and an interpreter at the American Embassy, in his monthly *Kaigai shimbun* (1865–66). The news items were therefore out of date, of little concern to the average Japanese, and too much resembling official announcements to be regarded as true newspapers. In 1867, however, the overthrow of the shogunate and the restoration of the Meiji led to the publication of more than a dozen newspapers concerned with domestic issues. Mainly issued by shogunate sympathizers, they included the *Koko shimbun,* whose publisher, the dramatist and educator Fukuchi Genichiro, had studied Western newspapers on his official travels abroad for the Japanese government (and who was later, in 1874, to preside over the *Nichi-Nichi shimbun,* which was closer to Western newspapers in style). The government soon suppressed these publications and promulgated the Newspaper Ordinance, which, in its 1871 version, decreed that the contents of a newspaper should always be "in the interest of governing the nation," a principle that was already anathema to many European and North American publishers.

Arrests of journalists and the supression of newspapers were common in the 1870s, but the decade nevertheless saw the birth of several giants of contemporary Japanese journalism. In 1870 the *Yokohama Mainichi,* the first daily in Japan, was started; it was also one of the first to use lead type. Two years later the *Tokyo Nichi-Nichi* appeared as one of the first truly modern Japanese newspapers, although it regarded itself as virtually an official gazette. The *Yomiuri shimbun,* one of the three leading national dailies in modern Japan, was founded in Tokyo

The first Japanese daily

in 1874, and it soon gained a reputation as a "literary" newspaper. The other two principal papers—the *Ōsaka Nippo* (1876) and the *Ōsaka Asahi* (1879)—were to become, respectively, the *Ōsaka Mainichi* and the *Asahi shimbun* (created in the early 1940s by a merger with the *Tokyo Asahi*, founded in 1888). They are associated with two of the fathers of modern Japanese newspaper publishing, Murayama Ryuhei (*Asahi*) and Motoyama Hikochi (*Mainichi*). Motoyama took full control of the *Mainichi* in 1903 and three years later added the *Tokyo Nichi-Nichi* to his publishing empire.

Other countries. In other parts of the world a familiar cycle took place, with prohibition or strict censorship gradually giving way to the demand for a free press, although colonial governments long exercised an especially tight control on political publications. Canada had its first newspapers as early as the 18th century. These developed regionally and catered to both English and French speakers in Montreal, Quebec, and Toronto. Fine standards of journalism were later set by the *Winnipeg Free Press* (1872).

Parts of India also had an early service, with newsletters being circulated from the 16th century. Under British rule, both English- and vernacular-language papers flourished—the latter under government control—and enviable standards were set by *The Times of India* (1838, formerly the *Bombay Times*) and *The Hindu* (1878).

The Australian press

Several Australian titles date to the early years of settlement, notable ones being the *Sydney Morning Herald* (1831), the Melbourne *Argus* (1846), and *The Age* (1854). Full censorship lasted until 1824 and the stamp tax until 1830, but one title (*The Sydney Gazette and New South Wales Advertiser*) was being published as early as 1803. The first issue of New Zealand's earliest newspaper, the *New Zealand Gazette,* was printed by emigrants even before their departure from London. The second issue awaited the installation of printing facilities in Wellington in 1840, when large-scale colonization was begun, but in the same year the *New Zealand Advertiser* was added to the list. The *Taranaki Herald* began publication in 1852.

In South Africa a press law was passed in 1828 to secure a modicum of publishing freedom, mainly through the efforts of the editor of the country's first paper, the *South African Commercial Advertiser.* Later papers, such as the *Cape Argus* (1857), were often tied to commercial and mining interests at first, but later their editors began to insist on freer commentary. The racially divided nature of the country has prevented, however, absolute freedom even in modern times. Similar restrictions continue to be applied to publishers in many other African and Asian countries, in eastern Europe, and in Latin America, although the political complexion of the various regimes may differ.

ERA OF THE POPULAR PRESS

The introduction of the Linotype machine in 1886 allowed production of the vast numbers of newspapers then in circulation in the industrialized countries, led as before by Britain and the United States. In the latter country new standards of sensationalism were set—and new sales records frequently announced—with the birth of the ruthless "yellow" journalism (an expression derived from a cartoon character called the "Yellow Kid," whose creator worked for the American newspaper publishers William Randolph Hearst and Joseph Pulitzer). In Britain the print runs of papers like *The Times* and the *Daily Telegraph* quickly reached the 100,000 mark in the second half of the 19th century. Newspapers were becoming part of mass-market industry, and in so doing they were shaking off many of their former ties with the literary world. This was evidenced in the revolutionary 1890s by the emergence of the "press baron," a businessman who owned chains of newspapers, by the increasing importance of advertising revenue, and by the use of unorthodox methods of winning more readers.

The United States. The number of American newspaper titles more than doubled between 1880 and 1900, from 850 to nearly 2,000. In addition to the weekly newspaper serving the smaller community, every major city had its own daily newspaper, and the metropolis had become the site of circulation battles between several titles. In New York City the newspaper business was shaken up by the arrival of Joseph Pulitzer, who is often credited with changing the course of American journalism. An immigrant from Hungary, Pulitzer had proved his ability in St. Louis, Mo., where he had bought and merged two local papers, the *Post* and the *Dispatch.* In New York City Pulitzer bought the failing *New York World* and in three years raised its circulation from 15,000 to 250,000, at that time the highest figure achieved by any newspaper in the world. With a series of stunts and campaigns, Pulitzer revitalized the established formulas of sensationalism and idealism, taking one step further the qualities of editorial independence and exciting journalism that had been introduced to an earlier generation of New Yorkers by Bennett's *Herald* and Greeley's *Tribune* (see above).

Joseph Pulitzer

Whereas Pulitzer was never afraid to unearth public wrongdoing and to crusade against it, the next press baron to influence New York City newspapers, William Randolph Hearst, was prepared to go to much further extremes in creating a headline story. Like Pulitzer, Hearst had learned about newspaper proprietorship in the brash, tough frontier West. His *San Francisco Examiner* (from 1880) had gained a reputation for exposing and cleaning up political corruption. By the time he came to New York City in 1895, however, Hearst was interested in circulation-building sensation at any price, even if it meant dressing up complete fabrications as news. This approach was revealed all too clearly in 1898, when Hearst's *Morning Journal* was challenging Pulitzer's *World* in the New York circulation battle. The *Journal* published exaggerated stories and editorials about the political tensions between the United States and Spain that stirred the country to a pitch of hysteria. Eventually, war—over Cuba—was triggered by the sinking of the U.S. battleship *Maine* in Havana harbour, but Hearst nevertheless claimed credit for the war in a banner headline: "How Do You Like the *Journal's* War?" Hearst is reported to have cabled his illustrator in Cuba, demanding pictures of atrocities for the *Journal.* The illustrator found no atrocities to illustrate and informed Hearst, who replied, "You furnish the pictures and I'll furnish the war." Scare headlines and attention-grabbing campaigns were only one of the tactics introduced by Hearst. Equally important in yellow journalism were a strong emphasis on the pictorial—photographs, cartoons, graphic illustrations—and the new Sunday supplements, which focused on human-interest stories and comic strips.

Hearst and the Spanish–American war

It was inevitable that some newspapers, and especially those that refrained from irresponsible tactics, would suffer circulation losses. One of these was *The New York Times,* which only recovered after its position as the city's leading serious journal was reestablished after the paper was taken over by the tycoon Adolph S. Ochs in 1896. The paper's slogans, "All the news that's fit to print" and "It will not soil the breakfast cloth," indicated its commitment to serious news reporting.

By 1900 there were half a dozen well-known newspaper barons in the United States. Hearst, whose collections at one time ran to 42 papers, was the most acquisitive of the early owners. Another early chain-builder was Edward Scripps, who began collecting newspapers in 1878. Scripps bought small, financially insecure newspapers and set them on their feet by installing capable young editors, who were given a share of the profits as an incentive to improve circulation. The editors were always urged "to serve that class of people and only that class of people from whom you cannot even hope to derive any other income than the one cent a day that they pay for your newspaper." Scripps wanted his papers to be of genuine service to the public, and though he succeeded in making money from them his motive was never exclusively profit. But the commercial advantage of owning newspaper chains soon became obvious, as it allowed newsprint to be bought on favourable terms and syndicated articles to be used to the fullest. Scripps's methods were adopted by his rivals and by newspaper proprietors in other countries as the idea of the chain spread. Inevitably, the profitable newspapers at-

Edward Scripps

tracted outside investors whose motives were commercial, not journalistic. This new type of proprietor was exemplified by Frank A. Munsey, who bought and merged many newspapers between 1916 and 1924, including the *Sun* and the *Herald* in New York City. In describing Munsey and others like him, the American author and editor William Allen White wrote that he possessed "the talent of a meat packer, the morals of a money changer, and the manners of an undertaker."

Commercial consolidation into larger publishing groups continued immediately after World War I, when the struggle for circulation intensified. First published in 1919, the *New York Daily News* was written to a ruthless recipe of sex and sensationalism by Joseph Medill Patterson, and it sparked off a war with Hearst's *Daily Mirror* and Bernarr Macfadden's *Daily Graphic,* both launched in 1924. The *Graphic* closed in 1932, and the *Daily News* survived to take over the *Mirror* in 1963 before itself becoming a part of the international media empire of the Australian-born Rupert Murdoch. Takeovers often led to title mergers or the complete disappearance of titles. In 1931 the *New York Morning, Evening,* and *Sunday World* titles were bought by the Scripps-Howard chain; the morning and Sunday editions were dropped, and the *Evening World* was merged with the *New York Evening Telegram,* an action that fitted in well with the popularity of afternoon papers in the United States. Newspapers with extensive circulations could command the attention of the larger advertisers, and this reinforced the disappearance of smaller titles in favour of a few high-circulation papers.

One outcome of the new ownership pattern was the gradual disappearance of the old press baron, who, as editor-proprietor, had tended to combine the roles of professional editor and management executive. Even the editor was to suffer a loss of personal impact as fame was increasingly won by columnists—men and women who were given regular columns to express forceful points of view or divulge society secrets. Among the most important political columnists of the 1920s were David Lawrence of the *United States News,* Frank Kent of the *Baltimore Sun,* Mark Sullivan of the *New York Herald-Tribune,* and Walter Lippmann of the *New York World.* Such writers could gain considerable national followings when their columns or articles were syndicated by major chains.

Great Britain. The British press was slower to emerge as a popular, sensational medium, but a major turning point came in 1855 when the stamp tax was abolished. This was preceded in 1853 by the abolition of the duty on advertisements, and the more liberal climate exposed a remarkable national appetite for newspapers of all kinds. The abolition of taxes and duties, including that on paper in 1861, brought down the prices of newspapers, and this alone was enough to create what were, for the time, very high circulations. By 1861 sales of the *Daily Telegraph* had risen to a daily average of 130,000, double that of *The Times.* Abolition of the tax on paper was said to have brought an additional £12,000 a year to the *Telegraph.* The *Telegraph*'s circulation had risen to more than 240,000 copies by 1877, then the highest in the world. The *Telegraph* could not be compared, however, with the more colourful New York papers. It was a worthy newspaper, more than half of it being taken up with reports of proceedings in Parliament, but its readers and those of *The Times* came almost exclusively from the growing mercantile middle class, for whom the two papers provided the writings of many of the best authors of the day at a comfortably affordable price. Journalistic independence was usually upheld, but as the party political hostility between Gladstone and Disraeli grew sharper, each paper became more partisan, a development that in turn stimulated sales.

Later in the century the British press began to adapt to the demand for less exacting reading matter. In 1888 the halfpenny evening *Star* was launched by the Irish nationalist politician T.P. O'Connor. Aiming at a wider public than any previous newspaper, the Star incorporated short, lively news items of human interest in a bold, attractive display. The new paper also gave good racing tips, thus endearing it to a group of men who have always contributed substantially to the circulation of what are known as the "populars." Another contemporary evening paper, the *Pall Mall Gazette,* adopted American tactics for some of its crusades. In a series of articles entitled "The Maiden Tribute to Modern Babylon," W.T. Stead exposed the prostitution of young girls in London by himself procuring one. (Indeed as a result he served a term in jail.) This early example of investigative journalism—in which the reporter creates hard news stories by investigating (sometimes clandestinely and by direct experience from inside) illegal or scandalous activities—led to the passing of the Criminal Law Amendment Act in 1885, which improved protection of minors. It also highlighted the power of the press to define what is unacceptable to society.

The first British press baron

At the turn of the century, popular journalism came into its own in Britain with the rise of Alfred Harmsworth (later Lord Northcliffe), who can be called the first of the British press barons both for his title and for his enduring influence on the press. During his lifetime he owned, at various times, the *Daily Mail, Mirror, The Times,* and the *Observer.* As his first effort he launched a cheap weekly magazine in 1888, when he was only 23. Using short sentences, short paragraphs, and short articles, the new style of editing was aimed at attracting a large following among those who had learned to read as a result of the 1870 Education Act making school compulsory for all British children. In 1894 Harmsworth bought the *Evening News,* and by combining his editing style with some of the methods of the American yellow press he quadrupled its circulation within a year. In 1896 came Harmsworth's main innovation, the *Daily Mail,* which within three years was selling more than 500,000 copies a day. This was more than twice the figure reached by any other paper thitherto. The *Daily Mail* went on to sell more than 1,000,000 copies a day during the Boer War (1899–1902).

As "A Penny Paper for One Halfpenny" the *Daily Mail* was sold to the reader at a low price only made possible by the paper's lucrative revenue from advertising. It was the first British paper to be based deliberately on advertising revenue rather than on sales revenue and the first to publish circulation figures audited independently by a chartered accountant. These figures gave advertisers evidence that the *Daily Mail* was reaching the public in sufficient numbers to warrant increasingly expensive advertising space. Another *Mail* slogan, "The Busy Man's Daily Journal," emphasized the snappy editorial style that followed the Harmsworth dictum of "Explain, simplify, clarify." This approach guided the new type of journalists known as subeditors, whose job was to rewrite stories in the "house" style, to compose headlines, and, if necessary, to add a little seasoning to the original story.

The tabloids

Another Harmsworth innovation was the tabloid newspaper, which was to revolutionize the popular press in the 20th century. The term tabloid was coined by Harmsworth when he designed and edited an experimental issue of the *New York World,* produced for New Year's Day, 1900. The tabloid halved the size of the newspaper page, which allowed easier handling by the reader, but it also suited the new, curtailed size of articles and the more numerous pages required per issue. In the long run, however, the term tabloid has come to define the popular newspaper more in style than in physical characteristics. The first successful tabloid was Harmsworth's *Daily Mirror* (1903). Originally launched as a newspaper for "gentlewomen," the *Mirror* had been a failure, but the tabloid format, together with a halfpenny cover price and numerous photographs, made the new picture paper an immediate success, with circulation running at more than 1,000,000 copies by 1914. Lord Northcliffe sold the *Mirror* to his brother Lord Rothermere in 1913. Meanwhile, the equally successful tabloid *Daily Sketch* had been begun in Manchester in 1909 by Sir Edward Hulton.

Like the American press barons, Northcliffe constantly intervened in the production of his newspapers, sending orders under his preferred epithet of "Chief" to the editors not only of the *Mail* and the *Mirror* but also of *The Times* (from 1908) and the *Observer* (from 1905), both of which he owned until his death in 1922. This was never more apparent than during World War I, when the British

Official Press Bureau was set up to control the amount of war information available to the public through the newspapers. Though accepting that a certain degree of censorship was necessary to conceal military intelligence from the enemy, Northcliffe nevertheless boldly defied the bureau over its cover-up of an ammunition shortage. Such defiance confirmed the independence of the press from government, but the influence of proprietors was itself to become an important issue in press freedom. This was typified after World War I by the intensive campaign for Empire Free Trade in Lord Beaverbrook's *Daily Express*. The preservation of the British Empire was the guiding passion of Max Aitken, who was raised to the peerage as Lord Beaverbrook in 1917. A Canadian-born journalist who took the *Express* into second place in national circulation behind the *Daily Mirror*, Beaverbrook continued to thrust his viewpoint on the editors of his papers for many years, although his campaigns for free trade within the empire and, after World War II, commonwealth trade preference were unsuccessful. Through the *Daily Express*, the *Sunday Express* (started in 1918), and the London *Evening Standard* (acquired 1923), Beaverbrook's opposition to Britain's attempts to join the European Community was given a regular airing. Beaverbrook admitted to the first Royal Commission on the Press that if an editor took a divergent view on, for example, the empire, he would be "talked out of it." So talented was Beaverbrook as a publisher and journalist that the *Express* newspapers gained and kept many readers for life, even though it is doubtful whether the issues of empire and Common Market membership were of passionate concern to them.

The circulation "war of the tabs" that struck New York City in the 1920s was copied in Britain in the 1930s, bringing with it numerous circulation-boosting stunts. Prizes for readers had been introduced as early as the 1890s, when Harmsworth offered a pound a week for life for the reader who could guess the value of gold in the Bank of England on a given day. In the 1920s one paper offered free insurance to subscribers, but this soon proved too costly to maintain. In 1930 the *Daily Herald* offered gifts to woo new readers. Although they were condemned by the Newspaper Proprietors' Association (now known as the Newspaper Publishers Association), gift schemes proliferated among other newspapers, with the *Herald* eventually achieving a circulation of 2,000,000, the highest in the world. Many of the new readers were stolen from other papers—the *Daily Mirror* saw its figure drop from more than 1,000,000 to 700,000 by 1934—but newspapers in general acquired 1,500,000 new readers, so that by the end of the decade there was a national newspaper aimed at every socioeconomic class. The *Daily Mirror* was revived by its editor, Harry Bartholomew, to become a true working-class paper with a radical political voice, although the winning of new readers—circulation eventually topped 4,000,000—was mostly due to the shameless use of the techniques of yellow journalism.

THE MODERN ERA

Since World War II, there have been radical changes in newspaper production on a par with those brought by the Industrial Revolution. Electronic technology has revolutionized the ways in which newspapers are written, edited, and printed, while radio and television have developed into serious competitors as sources of news, official information, and entertainment and as a vehicle for advertising.

Technological developments. Computers and telecommunications have transformed the production process for the modern newspaper. They have also led to changes in the quality of the newspaper itself, but their real impact has been on the finances of the newspaper industry and on the relevance of the traditional print workers. One of the first signs of technology's potential for change came in the 1930s, when Walter Morey developed the Teletypesetter (first demonstrated in 1928). This machine was an improvement on the telegraph, which was widely used by reporters in the field and by the wire services, such as Reuters and Associated Press, to send news items in draft form to editorial offices miles away. With the Teletypesetter, the impulses sent over the wire included encoded instructions to Linotype machines. The machines could then decode the instructions and automatically prepare whole pages ready for printing. It was therefore envisaged that the reporter would have the facility for "direct input" into the printing room, which would eliminate the need for retyping by a Linotype operator and thus save newspapers both time and money.

But direct input had to await the development of sophisticated computers and computer programs, which did not materialize until after World War II. In 1946 the first techniques of photocomposition were developed. With this method of typesetting, the images of pages are prepared for the printer photographically, as on a photocopier, instead of in lines of metal type. The new method was introduced gradually in newspapers, where the Linotype machines had worked well enough for more than half a century and where union opposition to the new technology was deeply entrenched. Technological advances were accelerated in the 1970s, introducing computers and computer programs that were tailor-made for the newspaper publisher, and many newspaper companies replaced their 19th-century printing systems with the new technology almost overnight. Photocomposition

In a modern newspaper office each journalist has a desktop terminal—*i.e.*, a keyboard and a visual display screen connected to the main computer. The visual display shows the current article or, in the case of a copy editor, the whole of the page being composed from various articles and pictures. While writing, the reporter can retrieve information stored in the computer, such as any previous articles on the same subject, which can be displayed on the screen alongside the new copy. This split-screen technology also allows the copy editor to move copy around the screen on special page-layout terminals until the copy fits the page. Once it is ready, a push of a button sends the complete page to the main computer for eventual transformation into an aluminum printing plate.

By this direct-input process the production of a page of news is accelerated. But the new technology can serve other production purposes. On some papers it is possible for an advertiser to send copy via the telephone to the newspaper office, where the computer automatically finds a suitable space for it and transmits it to the copy editor's screen. The reporter in the field, equipped with a portable terminal, can also input a story to the newspaper's computer directly and can gain access to the computer's library of information in the same way. If necessary, the editor can discuss the article with the reporter over the telephone as they both look at it on their screens. Similarly, items from press agencies can be located instantly; these may be transmitted to the computer terminal via cables or over the air by satellite, enabling news to reach the other side of the world within minutes. The electronic transmission of whole pages of news between remote locations also means that the printing plant does not have to be situated near the editorial offices. This can decrease real estate or rental costs, and it allows simultaneous editions of the same newspaper to be printed in different cities and even on different continents, an advantage that has been exploited by the British-based *Financial Times* and U.S.-based *Wall Street Journal*.

Financial developments. The introduction of new technology brought forth strong resistance from the unions of printing workers, which were traditionally among the most powerful labour unions. At first the operators of the obsolete Linotype machines were "brought upstairs" from the hot-metal shop to the newspaper offices, where they were retrained to compose copy on computer keyboards. But eventually even this function was no longer necessary as computers became more sophisticated, featuring word processing for journalists, graphics programs for illustrators, and editing programs designed specifically for newspaper editors. As the computer has taken over more and more of the basic functions of newspaper production, the proprietor has been able to cut down on highly skilled workers and take on a less qualified and lower-paid staff to do necessary jobs such as typing. New technology and the unions

Even before the introduction of the Linotype machine, however, the manual craft unions had been jealous

guardians of their trade as hand compositors. The compositors, together with the other craftsmen involved in printing, were well paid for their skills and for the night shifts they were obliged to work on morning papers. Overstaffing became common in newspaper printing departments when the unions laid down strict rules on the demarcation of labour (jobs that would be done only by particular employees), and working hours and conditions were precisely defined. The strike was used as a powerful weapon against the newspaper proprietor, since the loss of even one day's circulation might drive the reader to another paper. It was also feared that the regular reader might find that he did not miss his newspaper enough to start buying it again after the strike, especially when radio and television made news so readily available.

In the 1960s rising costs forced many newspapers to cease publication. The company owning the newspaper was often a conglomerate with various industrial interests. When this was the case, it was often not committed to maintaining an unprofitable title, regardless of the newspapers' history and tradition or its number of loyal readers. High circulation levels meant nothing if the paper did not attract advertising revenue and if inflated production costs prevented it from making any sort of profit. However, some well-known newspapers have been supported financially from the profits made by other parts of the conglomerate. Owning a venerable title can be a mark of prestige for a business enterprise, and there are still business tycoons who hope to emulate the press barons of earlier years—especially so in Britain, where a high percentage of newspaper proprietors have been raised to the peerage. It is not unusual, however, for prestigious papers to change ownership fairly frequently.

In Britain, for example, *The Times* was bought by a Canadian conglomerate, The Thomson Organisation, in 1966. Following a self-imposed closedown lasting nearly a year in 1979, Thomson sold *The Times* to News International, an Australian-based media company run by Rupert Murdoch, a businessman from a newspaper-owning family and who had a reputation for toughness. In a major attempt to break the power of the print unions, Murdoch moved *The Times* away from its headquarters near Fleet Street to new premises at Wapping in the London docklands, where he started to produce and print it using new technology. Another British newspaper proprietor, Eddie Shah, had already challenged the print unions in the provinces, and in 1986 he began *Today,* the first national British paper produced entirely on the new technology and without cooperation from the traditional print unions. *Today* used a production system similar to those of certain American papers and magazines. One of these, *USA Today,* was a success from its beginning in the early 1980s.

Newspaper popularity

Despite earlier fears of possible damage from competition by radio and, later, television, many newspapers have proved to be still attractive to consumers and hence profitable. Newspapers have, for instance, retained their importance as vehicles for advertising. The total amount of money that advertisers are willing to spend has continued to grow, which means that television has not greatly impinged on the value of the space offered by newspapers. In many countries radio and television are limited in the amount of advertising they may carry, which leaves a gap to be filled by newspapers and magazines. The press also offers classified space for the small advertiser, while regional and local papers are essential for small business wishing to advertise, especially in job recruitment, real estate, and automobile sales. One of the main developments of the 1970s and '80s has been the spread of free newspapers, or freesheets, which are delivered door-to-door or are distributed in public places. The free newspapers are usually printed by small newspaper enterprises using computerized technology and are entirely financed by advertising revenue. They pose a more serious threat to the existence of newspapers—particularly local ones—than do either radio or television. Because colour pictures are often essential to advertisers, many newspapers have come to offer occasional colour pages or, more commonly, to include colour magazines—Sunday supplements; by 1985 every British national Sunday newspaper included a free colour magazine. The costs of producing such a magazine are clearly worth the advertising revenue they generate.

Besides retaining their share of advertising, newspapers have had to compete for the attention of the consumer who can get the main points of the news from radio or television. Newspapers have done well to survive amid the proliferation of portable radios and radios in automobiles and of cable and satellite-broadcast television channels. The modern reader, in fact, is more likely to buy a newspaper to consult a special section than to read it from cover to cover. Readers may be attracted by the paper's sports reporting, racing tips, horoscopes, job advertisements, gossip columns, or, ironically, the daily listings of radio and television programming.

Another social change that has proved advantageous for newspapers is the increase in leisure time in developed countries. Shorter workweeks and longer vacations offer two opportunities for newspapers to exploit. First, there is simply more time to spend reading, encouraging sales to the individual of more than one newspaper or of a larger newspaper. Second, and of more importance for newspapers from the economic point of view, has been the growth of a diversity of leisure activities such as home improvement, gardening, and food and wine. Newspapers devote special features to these activities, particularly in their weekend editions. Foreign travel has also become more common, creating a demand for informative articles on popular tourist destinations. Even the sports pages, an essential part of the modern newspaper, have been affected by the changing leisure patterns—there has been an increase in the number of sports of general interest, allowing the expansion of the sports section to cover less popular sports. The economic advantage of covering more leisure activities and interests comes from the ability of newspapers to attract advertising revenue from commercial suppliers of leisure goods and services. The addition of new leisure pages does not normally lead to any reduction of editorial space, including the usual news and features, and so there is a tendency for newspapers to add more pages. Technological advances are allowing these larger newspapers to be printed. In expanding their coverage to include modern leisure interests, newspapers can be seen to continue to reflect the society of which they are a part.

Coverage of leisure activities

The role of the press. In the years leading up to World War II, newspapers had reflected the socially gloomy atmosphere engendered by economic and political restrictions. In Britain, *The Times* supported government policy, and only the *Daily Mirror* was prepared to investigate in depth European political, diplomatic, and economic events. The continental press reflected all too faithfully the unhappy state of affairs. In Germany, Italy, Spain, and Portugal the press was subject to the rigid censorship associated with Fascism, while in the Soviet Union a different political system induced the same controls. During the war, there was an expected clampdown on news coverage, although in the United States and Britain a greater flow of information was allowed than had been the case during World War I. The British Ministry of Information and the U.S. Office of War Information issued vast amounts of official news and propaganda. In the United States, the Office of Censorship defined a "Code of Wartime Practices for the American Press," while in Britain, newspapers voluntarily submitted any doubtful material for censorship, and an air of cooperative self-censorship thus prevailed. As in any modern war, journalist heroes were created, the standards of photography reached great heights, and the horror and tragedy of the war frequently inspired excellence in writing and editing.

In the developing countries of the Third World, newspapers have a vital role to play in disseminating a balanced picture of national affairs and in contributing to the growth of literacy. Repression of independent opinion is common in such countries, however. The freedom of the press is by no means universal even in the industrialized West.

Apart from constraints to do with defamation and national security, news blackouts or restrictions on information have occurred during military crises. Of even more concern is the growing number of threats made to journalists reporting from areas of political or military tension,

where at one time the press card gave the right to independent reporting. For more than half the world's population, an independent press is still an unattainable goal. The People's Republic of China holds the view that Western press freedom is illusory, because a wealthy minority controls what is to be printed, whereas access to the press is truly free in communist countries. Distortion of the truth can be said to arise wherever newspaper ownership approaches monopoly or even, as in some Western countries, if it is controlled by a small number of organizations. New technology does offer escape from this impasse, because it makes possible the commercial production of far more newspapers. Many of these new titles could publicize the views of minorities who at present frequently go unheard.

Above all, the serious newspaper has moved toward providing in-depth detail, analysis, and opinion on many current events. The quality of newspaper coverage of business affairs, the arts, and social issues is increasingly important as publishers deal with more sophisticated readers. The newspaper can still be a forum for thoughtful debate, a medium for creative expression, and a safeguard of the written language. (P.U./G.U./D.H.T.)

Magazine publishing

BEGINNINGS IN THE 17TH CENTURY

Though there may have been published material similar to a magazine in antiquity, especially perhaps in China, the magazine as it is now known began only after the invention of printing in the West. It had its roots in the spate of pamphlets, broadsides, ballads, chapbooks, and almanacs that printing made possible. Much of the energy that went into these gradually became channeled into publications that appeared regularly and collected a variety of material designed to appeal to particular interests. The magazine thus came to occupy the large middle ground, incapable of sharp definition, between the book and the newspaper.

The earliest magazine appears to have been the German *Erbauliche Monaths-Unterredungen* (1663–68; "Edifying Monthly Discussions"), started by Johann Rist, a theologian and poet of Hamburg. Soon after there appeared a group of learned periodicals: the *Journal des Sçavans* (later *Journal des Savants;* 1665), started in France by the author Denis de Sallo; the *Philosophical Transactions* (1665) of the Royal Society in England; and the *Giornale de' letterati* (1668), published in Italy and issued by the scholar and ecclesiastic Francesco Nazzari. A similar journal was started in Germany a little later, the *Acta eruditorum Lipsiensium* (Leipzig; 1682); and mention may also be made of the exile-French *Nouvelles de la République des Lettres* (1684), published by the philosopher Pierre Bayle mainly in Holland to escape censorship. These sprang from the revival of learning, the need to review its fruits, and the wish to diffuse its spirit as widely as possible.

The learned journals summarized important new books, but there were as yet no literary reviews. Book advertisements, by about 1650 a regular feature of the newssheets, sometimes had brief comments added, and regular catalogs began to appear, such as the English quarterly *Mercurius librarius, or A Catalogue of Books* (1668–70). But in the 17th century the only periodicals devoted to books were short-lived: the *Weekly Memorials for the Ingenious* (1682–83), which offered some critical notes on books, and the *Universal Historical Bibliothèque* (January–March 1686). The latter invited scholarly contributions and could thus be regarded as the true forerunner of the literary review.

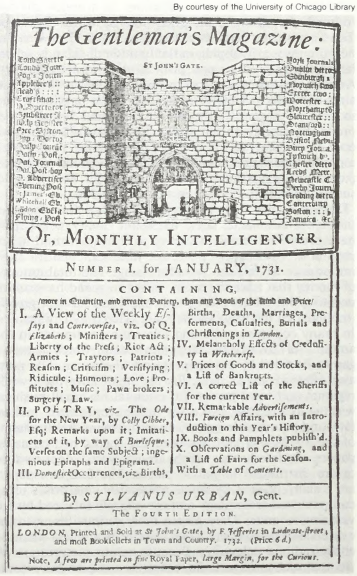
The lighter type of magazine, or "periodical of amusement," may be dated from 1672, which saw the first appearance of *Le Mercure Galant* (renamed *Mercure de France* in 1714). It was founded by the writer Jean Donneau de Vizé and contained court news, anecdotes, and short pieces of verse—a recipe that was to prove endlessly popular and become widely imitated. This was followed in 1688 by a German periodical with an unwieldy title but one that well expressed the intention behind many a subsequent magazine: "Entertaining and Serious, Rational and Unsophisticated Ideas on All Kinds of Agreeable and Useful Books and Subjects." It was issued in Leipzig by the jurist Christian Thomasius, who made a point of encouraging women readers. England was next in the field, with a penny weekly, the *Athenian Gazette* (better known later as the *Athenian Mercury;* 1690–97), run by a London publisher, John Dunton, to resolve "all the most Nice and Curious Questions." Soon after came the *Gentleman's Journal* (1692–94), started by the French-born Peter Anthony Motteux, with a monthly blend of news, prose, and poetry. In 1693, after devoting some experimental numbers of the *Athenian Mercury* to "the Fair Sex," Dunton brought out the first magazine specifically for women, the *Ladies' Mercury.* Finally, another note, taken up time and again later, was struck by *The London Spy* (1698–1700), issued by a tavern keeper, Ned Ward, and containing a running narrative of the sights and sounds of London.

DEVELOPMENTS IN THE 18TH CENTURY

Great Britain. With increasing literacy—especially among women—and a quickening interest in new ideas, the magazine filled out and became better established. In Britain, three early "essay periodicals" had enormous influence: Daniel Defoe's *The Review* (1704–13; thrice weekly); Sir Richard Steele's *The Tatler* (1709–11; thrice weekly), to which Joseph Addison soon contributed; and Addison and Steele's *The Spectator* (1711–12, briefly revived in 1714; daily). Though they resembled newspapers in the frequency of their appearance, they were more like magazines in content. *The Review* introduced the opinion-forming political article on domestic and foreign affairs, while the cultivated essays of *The Tatler* and *The Spectator,* designed "to enliven morality with wit, and to temper wit with morality," did much to shape the manners and taste of the age. The latter had countless imitators not only in Britain, where there were in addition the *Female Tatler* (1709–10) and the *Female Spectator* (1744–46), but also on the Continent and later in America. The Stamp Tax of 1712 had a damping effect, as intended, but mag-

The Spectator

By courtesy of the University of Chicago Library


The Gentleman's Magazine:

Or, Monthly Intelligencer.

Number I. for JANUARY, 1731.

CONTAINING,

I. A View of the Weekly Essays and Controversies, viz. Of Oaths; Ministers; Treaties; Liberty of the Press; Riot Act; Armies; Traytors; Patriots; Reason; Criticism; Versifying; Ridicule; Humours; Love; Prostitutes; Music; Fawn brokers; Surgery; Law.

II. POETRY, viz. The Ode for the New Year, by Colley Cibber, Esq; Remarks upon it; Imitations of it, by way of Burlesque; Verses on the same Subject; ingenious Epitaphs and Epigrams.

III. Domestick Occurrences, viz. Births, Deaths, Marriages, Preferments, Casualties, Burials and Christenings in London.

IV. Melancholy Effects of Credulity in Winchester.

V. Prices of Goods and Stocks, and a List of Bankrupts.

VI. A correct List of the Sheriffs for the current Year.

VII. Remarkable Advertisements.

VIII. Foreign Affairs, with an introduction to this Year's History.

IX. Books and Pamphlets publish'd.

X. Observations on Gardening, and a List of Fairs for the Season.

With a Table of Contents.

By SYLVANUS URBAN, Gent.

The FOURTH EDITION.

LONDON, Printed and Sold at St John's Gate, by F. Jefferies in Ludgate-street, and most Booksellers in Town and Country. 1732. (Price 6d.)

Note, A few are printed on fine Royal Paper, large Margin, for the Curious.

Front page of the first issue of *The Gentleman's Magazine* (bound volume, 4th ed.), 1731, the first general periodical in England, published until 1907.

azines proved endlessly resilient, easy to start and easy to fail, then as now.

So far various themes had been tried out; they were first brought together convincingly by the English printer Edward Cave, who began to publish *The Gentleman's Magazine* in 1731. It was originally a monthly collection of essays and articles culled from elsewhere, hence the term magazine—the first use of the word in this context. Cave was joined in 1738 by Dr. Johnson, who was later to publish his own *Rambler* (1750–52); thereafter *The Gentleman's Magazine* contained mostly original matter, including parliamentary reports. Rivals and imitators quickly followed, notably the *London Magazine* (1732–85) and the *Scots Magazine* (1739–1817; to 1826 published as the *Edinburgh Magazine*); and, among the increasing number of women's periodicals, there were a *Ladies' Magazine* (1749–53) and a *Lady's Magazine* (1770–1832). Their progenitor, however, outlived them all and perished only in 1907.

The literary and political rivalries of the day produced numerous short-lived periodicals, from which the critical review emerged as an established form. Robert Dodsley, a London publisher, started the *Museum* (1746–47), devoted mainly to books, and Ralph Griffiths, a Nonconformist bookseller, founded *The Monthly Review* (1749–1845), which had the novelist and poet Oliver Goldsmith as a contributor. To oppose the latter on behalf of the Tories and the Church of England, *The Critical Review* (1756–1817) was started by an Edinburgh printer, Archibald Hamilton, with the novelist Tobias Smollett as its first editor. Book reviews tended to be long and fulsome, with copious quotations; a more astringent note came in only with the founding of the *Edinburgh Review* in 1802 (see below).

Continental Europe. On the Continent development was similar but was hampered by censorship. French magazines containing new ideas had to appear in exile, such as the philosopher Pierre Bayle's *Nouvelles de la République des Lettres,* which was published largely in Holland; some 30 titles were published in Holland up to the time of the French Revolution. Within France, there were the short-lived *Spectateur Français* (1722–23) and *Spectateur Suisse* (1723); and *Le Pour et le Contre* (1733–40; "For and Against"), issued by the Abbé Prévost (author of *Manon Lescaut*). Of more literary interest were the *Gazette Littéraire de l'Europe* (1764–84) and *La Décade Philosophique, Littéraire et Politique* (1794–1804).

In Leipzig the poet and philosopher Johann Christoph Gottsched issued a periodical for women, *Die vernünftigen Tadlerinnen* (1725–26; "The Rational Woman-Critics"), and the first German literary review, *Beiträge zur kritischen Historie der deutschen Sprache* (1732–44; "Contributions to the History of the German Language"). German literary movements were connected with the production of new magazines to a greater extent than in Britain. Examples of such vehicles include Friedrich von Schiller's *Horen* (1795–97) and Johann Wolfgang von Goethe's *Propyläen* (1798–1800), the influence of which was often greater than their duration. Of more general and lasting influence was the *Allgemeine Literatur-zeitung* (1785–1849), founded by Friedrich Justin Bertuch, "the father of the German periodical."

The first Russian periodical, published by the Academy of Sciences, was a learned journal called "Monthly Works" (1755–64). The first privately published Russian magazine, a critical periodical with essays and translations from the British *Spectator,* was called "Industrious Bee" and began in 1759. Catherine II used her *Vsiakaia Vsiachina* (1769–70), also modeled on the *Spectator,* to attack opponents, among them Nikolay Novikov, whose "Drone" (1769–70) and "Windbag" (1770) were suspended and whose "Painter" (1770–72) escaped only by being dedicated to the Empress.

America. In America the first magazines were published in 1741. In that year appeared Andrew Bradford's *American Magazine,* the first publication of its kind in the colonies. It was joined, a mere three days later, by Benjamin Franklin's *General Magazine.* Both magazines appeared in Philadelphia; neither lasted very long, however—Bradford's magazine survived only three months and Franklin's six. Franklin was more widely known for another of his publications, *Poor Richard's Almanack* (1732–57), which contained maxims and proverbs. Before the end of the 18th century, some 100 magazines had appeared, offering miscellaneous entertainment, uplift, or information, mostly on a very shaky, local, and brief basis. Among the more important were, in Philadelphia, the *Pennsylvania Magazine* (1775–76), edited by Thomas Paine, and the *American Museum* (1787–92) of the bookseller Mathew Carey; the *Massachusetts Magazine* (1789–96), published in Boston; and the *New-York* (City) *Magazine* (1790–97).

THE 19TH CENTURY AND THE START OF MASS CIRCULATION

General periodicals. Most of the early periodicals were designed for the few who could afford them and can be fairly called "quality" magazines. In the 1830s, however, less expensive magazines, aimed at a wider public, began to appear. At first these magazines emphasized features that promoted improvement, enlightenment, and family entertainment, but, toward the end of the century, they evolved into popular versions that aimed at providing amusement.

The pioneers of the new type of magazine in Britain were Charles Knight, publisher for the Society for the Diffusion of Useful Knowledge, with his weekly *Penny Magazine* (1832–46) and *Penny Cyclopaedia* (1833–58); the Chambers brothers, William and Robert, with *Chambers's (Edinburgh) Journal* (1832–1956), which reached a circulation of 90,000 in 1845; and teetotaler John Cassell, with his *Working Man's Friend and Family Instructor* (1850–53) and the *Quiver* (1861). Besides popular magazines, many standard works appeared serially, often with illustrations. Typical of family entertainment were Charles Dickens' *Household Words* (1850), followed in 1859 by *All the Year Round*; several similar periodicals such as *Good Words* (1860); and, for young people, the *Boy's Own Paper* (1879) and the *Girl's Own Paper* (1880). Germany had its *Pfennigmagazin* (1833), edited by Johann Jakob Weber, and a family magazine modeled on that of Dickens. One example was the *Gartenlaube* (1853–1937; "Arbour"), which enjoyed great popular influence and a circulation of 400,000 in the 1870s. There were no national magazines in the United States before about 1850, but two of its best-known early periodicals were the *Saturday Evening Post* (1821–1969; revived 1971) and *Youth's Companion* (1827–1929). The latter, published in Boston, was typically wholesome in content, intended to "warn against the ways of transgression" and to encourage "virtue and piety."

Family magazines

By the last quarter of the century, largely as a result of compulsory education, the potential market for magazines had greatly increased, and the public was avid for miscellaneous information and light entertainment. The first man in Britain to discover this was George Newnes, who liked snipping out any paragraph that appealed to him. In 1881 he turned his hobby to advantage by publishing a penny magazine, *Tit-Bits from all the Most Interesting Books, Periodicals and Contributors in the World,* soon shortened to *Tit-Bits* (in 1968 restyled *Titbits*). It was a great success and formed the beginning of a publishing empire that was to include *Country Life* (founded 1897), *Wide World Magazine* (1898), and, above all, *The Strand Magazine* (1891–1950), one of the first monthly magazines of light literature with plenty of illustrations. *The Strand* became enormously popular and is perhaps most famous for its Sherlock Holmes stories by Arthur Conan Doyle. Among the early contributors to *Tit-Bits* was Alfred Harmsworth (later Lord Northcliffe), who had an appetite for odd bits of information similar to that of Newnes. In 1888, after editing *Youth* and *Bicycling News,* Harmsworth launched a rival to *Tit-Bits* called *Answers to Correspondents,* or *Answers,* which he successfully promoted by contests. Within five years he produced a string of inexpensive magazines for the same popular market, including *Comic Cuts* and *Home Chat.* A similar empire was built up by Arthur Pearson, another former *Tit-Bits* employee, with *Pearson's Weekly* and *Home Notes,* among others.

In the United States, magazine publishing boomed as part of the general expansion after the Civil War. It was also helped by favourable postal rates for periodicals (1879). But a gulf remained between expensive magazines aimed at the genteel, such as *Harper's* and *Scribner's* (see below *Literary and scientific magazines*), and cheaper weeklies and miscellanies. The first person to produce a popular monthly to fill this gap and thus spark off a revolution in the industry was Samuel Sidney McClure, who began publishing *McClure's Magazine* in 1893, which he sold for 15 cents an issue instead of the usual 25 or 35 cents. John Brisben Walker, who was building up *Cosmopolitan* (founded 1886) after acquiring it in 1889, cut his price to 12½ cents, and in October 1893 Frank A. Munsey reduced the price of *Munsey's Magazine* (1889–1929) to 10 cents. All three saw that, by keeping down the price and gearing contents to the interests and problems of the average reader, high circulations were attainable. Munsey estimated that, between 1893 and 1899, "the ten-cent magazine increased the magazine-buying public from 250,000 to 750,000 persons." This increase in circulation in turn led to high advertising revenue, making it possible to sell a magazine, like a newspaper, for less than its cost of production, a practice that was to become common in the next century. Technical development was also important; mass-production methods and the use of photoengraving processes for illustration enabled attractive magazines to be produced at ever lower unit costs.

Early popular monthly magazines in the United States

The first magazine published in Australia was the *Australian Magazine,* which began in 1821 and lasted for 13 monthly issues. The *South Asian Register* began as a quarterly in 1827 but only four issues appeared. The *Hobart Town Magazine* (1833–34) survived a bit longer and contained stories, poems, and essays by Australian writers. The *Sydney Literary News* (1837) was the first to contain serial fiction and advertisements. Illustrations were introduced in the 1840s; the *Australian Gold Digger's Monthly Magazine and Colonial Family Visitor* (1852–53) was followed by the *Melbourne Punch* (1855–1925; incorporated in *Table Talk,* 1885–1937).

In India the first magazines were published by the British. The earliest to appear was the *Oriental Magazine; or, Calcutta Amusement* (1785–86); it was followed by a number of short-lived missionary publications. The first periodical founded and edited by an Indian was the *Hindustan Review,* which commenced in 1900.

Magazines in China

Missionaries founded the first periodical in China; printed in Malacca, the *Chinese Monthly Magazine* lasted from 1815 to 1822. It was followed by the *East-West Monthly Magazine,* printed in Canton from 1833 to 1837 and in Singapore from 1837 until its end in 1847.

Illustrated magazines. The first man in Britain to notice the effect of illustrations on sales and grasp their possibilities was a newsagent in Nottingham, Herbert Ingram, who moved to London in 1842 and began publishing *The Illustrated London News,* a weekly consisting of 16 pages of letterpress and 32 woodcuts. It was successful from the start, winning the approval of the Archbishop of Canterbury and hence that of the clerical public. Though it suffered at first from the defect that its pictures were by well-known artists but were not taken from life, it later sent artists all over the world. Drawings made on the spot during the South African War, sometimes at considerable risk, were a great popular feature. Among its competitors was the monthly *English Illustrated Magazine* (1883–1913).

News in pictures

The idea of presenting the news largely in pictures was quickly taken up in France by *L'Illustration* (1843–1944) and in Germany by the *Leipziger illustrierte Zeitung* (1843) and *Die Woche* (1899–1940).

In the United States, the main early illustrated magazines were *Leslie's Weekly* (1855–1922) and *Harper's Weekly* (1857). Soon after its founding, *Leslie's* had a circulation of 100,000, which doubled or trebled whenever there was something sensational to portray. During the Civil War, of which it gave a good pictorial record, it had as many as 12 correspondents at the front.

The invention of photography and the development of the halftone block began to transform this type of magazine from the 1890s, with the artist increasingly being displaced by the camera.

Women's magazines. Women's magazines frequently reflect the changing view of women's role in society. In the 18th century, when women were expected to participate in social and political life, those magazines aimed primarily at women were relatively robust and stimulating in content; in the 19th, when domesticity became the ideal, they were inclined to be insipid and humourless. After about 1880, magazines began to widen their horizons again.

Typical of the late Georgian and Regency magazines in Britain were *The Lady's Magazine* (1770), a sixpenny monthly that, along with its literary contributions and fashion notes, gave away embroidery patterns and sheet music; *The Lady's Monthly Museum* (1798), which had a half-yearly "Cabinet of Fashion" illustrated by coloured engravings, the first to appear in a women's periodical; and *La Belle Assemblée* (1806), which encouraged its readers to unburden themselves in its correspondence columns. These three merged in 1832, the first instance of what was to become a common occurrence, but ceased publication in 1847. Later women's magazines included *The Ladies' Pocket Magazine* (1824–40), *The Ladies' Cabinet* (1832–52), *The New Monthly Belle Assemblée* (1847–70), and *The Ladies' Treasury* (1857–95). All contained verse, fiction, and articles of high moral tone but low intellectual content. There were attempts to swim against the tide, such as *The Female's Friend* (1846), which was one of the first periodicals to espouse women's rights, but they seldom lasted long.

In 1852 a wider market began to be tapped by *The Englishwoman's Domestic Magazine,* a monthly issued by Samuel Beeton at twopence instead of the usual one shilling; it was also the first women's periodical to concentrate on home management and offer practical advice to women rather than provide entertainment for the idle. Beeton's wife (author of the classic *Book of Household Management,* 1861) visited Paris regularly and acquired fashion plates from Adolphe Goubaud's *Moniteur de la Mode.* A feature of Beeton's magazine was the "Practical Dress Instructor," a forerunner of the paper dressmaking patterns. In 1861, Beeton followed up his success with *The Queen,* a weekly newspaper of more topical character.

The great expansion of women's magazines into a major industry may be dated in Britain from *Myra's Journal of Dress and Fashion* (1875–1912) and *Weldon's Ladies' Journal* (1875–1954), both of which supplied dressmaking patterns and met the needs of a mass readership. Several new quality magazines were started, such as *The Lady* (founded 1885) and *The Gentlewoman* (1890–1926), one of the first to acknowledge the financial necessity of advertisements, but there were many more cheap weeklies, such as *Home Notes* (1894–1957), *Home Chat* (1895–1958), and *Home Companion* (1897–1956); these were of great help in teaching women about hygiene, nutrition, and child care.

Among the earliest women's magazines in the United States was a monthly published in Philadelphia called *Godey's Lady's Book* (1830–98), which employed up to 150 women to hand-tint its fashion plates. Of the early national magazines, one of the best and hardiest was *Harper's Bazar* (1867; *Harper's Bazaar* after 1929), modeled on a Berlin women's periodical, *Der Bazar,* from which it obtained its fashion material. The practical trend was begun in 1863 by Ebenezer Butterick, who devised the tissue-paper clothing pattern and, to popularize it, brought out the *Ladies' Quarterly Review of Broadway Fashions* and, later, *Metropolitan.* These merged in 1873 into the *Delineator,* which had a highly successful career until 1937. The field of women's magazines was finally transformed, however, by Cyrus Curtis with his *Ladies' Home Journal* (founded 1883), edited by his wife, Louisa Knapp Curtis. This soon reached a circulation of 400,000 and, under the editorship of Edward W. Bok, from 1889, broke with sentimentality and piety to become a stimulating journal of real service to women. Other popular magazines were *Ladies' Home Companion* (1886; called *Woman's Home Companion,* 1897–1957), *McCall's Magazine* (founded 1897), and *Pictorial Review* (1899–1939).

Two requiring special mention were *Good Housekeeping* (founded 1885), which established a testing station for consumer goods early in the 20th century, and *Vogue* (founded 1892), a fashion weekly (later a monthly) dedicated to "the ceremonial side of life," which was designed for the elite of New York City and had Cornelius Vanderbilt among its backers.

Literary and scientific magazines. The critical review developed strongly in the 19th century, often as an adjunct to a book-publishing business. It became a forum for the questions of the day—political, literary, and artistic—to which many great figures contributed. There were also many magazines with a literary flavour, and these serialized some of the best fiction of the period. A few marked the beginning of specialization—*e.g.*, in science.

The *Edinburgh Review*

Britain was particularly rich in reviews, beginning with the *Edinburgh Review* (1802–1929), founded by a trio of gifted young critics: Francis Jeffrey, Henry Brougham, and Sydney Smith. The high and independent tone they adopted was said by Samuel Taylor Coleridge to mark an "epoch in periodical criticism." Though Tories, including at first Sir Walter Scott, wrote for it, the *Edinburgh Review* gradually became increasingly Whig in attitude. Scott accordingly transferred his allegiance to the *Quarterly Review* (1809–1967), the *Edinburgh Review*'s Tory rival, founded by the London publisher John Murray and first edited by William Gifford. Gifford had previously edited *The Anti-Jacobin* (1797–98), with which such figures as the Tory statesman George Canning were associated. In opposition to these, and more political than any of them, was the *Westminster Review* (1824–1914), started by Jeremy Bentham and James Mill as an organ of the philosophical radicals. Two other early reviews were the *Athenaeum* (1828–1921), an independent literary weekly, and the *Spectator* (founded 1828), a nonpartisan but conservative-leaning political weekly that nonetheless supported parliamentary reform and the cause of the North in the American Civil War. Later reviews included the *Saturday Review* (1855–1938), which had George Bernard Shaw and Max Beerbohm as drama critics (1895–1910); the *Fortnightly Review* (1865–1954), which had the Liberal statesman John Morley as editor (1867–83); the *Contemporary Review* (founded 1866); the *Nineteenth Century* (1877; later the *Twentieth Century*, until it closed in 1974); and W.T. Stead's *Review of Reviews* (1890–1936), a more limited version of *Reader's Digest*.

Of the closely related literary magazines, one of the earliest and best was *Blackwood's Edinburgh Magazine* (1817–1981), founded by a book publisher, William Blackwood, as a rival to the *Edinburgh Review*, but a less ponderous one than the *Quarterly*. It provoked in turn the founding of the *London Magazine* (1820–29), in which Charles Lamb's *Essays* first appeared. The rivalry between these two publications led to a duel in which John Scott, the first editor of the *London Magazine*, was mortally wounded. Other literary periodicals included the *Examiner* (1808–80), edited by the radical essayist Leigh Hunt, who introduced the poetry of Shelley and Keats to the public through its columns; the *New Monthly Magazine* (1814–84); *Bentley's Miscellany* (1837), which had Dickens as its first editor and *Oliver Twist* as one of its serials; and the *Cornhill* (1860–1975), first edited by William Thackeray and the first magazine of its kind to reach a circulation of 100,000. Finally, two rather different periodicals must be mentioned: *Nature* (founded 1869), which began to make scientific ideas more widely known and to which Charles Darwin and Thomas Huxley contributed; and *Punch* (founded 1841), which provided a weekly humorous comment on British life illustrated by many distinguished draftsmen.

Continental European reviews tended to be more literary than political, perhaps because of the persistence of censorship. The most notable in France were the *Revue des Deux Mondes* (founded 1829; later *La Nouvelle Revue des Deux Mondes*), with such contributors as Victor Hugo and the critic Charles-Augustin Sainte-Beuve, and its rival the (*Nouvelle*) *Revue de Paris* (founded 1829), which published authors disapproved of by the other, notably Gustave Flaubert. In Germany, F.A. Brockhaus, the book publisher, tried to emulate the *Edinburgh Review* with *Hermes* (1819–31) but had more success with *Literarisches Wochenblatt* (1820–98). Two later reviews were the conservative *Deutsche Rundschau* (founded 1874) and the liberal *Freie Bühne* (1890). Two influential Italian reviews were the *Nuova Antologia* (founded 1866) and *La Cultura* (1881–1935).

Early U.S. literary magazines

The early literary magazines in the United States included, among many others often of more local interest, the *Philadelphia Literary Magazine* (1803–08); the *Monthly Anthology* (Boston, 1803–11), which became the quarterly *North American Review* (1815–1940), with a host of famous contributors; the *New York Monthly Magazine* (1824); *Dial* (1840–44), the organ of the New England essayist Ralph Waldo Emerson's Transcendental Club (there was a second, literary *Dial*, 1880–1929); and *De Bow's Review* (New Orleans, 1846–80). The cultured weekly *Home Journal* (1846–1901; then continuing as *Town and Country*) introduced Swinburne and Balzac to Americans, while *Harper's New Monthly Magazine* (New York City, 1850; later called *Harper's Magazine*), founded by the book-publishing Harper brothers, serialized many of the great British novels and became one of America's finest quality magazines. It was rivaled only by the *Atlantic* (Boston, 1857; later called *Atlantic Monthly*), which had a long line of distinguished editors, beginning with James Russell Lowell, and published most of the great American writers, from Ralph Waldo Emerson, Henry Wadsworth Longfellow, and Oliver Wendell Holmes onward; it seemed to enjoy "a perpetual state of literary grace." Similar in quality was *Scribner's Monthly* (1870), which became the *Century* (1881–1930) but was restarted as *Scribner's Magazine* (1887–1939). A fine magazine in the Far West was *Overland Monthly* (San Francisco, 1868–1935), first edited by Bret Harte. Non-literary specialized magazines included *Scientific American*, which was founded in 1845 by Rufus Porter, a talented inventor whose magazine encouraged other inventors; *Popular Science Monthly*, which was founded in 1872, to spread scientific knowledge and which had the philosophers William James and John Dewey among its contributors; and the ever-popular *National Geographic Magazine*, founded in 1888 and published ever since by the National Geographic Society, which used some of the proceeds to sponsor scientific expeditions.

Scholarly journals. The publishing of scholarly journals, begun in the 17th century, expanded greatly in the 19th as fresh fields of inquiry opened up or old ones were further divided into specialties. Numerous learned societies were formed in such fields as classical studies, biblical studies, archaeology, philology, Egyptology, the Orient, and all the branches into which science was dividing, and each society published a regular bulletin, proceedings, or "transactions," which enabled scholars to keep in touch with what others were doing. In the sober pages of these journals, seldom read by the general public, some of the most far-reaching discoveries were first made known. Among the many notable publications were *Annali del Istituto di Corrispondenza Archeologica* (1829), the *Revue Archéologique* (founded 1844), *Philologus* (1846), *Mind* (founded 1876), the *Journal of Hellenic Studies* (founded 1880), the *American Journal of Philology* (founded 1880), the *Asiatic Quarterly* (1886; later called *South Asian Review*), the *Geographical Journal* (1893), and an interesting informal aid to scholars, *Notes and Queries* (1849), with the motto: "When found, make a note of." In every advanced country the professions too began to have journals, such as medicine's *Lancet* (founded 1823), in Britain, originally started to attack abuses in hospital administration; the *Mining Journal* (founded 1835); the *British Medical Journal* (founded 1840); *The Engineer* (founded 1856); and the *Solicitors' Journal* (founded 1857), to cite only a few examples. In the course of time, these developed endless technical ramifications. The economics of all such journals are based on necessity. Though their circulation is small, anyone working in a particular field generally subscribes to them or at least has access to them in appropriate libraries. They can be described as reference books in installments.

THE 20TH CENTURY

The advertising revolution in popular magazines. There was a certain resistance to advertising in magazines, in keeping with their literary affinities. When the advertisement tax in Britain was repealed in 1853 and more advertising began to appear, the *Athenaeum* thought fit to say: "It is the duty of an independent journal to protect as far as possible the credulous, confiding and unwary from the wily arts of the insidious advertiser." In the United States many magazines, such as *Harper's,* took a high line with would-be advertisers until the 1880s; and *Reader's Digest,* with its mammoth circulation, admitted advertisements to its American edition only in 1955. Yet today some sectors of the magazine industry are dominated by advertising, and few are wholly free from its influence.

Magazine advertising economics. In the United States Cyrus Curtis showed what could be achieved in attracting advertising revenue with the *Saturday Evening Post.* He bought the magazine for $1,000 in 1897, when it was on its last legs, and invested $1,250,000 of his profits from the *Ladies' Home Journal* before it finally caught on. But when it did, through an appeal based on well-founded stories and articles about the business world, a prime interest at the time, its success was enormous; by 1922 it had a circulation of more than 2,000,000 and an advertising revenue in excess of $28,000,000. It was a classic demonstration of modern magazine economics: as circulation rose in the initial phase of low advertising rates, money had to be poured in to meet the cost of producing more copies; but, as soon as high advertising rates could be justified by a high circulation, profitability was assured. Conversely, when high rates are maintained on a falling circulation, it is the advertisers who lose, until they withdraw their support.

Verifying circulation

Once circulation figures became all-important, advertisers naturally asserted their right to verify them. The first attempt, made in 1899 by the Association of American Advertisers, only lasted until 1913, but fresh initiatives in 1914 created the Audit Bureau of Circulation. Though resented at first by publishers, it was eventually seen as a guarantee of their claims. Interest in circulation led publishers into market research. The first organization for this purpose was set up by the Curtis Publishing Company in 1911; but such research did not become general until the 1930s. Reader research, to ascertain what readers wanted from magazines, was also developed in the 1930s and proved to be a useful tool, though no substitute for editorial flair. As was once observed by the features editor of *Vogue:* "If we find out what people want, it's already too late."

Thus the popular magazine in the United States, expanding with the economy, became part of the marketing system. By 1900 advertisements might form up to 50 percent of its contents; by 1947, the proportion was more often 65 percent. A proprietor was no longer just selling attractive editorial matter to a segment of the public; he was also selling a well-charted segment of the public to the advertiser. Though the process was most pronounced in the United States, a vast country where, in the absence of national newspapers, national magazines had a special function, the same principles came to apply, in varying degrees, in Europe.

The effects of advertising on the appearance of the magazine have been, on the whole, stimulating. At the turn of the century, advertisements began to move forward from the back pages into greater prominence among the editorial matter, and this was often regretted by readers. At the same time, advertising agencies were developing from mere space sellers into copywriters and designers; their efforts to produce work of high visual appeal forced editors to make their own editorial typography and layout more attractive. The use of colour, in particular, was greatly fostered by advertisers once they discovered its effectiveness. In the 1880s colour printing was rare, but, after the development of the multicolour rotary press in the 1890s, it steadily became more common. By 1948 nearly half the advertising pages of the leading American magazines were in two or more colours.

The effect of advertising on editorial content is harder to analyze. Advertisers have not been slow to exercise financial pressure and have often succeeded in suppressing material or modifying policy. In 1940, for instance, *Esquire* lost its piano advertisements after publishing an article recommending the guitar for musical accompaniment; six months later it tried to win them back with a rueful editorial apology. Yet many magazines, notably the *Saturday Evening Post, Time,* and *The New Yorker,* have persistently asserted editorial independence. Something like a balance of power has come into being, which can tip either way. What can safely be said is that advertising pressure as a whole has been a socially conservative force, playing on conformity, inclining magazines to work on the principle of "minimum offense," and holding them back from radical editorial departures until they are clearly indicated by changes in public taste. This has tended to make the large-circulation magazine an exploiter rather than a discoverer of fresh talent or new ideas. Yet in the last analysis, advertisers have been forced to recognize that magazines, like newspapers, cannot forgo too much of their independence without forfeiting the loyalty of their readers and hence their value as an advertising medium.

Advertising and editorial content

Women's magazines in the United States. The bond with advertising is probably most evident in magazines for women, since they are the greatest buyers of consumer goods. In the United States, up to the mid-1930s, such magazines were largely "trade-papers for home-makers." There were exceptions, such as *True Story* (founded 1919), which concentrated on entertainment, and *Vogue,* which introduced readers to a wider world, but more typical was *Better Homes and Gardens* (founded 1922), which gave fresh impetus to the trend toward "service" by helping both men and women in the running of their homes. In this area, of course, advertising pressure can be considerable—*e.g.,* for editorial support of a new product—but editors have usually contained it within some limits.

An innovation in the 1930s was the store-distributed magazine. One of the first and most successful was *Family Circle* (founded 1932), given away in Piggly Wiggly supermarkets until 1946, when it was sold as a family monthly. Equally successful were *Woman's Day* (founded 1937), published by a subsidiary of the Great Atlantic and Pacific Tea Company, and *Better Living* (founded 1951), sponsored by the Super Market Institute. During the 1930s women's magazines broadened their base to combat falling circulations and to meet changes in taste, as they did again in the 1950s, in a similar crisis.

By the late 1980s scores of political and literary magazines of broadly feminist sympathies had been established, one of the most prominent being *Ms.* (founded 1972), a nonprofit magazine with a circulation of about 500,000. Another general trend has been to direct appeal toward younger women, not only in the old magazines but also in such newer ones as *Seventeen* (founded 1944), *Ingenue* (founded 1959), and *Teen* (founded 1957).

Advertising in Britain and Europe. Though the advertising revolution began in Britain at much the same time as in the United States, its course has been less explosive. By 1898, *The Gentlewoman* was pointing out in its first issue that every copy cost "nearly double the price for which it is sold." Yet Britain's Audit Bureau of Circulations was not set up until 1931, and membership remained small until the 1960s; for it was only then that consumer spending in Britain (and hence advertising) really began to soar, to be reflected in a boom in women's magazines. In the early part of the century, the old general magazines continued to flourish, with such additions as the *Windsor Magazine* (1895–1939), *Pearson's Magazine* (1896–1909), *Argosy* (founded 1926), which published only fiction, and the popular weekly *John Bull* (1906–64), which thrived on "revelations." Several American magazines, especially women's, began to come out in British editions, such as *Vogue* (1916), *Good Housekeeping* (1922), and *Harper's Bazaar* (1929; in 1970 amalgamated with *Queen* as *Harpers & Queen*). Society periodicals lost ground after World War I to those catering to the so-called new poor and new rich, although snobbery still proved a lucrative element in magazine publishing, notably with the *Tatler,* which became highly successful under a new edi-

tor in the early 1980s. The fortnightly *Queen, Woman's Weekly* (founded 1911), and the monthly *Woman and Home* (founded 1926) and *Woman's Journal* (founded 1927) were joined by such popular weeklies as *Woman's Own* (founded 1932), *Woman's Illustrated* (1936–61), and, above all, *Woman* (founded 1937), the first to be printed by colourgravure. During World War II some of these magazines gave valuable practical advice on how to cope with shortages. In postwar Britain magazines began to be distributed through retail outlets—mostly supermarkets—other than bookshops or newsagents. The chief examples were *Family Circle* (founded 1964), an Anglo-American production, and its sister publication, *Living* (founded 1967). The trend toward youthful markets was indicated by *She* (founded 1955), broad and robust in outlook; *Honey* (founded 1960); *Annabel* (founded 1966), for younger married women in particular; *Petticoat* (1966–75), for girls 14 to 19 years old; and *19* (1968), a market leader. The death of many of the old general magazines, under the pressure of paperbacks and television, and the dearth of illustrated weeklies (see below *Picture magazines*) left room for a new advertising vehicle. The first to perceive this was Lord Thomson, who in 1962 brought out a colour magazine as supplement to the *Sunday Times* (London). Its eventual success forced the *Observer* and the *Daily Telegraph* to follow suit (the colour supplement was eventually removed from the latter paper and issued instead with its sister publication, the *Sunday Telegraph*). In the early 1980s the popular Sunday papers also started supplements.

Youth magazines

In the rest of Europe the impact of advertising on magazines has been more delayed and less pronounced, partly because market prices of continental magazines tend to be closer to the production cost. General magazines were fairly limited before World War II, but since then, as part of the economic expansion, there has been a rich crop, including many newsmagazines similar to *Time* and *Life* and also a number of magazines for women. France has several of the latter with large circulations, including *Nous Deux, Elle,* and *Intimité,* while those in Germany include entries for all age groups, such as *Jasmin* for newlyweds and *Eltern* for parents. Though the northern European countries have fewer periodicals, it is worth noting that in Finland *Pirkka,* a giveaway distributed through grocery stores, achieved one of the largest magazine circulations.

Magazines outside Europe and the United States. *Japan.* The outstanding early 20th-century personality in Japanese magazine publication was Noma Seiji, who published nine magazines, nearly all with six-figure circulations. World War II did not seriously affect periodicals; and, at the end of occupation in 1952, there were more than 2,000 of all kinds, including *Shufu No Tomo* (1917–56; "Woman's Friend"), *Yoiko No Tomo* (1924–57; "Child's Friend"), and *Le-no-Hikari* (founded 1925; "Light of Home").

Africa. Important publications in Africa have included the quarterly East African *Africana* (founded 1962); the Zimbabwean *Africa Calls* (1960), published every two months; the quarterly *Nigeria Magazine* (1933–66); the quarterly *Pan African Journal* (1967), published in Kenya; and, in South Africa, journals in Afrikaans. Elsewhere, magazines in African languages have increased, as have those in English and French—*e.g.,* the Nigerian *Black Orpheus* (founded 1957), containing creative writing by Africans and West Indians.

India, Bangladesh, and Pakistan. Important 20th-century magazines in India include the *Illustrated Weekly of India* (founded 1880), a topical review for educated readers; the *Statesman Weekly* (founded 1924), an illustrated digest of Indian news and views; the monthly general review *Current Events* (founded 1955); *Thought* (New Delhi, 1949–78/79), a political and economic weekly; the monthly *Akhand Anand* (founded 1947); and the weekly *Akashvani* (founded 1936), *Dharmayug* (founded 1950), and *Mukhabir-I-Alam* (1903). *Sport and Pastime* (1947), with offices in several cities, is well illustrated. *Eve's Weekly* (founded 1947), in English, Urdu, and Hindi, is a popular women's magazine. Bangladesh weeklies include *Bangladesh Sangbad* (founded 1972). Pakistani periodicals include the monthly *Subrang Digest* (1970) and the weekly *Muslim World* (1961).

South America. Argentina had a greater magazine circulation than any other nation in South America until the mid-1970s, when total circulation decreased by almost one-half (it later began to recover slowly). The weekly rotogravure *Maribel* (1932–56) long had the highest periodical circulation in that country, closely followed by that of the women's weekly *Para ti* (founded 1922). Mexico's leading magazine in the early 1980s was the weekly *Selecciones del Reader's Digest;* others included the weeklies *El Libro Semanal* (1954) and *Alarma* (1963). Venezuelan periodicals include the weekly *Resumen* (founded 1973) and *Elite* (1925).

News and photo magazines. The accelerated tempo of life in the 20th century, coupled with the bewildering amount of information appearing in print, suggested the need for more concise ways of presenting it. The first to show how it could be done and so give rise to a whole new class of periodical was the U.S. newsmagazine *Time,* founded in 1923 by Briton Hadden and Henry Luce.

The Luce magazines

Time magazine. There had, of course, been newsmagazines before, in both Europe and the United States. *Time* magazine's immediate forerunner was the *Pathfinder* (1894–1954), a weekly rewriting of the news for rural readers. There had also been attempts at compression of the digest type (see below *Digests and pocket magazines*). But *Time* was the first to aim at a brief and systematic presentation of the whole of the world's news. It was based on the proposition that "people are uninformed because no publication has adapted itself to the time which busy men are able to spend simply keeping informed." Its beginning was amateurish and precarious; neither Hadden nor Luce had much experience when they started summarizing the news from bundles of daily papers (copyright provisions on newspapers allowing this use). But after 1928 it grew steadily, finding its market chiefly among the rising number of college graduates. What came to be known as the *Time* style was characterized, in the words of a later critic, by two great democratic ideals, disrespect for authority and reverence for success. *Time* presented the news in tightly packed sentences, well researched and checked, and with a general air of omniscience. In the 1930s, to ensure adequate sources of information, Time Inc. built up a large news-gathering organization of its own. It also branched out into other publications, including *Fortune* (founded 1930), summarizing business news, *Life* (see below), and *People,* a weekly begun in 1974.

Among the direct followers of *Time* in the United States were *Business Week* (founded 1929), *United States News* (founded 1933), and *Newsweek* (founded 1933), its nearest rival. Similar magazines appeared in Shanghai (*East,* 1933) and in Britain (the *News Review,* 1936), though the latter did not have a comparable success, partly because Britain was so well supplied with national dailies. After World War II the United States had several newsmagazines of a regional nature, such as *Fortnight* (1946) in California and *Texas Week* (1946). *Time* has had its greatest influence, however, in postwar Europe, where such magazines as *L'Express* (founded 1953) in France, *Der Spiegel* (founded 1947) in Germany, and *Panorama* (founded 1962) in Italy derived directly from it. Such magazines did not always develop in exactly the direction that *Time* had taken, but *L'Express* was radically changed at least twice by its owners; the first time it followed *Time* fairly closely. *Der Spiegel* ("The Mirror") became famous for its aggressive, antiauthoritarian exposures of scandal and malpractice, while *Panorama* achieved a high standing and a reputation for reliability. The influence of *Time* can probably be traced in most newsmagazines, as in *Tiempo* (founded 1942) in Mexico or *Primera Plana* (founded 1962) in Argentina.

Picture magazines. Conciseness can also be achieved through pictures, which obviate the need for description. Illustrated newsmagazines began in the 19th century, but they took an altogether new form as photography developed. The most influential, though by no means the first of the modern type, was undoubtedly the American weekly *Life* (1936–72), started by Henry Luce.

Pictorial journalism grew up alongside advertising techniques, the tabloid, and the documentary film. Modern cameras enabled top-grade photographs to be taken quickly under almost any conditions. Photojournalists were particularly active in Germany, until many had to flee the Nazis. One of them was the Hungarian Stefan Lorant, who developed the photo essay (a story reported through pictures) with *Bilder Courier* in Berlin in 1926 and with the *Münchener illustrierte Presse* in the period 1927–33. He then went to Britain, where he started a pocket picture magazine, *Lilliput* (1937–60), and was the first editor of *Picture Post* (1938–57). Another pioneer was a German, Erich Salomon, who became celebrated for his photographs of the famous, particularly politicians, in unguarded moments. Salomon's pictures in the London *Tatler* in 1928 prompted *Fortune* to invite him to the United States, where he inspired the *Life* photographer Thomas McAvoy.

In November 1936, therefore, when *Life* first appeared, picture magazines were already fairly common. Only a month before, *Mid-Week Pictorial* (1914–37), an American weekly of news pictures, had been restyled along the lines *Life* was to take, but *Life* quickly overwhelmed it. Though expected to have a circulation of well under 500,000 copies, *Life* was running at 1,000,000 within weeks. Its first issue, 96 large pages of pictures on glossy paper for 10 cents, was a sellout, the opening picture brilliant: an obstetrician holding a newborn baby, with the caption "Life begins." Over the years, it kept the promise of its prospectus: "To see life; to see the world; to witness great events; to watch the faces of the poor and the gestures of the proud; to see strange things. . . ." During World War II, which it covered with great accomplishment, it enlarged its operations with a fortnightly international edition, and in 1952 a Spanish-language edition was added for Latin America, *Life en Español.* In 1971 *Life* magazine's circulation was about 7,000,000, but its high costs were no longer being met by advertising income, and it ceased publication in December 1972; it was revived as a monthly in October 1978.

Imitators of *Life* magazine

Of the countless imitators of *Life,* many were American, such as *Focus, Peek, Foto,* and two of longer duration, *Pic* (1937–48) and *Click* (1938–44). Best known was *Look* (1937–71; briefly revived 1979), a popular biweekly. It was founded by Gardner Cowles, Jr., who also started *Quick* (1949–53), a miniature magazine. Britain had two news picture magazines, *Picture Post* (1938–57), which acquired much prestige through its social conscience, and *Illustrated* (1939–58); their place was taken to some extent by the Sunday colour supplements. The French *Paris-Match* (founded 1949), exceptionally well-produced and well-supplied with photographers, gained preeminence throughout Europe; while West Germany produced *Stern* (founded 1948), a glossy blend of light and serious material, and Italy, where magazines are read more than newspapers, produced *Oggi Illustrato* (founded 1945), thriving on not-too-sensational disclosures, and the elegant *Epoca* (founded 1950). Magazines similar to *Life* appeared in a number of other countries, such as *Cruzeiro* (founded about 1908) in Brazil and *Perspektywy* (founded 1969) in Poland, and still more that follow the style of *Look,* such as *Manchete Esportiva* (founded 1952) in Brazil, *Caretas* (founded 1950) in Peru, or the Australian *Pix–People.*

Digests and pocket magazines. *Reader's Digest magazine.* The need for concise reading matter, so well met by *Time* and *Life,* was met even more successfully, in terms of circulation, by an American magazine that reprinted in condensed form articles from other periodicals. This was the pocket-size *Reader's Digest,* founded in 1922 by DeWitt Wallace.

Its forerunners in the United States were the *Literary Digest* (1890–1938), started by two former Lutheran ministers, Isaac K. Funk and Adam W. Wagnalls; the *Review of Reviews* (1890–1937), founded by Albert Shaw to condense material about world affairs; and Frank Munsey's *Scrap Book* (1906–12), "a granary for the gleanings of literature." The *Literary Digest,* in particular, with a circulation of more than 1,000,000 in the early 1920s, was something of an American institution. Its famous straw votes successfully predicted the result of the presidential elections after 1920, and its highly publicized wrong prediction of the outcome of the 1936 election played a decisive part in its collapse. *Reader's Digest,* however, was more specific in content and more universal in appeal. It aimed to supply "An article a day from leading magazines in condensed, permanent, booklet form." Each article, moreover, satisfied three criteria: "applicability" (it had to be of concern to the average reader); "lasting interest" (it had to be readable a year later); and "constructiveness" (it had to be on the side of optimism and good works).

After three years' preparation, Wallace began to produce the magazine (first issue February 1922) from a basement office in New York City. After a year, subscriptions were running at about 7,000. In 1939, when circulation had reached 3,000,000, *Reader's Digest* moved into large premises at nearby Chappaqua. Until 1930 it was produced entirely by amateurs. Condensed books began to be added at the end of the magazine in 1934, and from this grew the Reader's Digest Condensed Book Club, with 2,500,000 members four years later. Overseas editions were started in 1939 (British), and foreign-language editions in 1940 (Spanish), others being steadily added over the following 10 years. In the late 1980s, *Reader's Digest* had one of the largest circulations of any magazine in the world.

Problems of condensing articles

This success was not achieved entirely without setbacks and criticism. At first, permission to reprint was easy to obtain and was without charge; but after a while, and especially after competitors entered the field and sometimes reprinted without permission, magazines began to regard the digests as parasitic. Payments were required, which rose steadily, and the major proprietors withheld their permission at various times. To guard against this and because articles of the sort he wanted were in short supply, Wallace began to print original material in the *Digest* in 1933. To keep up the appearance of a digest, articles were commissioned and then offered to other magazines in exchange for the right to "condense" and reprint them. Such articles, "cooperatively planned" according to the *Digest,* "planted" according to critics, were naturally welcome to many magazines with slender budgets, but they did lead to controversy. In 1944 *The New Yorker,* fearing that *Reader's Digest* was generating too big a fraction of magazine articles in the United States, attacked the system as "a threat to the free flow of ideas and to the independent spirit"; but, in the more general view, the matter was regarded as a private one for the parties concerned. Internationally, too, the *Digest* was attacked by some after World War II for its part in "American cultural imperialism"; but it has continued to find favour with the magazine public.

The digest idea was soon taken up by others, often in direct competition but also in more limited areas, such as *Science Digest, Catholic Digest, Negro Digest,* and *Children's Digest.* There was also a *Cartoon Digest* (1939), an *Editorial Digest* (1947), and a *Column Digest* (1949). Most of the general digests used original articles, since competition for the limited amount of highly popular reprinted material became too keen, and *Reader's Digest,* as first in the field, was always able to outbid its competitors. One of the more successful was *Magazine Digest* (founded 1930), which was based in Canada and contained a good deal of scientific and technical matter. One that tried a new formula, based on timeliness and a liberal slant, was *Reader's Scope* (1943–48). The most successful book digest was probably *Omnibook* (1938–57), each issue of which contained abridgments of several popular works of fiction and nonfiction. The digests originally carried no advertising, but after World War II they were gradually driven to it by rising costs. One of the last to capitulate was *Reader's Digest* in 1955; the proportion of advertising was restricted, however, to 20 percent.

Types of pocket magazines. The success of *Reader's Digest* also had an influence through its format; it popularized the pocket magazine as a type. Several of the self-improving variety, such as *Your Life* (founded 1937) and *Success Today* (1946–50), were started by Wilfred J. Funk on the proceeds from his father's *Literary Digest* (sold to *Time* in 1938). Of those more directly inspired by *Read-*

er's Digest, Coronet (1936–61), an offshoot of Esquire Inc., built up a large circulation during World War II, and when it closed, a victim of the promotion race, it was still running at more than 3,000,000. Somewhat livelier and glossier was *Pageant,* first published in 1944. Britain had several pocket magazines, such as *London Opinion, Men Only,* and *Lilliput,* but these owed rather less to *Reader's Digest.* Finally, there have been a few "superdigests," miniature newsmagazines with pictures and a minimum of text, such as *Tempo* (1950), *People Today* (founded 1950), and *Jet* (founded 1951).

Specialized magazines. Though general magazines have the largest circulations, most magazines cater to specialist interests or pursuits. Circulation varies, but, even where it is small, it is usually stable over the short term and offers an advertiser a well-defined market. Such magazines may be broadly classified into professional (including trade and technical) and nonprofessional journals.

Professional types. The professional magazine, often the organ of an association, keeps members informed of the latest developments, helps them to maintain standards, and defends their interests. Some were started in the 19th century, but specialization and different viewpoints within specialties have encouraged proliferation. Instead of two or three medical journals, for instance, there are now likely to be dozens, besides those in specialized areas such as dentistry, ophthalmology, and psychiatry. Though most of these magazines are of little interest to the general public, a few print authoritative articles of broader scope.

Trade and technical journals serve those working in industry and commerce. They too have grown enormously in numbers. Major discoveries in science, manufacturing methods, or business practice tend to create a new subdivision of technology, with its own practitioners and, more often than not, its own magazine. Articles in these magazines tend to be highly factual and accurately written, by people deeply immersed in their subjects. Most are well produced, often on art paper for the sake of the illustrations, and heavily dependent on advertising. Indeed, many are issued for a controlled circulation; *i.e.,* a publisher undertakes to distribute a magazine free of charge to a given number of specialist concerns, which can be relied upon to want a certain range of products. The manufacturers of these products, for their part, are naturally glad to have an advertising medium guaranteed to reach their particular market. The business papers may lack glamour, but they play a vital and highly influential part in economic life.

Controlled circulation magazines

Nonprofessional types. Of the nonprofessional magazines, quite a number serve broad interest groups, religious, political, or social. Most religious denominations have journals, often more than one. Though some of these magazines are subsidized as part of a drive to spread their message, most of them merely aim to foster corporate feeling among coreligionists. Much the same applies to political magazines in the narrow sense—*i.e.,* where they are issued by political organizations: they discuss doctrine, give news of activities, and forge links among members. Political discussion on less partisan matters and in a less partisan tone tends to take place in more general magazines. Certain periodicals spring from the needs of particular groups, an example being student magazines.

Specialized magazines for the layman may fall into the hobby category. Very often a professional magazine has an amateur counterpart, as, for instance, in electronics, where the amateur finds a wide range of technical magazines on radio, television, hi-fi, and tape recording. Other popular subjects are photography (the British *Amateur Photographer* was founded in 1884) and motoring (Hearst's *Motor* was founded, as *Motor Cycling and Motoring,* in 1902); specialization even extends to types of camera and makes of car. Virtually no hobby or sport is without its magazine. As soon as any activity becomes sufficiently popular, a magazine appears to cater to its adherents and to provide an advertising medium, not only for manufacturers and suppliers but also for readers, to help them buy and sell secondhand equipment, for instance.

"Pulp" and "comic" magazines

Some special tastes in entertainment are met by the "pulp" and "comic" magazines. In 1896 Frank Munsey turned his *Argosy* into an all-fiction magazine using rough wood-pulp paper. The "dime novel" did not qualify for inexpensive postal rates in the United States, but the pulp magazine did, and so an industry was born. Pulps began as adventure magazines but soon split up into further categories: love, detective, and western. Such magazines sold in the millions up to the mid-1930s, when they gradually lost ground to the comics. These began as collections reprinted from the comic strips in newspapers; the first to appear regularly was *Famous Funnies* (1934). After 1937, however, with *Detective Comics,* they came into their own as original publications, and, like the pulps, they grew into a major industry, dividing up into much the same types. They may be seen, in effect, as pictorial condensations of the pulps. Though mainly for children, they were widely read by adults. "Comic" rapidly became a misnomer, as they played increasingly on horror and violence. While some defended them as harmless and even cathartic, others condemned them as incitements to imitation. Attempts at control were made through legislation in the United States and elsewhere, and the industry itself tried to set standards. Television has since drawn much of the criticism, and the demand, to itself, but comics remain big business. One type of magazine, originally classed as pulp but attaining with the years a certain respectability, is the science-fiction magazine, the first example of which was Hugo Gernsback's *Amazing Stories,* first published in 1926.

The "fan" magazines offer glimpses of life behind the scenes in the world of entertainment and sport. In the heyday of motion pictures, many magazines on films and their stars appeared, beginning with *Photoplay* (1911–77) and *Picture Play* (1915) and later others, such as *Movie Mirror* (1930) and *Movieland* (1942). When radio and television became popular, similar magazines sprang up centring on programs and their personalities. One of their functions was to provide a weekly timetable of programs.

Finally, there are a number of "special service" magazines—*e.g.,* financial magazines to help the private investor, magazines of advice issued by consumer associations, magazines specifically for house hunters, racegoers, or for trading in secondhand goods, and so on.

Scholarly, cultural, and literary magazines. As the 20th century progressed, the old critical review lost some of its former glory, but it often wielded an influence quite out of proportion to its circulation. One may distinguish broadly between the scholarly type of review, the more widely read politico-cultural periodical, and the purely literary magazine.

Britain. Many of the British reviews founded in the 19th century have continued to flourish. Among additions of the scholarly type were the *Hibbert Journal* (1902–70), a nonsectarian quarterly for the discussion of religion, philosophy, sociology, and the arts; the *Times Literary Supplement* (founded 1902), important for the completeness of its coverage of all aspects of books and bibliographical matters; *International Affairs* (founded 1922), the journal of Chatham House, the Royal Institute of International Affairs; and *The Political Quarterly* (founded 1930), for the discussion of social and political questions from a progressive but nonparty point of view. Of the weekly political reviews, the *Spectator* (founded 1828), was representative of the right, and the *New Statesman* (founded 1913), founded by Sidney and Beatrice Webb, of the left, though both in a broad context; while *Time and Tide* (1920–79), originally founded by Lady Rhondda as an independent journal, was an influential newsmagazine. Several other periodicals met the need for serious articles on current questions; among them are *The Economist* (founded 1843); *The Listener* (founded 1929), published by the British Broadcasting Corporation and consisting mainly of radio talks in printed form; the *New Scientist* (founded 1956), drawing attention to current scientific work; and *New Society* (founded 1962), concentrating on sociology. Literary magazines came and went, but not without leaving their mark. They included the *Egoist* (1914–19), associated with Ezra Pound and the Imagists; the *London Mercury* (1919–39), started by J.C. (later Sir John) Squire, one of the Georgian poets; the *Criterion* (1922–39), founded and edited by T.S. Eliot; the *Adelphi* (1923–55), of John Middleton Murry; *New Writing*

Political reviews

(1936–46), edited by John Lehmann, who also later revived the old *London Magazine* (from 1954); and *Horizon* (1940–50; revived 1958), which Cyril Connolly started as a medium for literature during the war years. Later, *Encounter* (founded 1953), an international review originally sponsored by the Congress for Cultural Freedom, proved to be an intellectual magazine of value and distinction. In addition, many "little magazines" have struggled along, as always, providing essential seedbeds for new writers.

American scholarly journals

The United States. American counterparts to British scholarly journals include the *Political Science Quarterly* (founded 1886), edited by the political science faculty of Columbia University; the *American Scholar* (founded 1932), "a quarterly for the independent thinker" edited by the united chapters of Phi Beta Kappa; *Foreign Affairs* (founded 1922), a quarterly dealing with the international aspects of America's political and economic problems; and *Arts in Society* (founded 1958), a forum for the discussion of the role of art, which also publishes poetry and reviews. Of general political journals, the oldest still in publication at the turn of the 21st century was *The Nation,* founded in 1865 by E.L. Godkin and edited in the period 1918–34 by Oswald Garrison Villard. By tradition it adopted a critical stand on most matters, disdaining approval by the majority; it was notable for the "casual brilliance" of its literary reviews. When the muckraking phase in the popular magazines died down, zeal for reform was left to a succession of little magazines that led precarious lives, often needing extra support from loyal readers or rich individuals. Such were the *Progressive* (founded 1909), of the La Follette family; *The Masses* (1911–17), run by the Greenwich Village socialists; and *The New Republic* (founded 1914), which was started by Herbert Croly with the backing of the Straight family as "frankly an experiment" and "a journal of opinion to meet the challenge of the new time" and which survived as a liberal organ after many triumphs and vicissitudes. Between the wars came the Marxist *Liberator* (1918–24); the *Freeman* (1920–24 and 1950–54), founded to recommend the single-tax principle of Henry George and later revived as a Republican journal; the *New Leader* (founded 1927), for 10 years the organ of the American Socialist Party; and the extreme left *New Masses* (1926–48). Postwar foundations included the anticommunist *Plain Talk* (1946–50); the fortnightly *Reporter* (1949–68), strong on "facts and ideas"; and the conservative *National Review* (founded 1955). Of the literary magazines, the *Atlantic* and *Harper's* were joined by the *American Mercury* (founded 1924), which had a brilliant initial period under H.L. Mencken and George Jean Nathan, when it published work by many distinguished writers of the time; and the *Saturday Review* (founded 1924), which began as a purely literary magazine but broadened its scope in the 1940s. In 1972 a new ownership brought more changes. A powerful influence on American writing has been exerted by *The New Yorker* (founded 1925), mainly through its founder Harold Ross, a perfectionist among editors. It became famous for its cartoons and biographical studies. Finally, there has been no lack of "little magazines" to foster talent.

Literary magazines

Continental Europe. Among the numerous literary magazines in Europe, several in France and Germany in particular may be mentioned. The *Mercure de France* was revived in 1890 as an organ of the Symbolists; the influential *Nouvelle Revue Française* (1909) aimed at a fresh examination of literary and intellectual values; and the *Nouvelles Littéraires* (1922) was founded by André Gillon as a weekly of information, criticism, and bibliography. After World War II there appeared Jean-Paul Sartre's left-wing monthly *Les Temps Modernes* (founded 1945), *La Table Ronde* (1948), and *Les Lettres Nouvelles* (1953). In Germany, political magazines included the radical *Die Fackel* (1899; "The Torch") and *Die neue Gesellschaft* (1903–07; "The New Society") of the Social Democrats. An important literary influence was *Blätter für die Kunst,* associated with the Neoromantic movement of Stefan George. The Nazi period imposed a break in development, but after World War II the liberal weekly *Die Zeit* and a number of literary journals, such as *Westermanns Monatshefte, Neue deutsche Hefte,* and *Akzente,* appeared.

The political involvement of the literary review was especially marked in the Soviet Union and Soviet-bloc countries. The *Literaturnaya Gazeta* (founded 1929) and the influential *Novy Mir* (founded 1925; "New World") often became the centre of controversy in the Soviet Union when writers were condemned for their views or denied the opportunity to publish. This led to a strong underground press. In Czechoslovakia the *Literárne Listy* played a prominent part in the freedom movement of 1968 and was later suppressed at Soviet insistence, along with the *Reportér* and *Student,* leading to the start of several underground magazines. *Sinn und Form* (founded 1949), a Marxist critical journal in Berlin, was subject to temporary suspensions for publishing such authors as Sartre, Kafka, and Hemingway, whose works had been banned in East Germany. (P.U./G.U./Ed.)

BIBLIOGRAPHY

General works. General discussions can be found in ALBERTO MANGUEL, *A History of Reading* (1996); and W. SOMERSET MAUGHAM, *Books and You* (1940, reprinted 1977). DAVID M. BROWNSTONE and IRENE M. FRANCK, *The Dictionary of Publishing* (1982); and JEAN PETERS (ed.), *The Bookman's Glossary,* 6th rev. and enlarged ed. (1983), explain terminology, the former emphasizing business aspects. COLIN CLAIR, *A History of Printing in Britain* (1966), and *A History of European Printing* (1976), are useful for early periods. S.H. STEINBERG, *Five Hundred Years of Printing,* 3rd ed. (1974), is comprehensive; it may be supplemented by ROBIN MYERS and MICHAEL HARRIS (eds.), *A Millennium of the Book: Production, Design & Illustration in Manuscript & Print, 900–1900* (1994), which, despite its title, treats more than books. JOHN W. SEYBOLD, *The World of Digital Typesetting* (1984), charts the history of various printing techniques from the earliest days to the 1980s and emphasizes the importance of computers. Other developments are discussed in BRIAN KAHIN and HAL R. VARIAN (eds.), *Internet Publishing and Beyond: the Economics of Digital Information and Intellectual Property* (2000); and INGO VOGELSANG and BENJAMIN M. COMPAINE (eds.), *The Internet Upheaval: Raising Questions, Seeking Answers in Communications Policy* (2000).

Legal aspects of the industry are explored in W.J. LEAPER, *Copyright and Performing Rights* (1957), an early history of copyright in England and the implications of the Berne Convention and the Universal Copyright Convention; ALLEN KENT and HAROLD LANCOUR (eds.), *Copyright: Current Viewpoints on History, Laws, Legislation* (1972), a collection of essays from professional sources; RICHARD WINCOR and IRVING MANDELL, *Copyright, Patents, and Trademarks: The Protection of Intellectual and Industrial Property* (1980), a history of copyright in the United States; DENIS DE FREITAS, *The Copyright System: Practice and Problems in Developing Countries* (1983), a survey of key principles and practices; LEE BOAZ HALL, *International Magazine and Book Licensing* (1983); JACQUELINE M.B. SEIGNETTE, *Challenges to the Creator Doctrine: Authorship, Copyright Ownership and the Exploitation of Creative Works in the Netherlands, Germany and the United States* (1994); and RICHARD MCCRACKEN and MADELEINE GILBART, *Buying and Clearing Rights: Print, Broadcast and Multimedia* (1995). *Copyright Bulletin* (quarterly), published by UNESCO, presents current information on worldwide copyright practices.

ALLEN KENT *et al.* (eds.), *Encyclopedia of Library and Information Science,* 35 vol. (1968–83), continued with supplemental volumes, provides comprehensive information on many aspects of publishing. Another valuable reference source for current information is *The Bowker Annual of Library & Book Trade Information.* VITO J. BRENNI, *The Art and History of Book Printing: A Topical Bibliography* (1984), *Book Illustration and Decoration: A Guide to Research* (1980), *Book Printing in Britain and America: A Guide to the Literature and a Directory of Printers* (1983), and *Bookbinding, a Guide to the Literature* (1982), are bibliographical guides for further study.

Book publishing. The literature on book publishing and the culture of print is reviewed in two essays in ROBERT DARNTON, *The Kiss of Lamourette: Reflections in Cultural History* (1990): "What Is the History of Books?," pp. 107–135, and "First Steps Toward a History of Reading," pp. 154–187. The impact of literacy on Western modes of thought is discussed in ERIC A. HAVELOCK, *Origins of Western Literacy* (1976); and WALTER J. ONG, *Orality and Literacy: The Technologizing of the Word* (1982, reissued 1991). General works on the history of the book, which examine technical and commercial developments and the impact of the book on European culture, include LUCIEN FEBVRE and HENRI-JEAN MARTIN, *The Coming of the Book: The Impact of Printing, 1450–1800* (1976, reissued 1990; originally published in French, 1958); and ELIZABETH EISENSTEIN, *The Printing Press as an Agent of Change,* 2 vols. (1979). The last has been influential but also controversial; a penetrating critique is ANTHONY T.

GRAFTON, "The Importance of Being Printed," *Journal of Interdisciplinary History*, 11(2):265–286 (Autumn 1980). Other discussions of the history of the book include RICHARD D. ALTICK, *The English Common Reader: A Social History of the Mass Reading Public, 1800–1900* (1957, reissued 1983); ROGER CHARTIER, *The Cultural Uses of Print in Early Modern France*, trans. from French (1987); ROBERT DARNTON, *The Literary Underground of the Old Regime* (1982), and *The Forbidden Bestsellers of Pre-Revolutionary France* (1995); CARLA HESSE, *Publishing and Cultural Politics in Revolutionary Paris, 1789–1810* (1991); and ALVIN KERNAN, *Samuel Johnson & the Impact of Print* (1989). The impact of electronic media is treated in ALVIN KERNAN, *The Death of Literature* (1990); GEOFFREY NUNBERG, "The Places of Books in the Age of Electronic Reproduction," *Representations*, 42:13–37 (Spring 1993); and GEOFFREY NUNBERG (ed.), *The Future of the Book* (1996); JAMES J. O'DONNELL, *Avatars of the Word: From Papyrus to Cyberspace* (1998, reissued 2000); and FRANCES CAIRNCROSS, *The Death of Distance* (1998, reissued 2001).

The history of the publishing industries of various countries is chronicled in JOHN FEATHER, *The Provincial Book Trade in Eighteenth-Century England* (1985); JOHN FEATHER, *A History of British Publishing* (1988); GARY MARKER, *Publishing, Printing, and the Origins of Intellectual Life in Russia, 1700–1800* (1985); K.S. DUGGAL, *Book Publishing in India* (1980); VINOD KUMAR (ed.), *Book Industry in India: Problems & Prospects* (1980); EDUARD KIMMAN, *Indonesian Publishing: Economic Organizations in a Langganan Society* (1981); S.I.A. KOTEI, *The Book Today in Africa* (1981); GEORGE L. PARKER, *The Beginnings of the Book Trade in Canada* (1985); HELLMUT LEHMANN-HAUPT, *The Book in America: A History of the Making and Selling of Books in the United States*, 2nd rev. ed. (1951; originally published in German, 1937); JOHN TEBBEL, *A History of Book Publishing in the United States*, 4 vol. (1972–81); and DONALD FRANKLIN JOYCE, *Gatekeepers of Black Culture: Black-Owned Book Publishing in the United States, 1817–1981* (1983). Comprehensive information on the history and character of American book publishers is gathered in two reference sources, both prepared by PETER DZWONKOSKI (ed.), *American Literary Publishing Houses, 1638–1899*, 2 vol. (1986), and *American Literary Publishing Houses, 1900–1980: Trade and Paperback* (1986).

Publishing of paperbacks is the subject of ALLEN BILLY CRIDER (ed.), *Mass Market Publishing in America* (1982); KENNETH C. DAVIS, *Two-Bit Culture: The Paperbacking of America* (1984); CLARENCE PETERSEN, *The Bantam Story: Thirty Years of Paperback Publishing*, 2nd rev. ed. (1975); and WILLIAM H. LYLES, *Putting Dell on the Map: A History of the Dell Paperbacks* (1983). Production of special kinds of books is discussed in JOAN LYONS, *Artists' Books: A Critical Anthology and Sourcebook* (1985), an overview of the genre of book art; WALTER W. POWELL, *Getting into Print: The Decision-Making Process in Scholarly Publishing* (1985); INTERNATIONAL CONFERENCE ON SCHOLARLY PUBLISHING, *Proceedings from the 3rd International Conference on Scholarly Publishing* (1983); and ALAN MARSHALL MECKLER, *Micropublishing: A History of Scholarly Micropublishing in America, 1938–1980* (1982).

The following are histories of individual publishing firms, some compiled by the companies themselves: BUTTERWORTHS (FIRM), *Butterworths: Yesterday, Today, Tomorrow* (1977); PHILIP WALLIS, *At the Sign of the Ship: Notes on the House of Longman, 1724–1974* (1974); PETER SUTCLIFFE, *The Oxford University Press: An Informal History* (1978); M.H. BLACK, *Cambridge University Press, 1584–1984* (1984), a definitive history, supplemented by DAVID MCKITTERICK, *Four Hundred Years of University Printing and Publishing in Cambridge, 1584–1984* (1984), an exhibition catalogue; EUGENE EXMAN, *The House of Harper: One Hundred and Fifty Years of Publishing* (1967), with coverage of early U.S. copyright complications; THOMAS BONAVENTURE LAWLER, *Seventy Years of Textbook Publishing: A History of Ginn and Company* (1938); RUSSELL FREEDMAN, *Holiday House: The First Fifty Years* (1985), the history of a publisher of children's books; JOHN HAMMOND MOORE, *Wiley, One Hundred and Seventy Five Years of Publishing* (1982); and PETER SCHWED, *Turning the Pages: An Insider's Story of Simon & Schuster, 1924–1984* (1984).

Marketing aspects are emphasized in CHARLES LEE, *The Hidden Public: The Story of the Book-of-the-Month Club* (1958, reprinted 1973), a cultural and business history; WILLIAM M. CHILDS and DONALD E. MCNEIL (eds.), *American Books Abroad: Toward a National Policy* (1986), with information on cultural diplomacy; ALBERTO E. AUGSBURGER, *The Latin American Book Market: Problems and Prospects* (1981); and WILLIAM E. FREEMAN, *Soviet Book Exports, 1973–82* (1984), a research report published by the U.S. Information Agency.

Newspaper publishing. General accounts of the world press are offered in FRANCIS WILLIAMS, *The Right to Know: The Rise of the World Press* (1969); JOHN C. MERRILL, CARTER R. BRYAN, and MARVIN ALISKY, *The Foreign Press: A Survey of the World's Journalism* (1970), concentrating on newspapers but also containing some data on magazines; WILLIAM LUDLOW CHENERY, *Freedom of the Press* (1955, reprinted 1977); *World Communications: A 200-Country Survey of Press, Radio, Television, and Film*, 5th ed. (1975); ANTHONY SMITH, *The Newspaper: An International History* (1979); ANTHONY SMITH (ed.), *Newspapers and Democracy: International Essays on a Changing Medium* (1980); CYRIL BAINBRIDGE (ed.), *One Hundred Years of Journalism: Social Aspects of the Press* (1984); OLIVER BOYD-BARRETT and TERHI RANTANEN (eds.), *The Globalization of News* (1998); and BEN H. BAGDIKIAN, *The Media Monopoly*, 6th ed. (2000). Business aspects are discussed in W. PARKMAN RANKIN, *The Practice of Newspaper Management* (1986); and BENJAMIN M. COMPAINE, *The Newspaper Industry in the 1980s: An Assessment of Economics and Technology* (1980).

Newspaper publishing in Britain is discussed in MICHAEL HARRIS and ALAN LEE (eds.), *The Press in English Society from the Seventeenth to Nineteenth Centuries* (1986); LUCY BROWN, *Victorian News and Newspapers* (1985); GRAHAM STOREY, *Reuters' Century, 1851–1951* (1951, reprinted 1969), a history including information on important U.S. agencies; JAMES CURRAN, *The British Press: A Manifesto* (1978); SIMON JENKINS, *Newspapers: The Power and the Money* (1979), and *The Market for Glory: Fleet Street Ownership in the Twentieth Century* (1986); ALASTAIR HETHERINGTON, *News, Newspapers, and Television* (1985); and DAVID GOODHART and PATRICK WINTOUR, *Eddie Shah and the Newspaper Revolution* (1986), an account of the first electronically produced national newspaper.

The press of the United States is analyzed in BENJAMIN M. COMPAINE *et al., Who Owns the Media?: Concentration of Ownership in the Mass Communications Industry*, 2nd rev. ed. (1982); LOREN GHIGLIONE (ed.), *The Buying and Selling of America's Newspapers* (1984); PETER BENJAMINSON, *Death in the Afternoon: America's Newspaper Giants Struggle for Survival* (1984); RICHARD KLUGER, *The Paper: The Life and Death of the New York Herald Tribune* (1986); MARILYN MCADAMS SIBLEY, *Lone Stars and State Gazettes: Texas Newspapers Before the Civil War* (1983); and DANIEL F. LITTLEFIELD, JR., and JAMES W. PARINS, *American Indian and Alaska Native Newspapers and Periodicals*, 3 vol. (1984–86).

SUSAN GOLDENBERG, *The Thomson Empire* (1984), is a business history of one of the largest Canadian newspaper corporations. LES CARLYON, *Paper Chase: The Press Under Investigation* (1982), is a study of newspaper publishing in Australia. The press of Third World countries is the subject of E. LLOYD SOMMERLAD, *The Press in Developing Countries* (1966); and JOHN A. LENT (ed.), *Newspapers in Asia: Contemporary Trends and Problems* (1982).

Magazine publishing. RUARI MCLEAN, *Magazine Design* (1969), presents a collection of the covers of famous American and European magazines. Studies of magazine publishing in various individual countries include, on Great Britain, CYNTHIA L. WHITE, *Women's Magazines, 1693–1968* (1970), and *The Women's Periodical Press in Britain, 1946–1976* (1977); and ALVIN SULLIVAN (ed.), *British Literary Magazines*, 4 vol. (1983–86); on the United States, THEODORE PETERSON, *Magazines in the Twentieth Century*, 2nd ed. (1964); WALTER C. DANIEL, *Black Journals in the United States* (1982); JAMES P. DANKY (ed.), *Native American Periodicals and Newspapers, 1828–1982: Bibliography, Publishing Record, and Holdings* (1984); EDWARD E. CHIELENS (ed.), *American Literary Magazines: The Eighteenth and Nineteenth Centuries* (1986); and JOANNE P. SHARP, *Condensing the Cold War: Reader's Digest and American Identity* (2000); on Canada, NOEL ROBERT BARBOUR, *Those Amazing People!: The Story of the Canadian Magazine Industry, 1778–1967* (1982); and, on Germany, ERNST BEHLER, *Die Zeitschriften der Bruder Schlegel: Ein Beitrag zur Geschichte der deutschen Romantik* (1983).

Scholarly journals are discussed in E.C. SLATER, *Biochimica et Biophysica Acta: The Story of a Biochemical Journal* (1986); JILL LAMBERT, *Scientific and Technical Journals* (1985); and IRVING LOUIS HOROWITZ, *Communicating Ideas: the Politics of Scholarly Publishing*, 2nd expanded ed. (1991). MICHAEL L. COOK, *Mystery, Detective, and Espionage Magazines* (1983), describes more than 400 American, British, and Canadian magazines of the genre, with brief listings for several other countries.

Business aspects of magazine publishing are the subject of J. WILLIAM CLICK and RUSSELL N. BAIRD, *Magazine Editing and Production*, 4th ed. (1986); BENJAMIN M. COMPAINE, *The Business of Consumer Magazines* (1982); and W. PARKMAN RANKIN and EUGENE SAUVE WAGGAMAN, JR., *Business Management of General Consumer Magazines*, 2nd ed. (1984). Current coverage is found in *Folio: The Magazine for Magazine Management* (monthly, with special issues).

(G.U./P.U./D.H.T./Ed.)

Puppetry

A puppet is an inanimate object moved by human agency in some kind of theatrical show, and the puppet theatre includes any kind of theatrical show that is presented through the medium of puppets. These definitions are wide enough to include an enormous variety of shows and an enormous variety of puppet types, but they do exclude certain related activities and figures. A doll, for instance, is not a puppet, and a girl playing with her doll as if it were a living baby is not giving a puppet show; but, if before an audience of her mother and father she makes the doll walk along the top of a table and act the part of a baby, she is then presenting a primitive puppet show. Similarly, automaton figures moved by clockwork that appear when a clock strikes are not puppets, and such elaborate displays of automatons as those that perform at the cathedral clock in Strasbourg, Fr., or the town hall clock in Munich, Ger., must be excluded from consideration.

Puppet shows seem to have existed in almost all civilizations and in almost all periods. In Europe, written records of them go back to the 5th century BC (*e.g.*, the *Symposium* of the Greek historian Xenophon). Written records in other civilizations are less ancient, but in China, India, Java, and elsewhere in Asia there are ancient traditions of puppet theatre, the origins of which cannot now be determined. Among the American Indians, there are traditions of puppetlike figures used in ritual magic. In Africa, records of puppets are meagre, but the mask is an important feature in almost all African magical ceremonies, and the dividing line between the puppet and the masked actor, as will be seen, is not always easily drawn. It may certainly be said that puppet theatre has everywhere antedated written drama and, indeed, writing of any kind. It represents one of the most primitive instincts of the human race.

This article discusses the various types of puppets as well as historical and contemporary styles of puppet theatre around the world. Some specific national styles of puppetry are treated in the articles EAST ASIAN ARTS and SOUTHEAST ASIAN ARTS.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, section 622, and the Index.

The article is divided into the following sections:

Character of puppet theatre 450
Types of puppets 451
  Hand or glove puppets
  Rod puppets
  Marionettes or string puppets
  Flat figures
  Shadow figures
  Other types
Styles of puppet theatre 454
Puppetry in the contemporary world 456
Bibliography 457

CHARACTER OF PUPPET THEATRE

It may well be asked why such an artificial and often complicated form of dramatic art should possess a universal appeal. The claim has, indeed, been made that puppet theatre is the most ancient form of theatre, the origin of the drama itself. Claims of this nature cannot be substantiated, nor can they be refuted; it is improbable that all human dramatic forms were directly inspired by puppets, but it seems certain that from a very early period in man's development puppet theatre and human theatre grew side by side, each perhaps influencing the other. Both find their origins in sympathetic magic, in fertility rituals, in the human instinct to act out that which one wishes to take place in reality. As it has developed, these magical origins of the puppet theatre have been forgotten, to be replaced by a mere childlike sense of wonder or by more sophisticated theories of art and drama, but the appeal of the puppet even for modern audiences lies nearer a primitive sense of magic than most spectators realize.

The appeal of puppetry

Granted the common origin of human and puppet theatre, one may still wonder about the particular features of puppet theatre that have given it its special appeal and that have ensured its survival over so many centuries. It is not, for instance, simpler to perform than human theatre; it is more complicated, less direct, and more expensive in time and labour to create. Once a show has been created, however, it can provide the advantage of economy in personnel and of portability; one man can carry a whole theatre (of certain types of puppet) on his back, and a cast of puppet actors will survive almost indefinitely. These are clear advantages, but it would be a mistake to imagine that they can explain the whole popularity of puppet theatre. They do not apply to every kind of puppet—some puppets need two or even three manipulators for each figure, and many puppets need one manipulator for each figure. The company employed by a major puppet theatre, whether it be a traditional puppet theatre from Japan or a modern one from eastern Europe, will not be fewer than for an equivalent human theatre. The appeal of the puppet must be sought at a deeper level.

The essence of a puppet is its impersonality. It is a type rather than a person. It shares this characteristic with masked actors or with actors whose makeup is so heavy that it constitutes a mask. Thus, the puppets have an affinity with the stock characters of ancient Greek and Roman drama, with the masked characters of the Renaissance commedia dell'arte, with the circus clown, with the ballerina, with the mummers, and with the witch doctor and the priest.

In an impersonal theatre, where the projection of an actor's personality is lacking, the essential rapport between the player and his audience must be established by other means. The audience must work harder. The spectators must no longer be mere spectators; they must bring their sympathetic imagination to bear and project upon the impersonal mask of the player the emotions of the drama. Spectators at a puppet show will often swear that they saw the expression of a puppet change. They saw nothing of the kind; but they were so wrapped up in the passion of the piece that their imaginations lent to the puppets their own fears and laughter and tears. The union between the actor and the audience is the very heart and soul of the theatre, and this union is possible in a special way, indeed in a specially heightened way, when the actor is a puppet.

The unreality and universality of puppets

The impersonality of the puppet carries other characteristics. There is the sense of unreality. In the traditional English Punch-and-Judy puppet shows, for instance, no one minds when Punch throws the Baby out of the window or beats Judy until she is dead; everyone knows that it is not real and laughs at things that would horrify if they were enacted by human actors. Psychologists agree that the effect is cathartic—one's innate aggressive instincts are released through the medium of these little inanimate figures.

The puppet also carries a sense of universality. This, too, springs from its impersonality. A puppet Charlemagne in a Sicilian puppet theatre is not merely an 8th-century Frankish king but a symbol of royal nobility; and the leader of his rear guard dying on the pass of Roncesvalles is not merely a petty knight ambushed in a skirmish but a type representing heroism and chivalry. Similarly, in the Javanese puppet theatre, a grotesque giant is a personification of the destructive principle, while an elegantly elongated local deity is a personification of the construc-



An English Punch-and-Judy show. Detail from "Punch or May Day," oil painting by Benjamin Robert Haydon, 1829. In the Tate Gallery, London.

By courtesy of the trustees of the Tate Gallery, London; photograph, A.C. Cooper Ltd.

tive principle. Here the puppet theatre reveals its close relationship with the whole spirit of folklore and legend.

The puppet achieves its elemental qualities of impersonality, unreality, and universality through the stylizations imposed upon it by its own limitations. It is a mistake to imagine that the more lifelike or natural a puppet can be, the more effective it is. Indeed, the opposite is often the case. A puppet that merely imitates nature inevitably fails to equal nature; the puppet only justifies itself when it adds something to nature—by selection, by elimination, or by caricature. Some of the most effective puppets are the crudest: at Liège, Belg., for instance, there is a tradition of puppets whose arm and leg movements are not controlled but purely accidental. The Rājasthānī puppets of India have no legs at all. Even less naturalistic are the hunchbacked grotesques of the European tradition, the birdlike profiles of the Indonesian shadow figures, and the intricately shaped leather cutouts of Thailand, but it is precisely among these most highly stylized types of puppets that the art reaches its highest manifestations.

While these puppets that exist furthest from nature can be admired, it cannot be denied that there is a charm and a fascination in the miniaturization of life. Much of the appeal of the puppet theatre has come from the spectators' delight in watching a world in miniature. This can be appreciated best of all in a toy theatre, in which a tiny stage on a drawing room table can be filled with choruses of peasants, troops of banditti, or armies locked in combat, while the scenery behind them depicts far vistas of beetling cliffs or winding rivers.

Manipulation—seen and unseen

And to the appreciation, often instinctive, of these characteristics that mark the puppet theatre, there must be added admiration for the sheer human skill that has gone into the making and manipulation of the figures. The manipulator is usually unseen; his art lies in hiding his art, but the audience is aware of it, and this knowledge adds an element to the dramatic whole. In some kinds of presentation—for instance, in a type of cabaret floor show that became popular in the mid-20th century—the manipulator works in full view of the audience, who may, if they wish, study his methods of manipulation. This is a far cry from the philosophy of the traditional European puppet players of earlier generations, who guarded the secrets of their craft as if they were conjuring tricks. It is, indeed, fair to say that any presentation that deliberately draws attention to the mechanics of how it is done is distorting the art of puppetry, but the realization, nevertheless, of the expertise involved in a performance and some knowledge of the technical means by which it is achieved do add an extra dimension to the appreciation of this difficult and highly skilled art.

TYPES OF PUPPETS

There are many different types of puppets. Each type has its own individual characteristics, and for each there are certain kinds of suitable dramatic material. Certain types have developed only under specific cultural or geographic conditions. The most important types may be classified as follows:

Hand or glove puppets. These have a hollow cloth body that fits over the manipulator's hand; his fingers fit into the head and the arms and give them motion. The figure is seen from the waist upward, and there are normally no legs. The head is usually of wood, papier-mâché, or rubber material, the hands of wood or felt. One of the most common ways to fit the puppet on the hand is for the first finger to go into the head, and the thumb and second finger to go into the arms. There are, however, many variants of this. The "two-fingers-and-thumb" method is used for Punch-type figures; it allows the puppet to pick up and grasp small props very well and is obviously useful when wielding the stick that plays a big part in the show, but it tends to produce a lopsided effect, with one arm higher than the other. The performer normally holds his hands above his head and stands in a narrow booth with an opening just above head height. Most of the traditional puppet folk heroes of Europe are hand puppets; the booth is fairly easily portable, and the entire show can be presented by one person. This is the typical kind of puppet show presented in the open air all over Europe and also found in China. But it need not be limited to one manipulator; large booths with three or four manipulators provide excellent scope for the use of these figures. The virtue of the hand puppet is its agility and quickness; the limitation is small size and ineffective arm gestures.

By courtesy of Pollock's Toy Museum, London; photograph, A.C. Cooper Ltd.



An English toy theatre, 1850. In Pollock's Toy Museum, London.



"Faun" and "Nymph," rod puppets by Richard Teschner, 1914. In the Puppet Theatre Collection, Munich.
By courtesy of the Puppentheatersammlung, Munich

Rod puppets. These figures are also manipulated from below, but they are full-length, supported by a rod running inside the body to the head. Separate thin rods may move the hands and, if necessary, the legs. Figures of this type are traditional on the Indonesian islands of Java and Bali, where they are known as *wayang golek*. In Europe they were for a long time confined to the Rhineland; but in the early 20th century Richard Teschner in Vienna developed the artistic potentialities of this type of figure. In Moscow Nina Efimova carried out similar experimental productions, and these may have inspired the State Central Puppet Theatre in Moscow, directed by Sergey Obraztsov, to develop this type of puppet during the 1930s. After World War II Obraztsov's theatre made many tours, especially in eastern Europe, and a number of puppet theatres using rod puppets were founded as a result. Today the rod puppet is the usual type of figure in the large state-supported puppet theatres of eastern Europe. In a similar movement in the United States, largely inspired by Marjorie Batchelder, the use of rod puppets was greatly developed in school and college theatres, and the hand-rod puppet was found to be of particular value. In this figure the hand passes inside the puppet's body to grasp a short rod to the head, the arms being manipulated by rods in the usual way. One great advantage of this technique is that it permits bending of the body, the manipulator's wrist corresponding to the puppet's waist. Although in general the rod puppet is suitable for slow and dignified types of drama, its potentialities are many and of great variety. It is, however, extravagant in its demands on manipulators, requiring always one person, and sometimes two or three, for each figure on stage.

Prevalence in eastern Europe

Marionettes or string puppets. These are full-length figures controlled from above. Normally they are moved by strings or more often threads, leading from the limbs to a control or crutch held by the manipulator. Movement is imparted to a large extent by tilting or rocking the control, but individual strings are plucked when a decided movement is required. A simple marionette may have nine strings—one to each leg, one to each hand, one to each shoulder, one to each ear (for head movements), and one to the base of the spine (for bowing); but special effects will require special strings that may double or treble this number. The manipulation of a many-stringed marionette is a highly skilled operation. Controls are of two main types—horizontal (or aeroplane) and vertical—and the choice is largely a matter of personal preference.

The string marionette does not seem to have been fully developed until the mid-19th century, when the English marionettist Thomas Holden created a sensation with his ingenious figures and was followed by many imitators. Before that time, the control of marionettes seems to have been by a stout wire to the crown of the head, with subsidiary strings to the hands and feet; even more primitive methods of control may still be observed in certain traditional folk theatres. In Sicily there is an iron rod to the head, another rod to the sword arm, and a string to the other arm; the legs hang free and a distinctive walking gait is imparted to the figures by a twisting and swinging of the main rod; in Antwerp, Belg., there are just rods to the head and to one arm; in Liège there are no hand rods at all, merely one rod to the head. Distinctive forms of marionette control are found in India: in Rājasthān a single string passes from the puppet's head over the manipulator's hand and down to the puppet's waist (a second loop of string is sometimes used to control the arms); in southern India there are marionettes whose weight is supported by strings attached to a ring on the manipulator's head, rods controlling the hands.

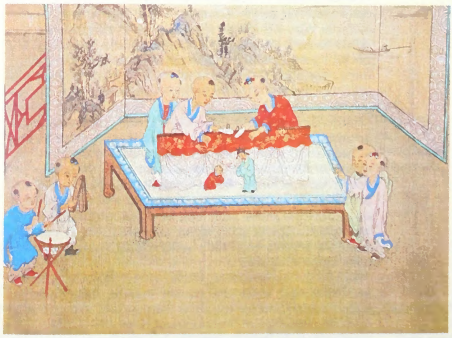
Primitive and developed forms

In European history the marionette represents the most advanced type of puppet; it is capable of imitating almost every human or animal gesture. By the early 20th century, however, there was a danger that it had achieved a sterile naturalism that allowed no further artistic development; some puppeteers found that the control of the marionette figure through strings was too indirect and uncertain to give the firm dramatic effects that they required, and they turned to the rod puppet to achieve this drama. But, in the hands of a sensitive performer, the marionette remains the most delicate, if the most difficult, medium for the puppeteer's art.

Flat figures. Hitherto, all the types of puppets that have been considered have been three-dimensional rounded figures. But there is a whole family of two-dimensional flat figures. Flat figures, worked from above like marionettes, with hinged flaps that could be raised or lowered, were sometimes used for trick transformations; flat jointed figures, operated by piston-type arms attached to revolving wheels below, were used in displays that featured processions. But the greatest use of flat figures was in toy theatres. These seem to have originated in England by a printseller in about 1811 as a kind of theatrical souvenir; one bought engraved sheets of characters and scenery for popular plays of the time, mounted them and cut them out, and performed the play at home. The sheets were sold, in a phrase that has entered the language, for "a penny plain or twopence coloured," the colouring by hand in rapid, vivid strokes of the brush. During a period of about 50 years some 300 plays—all originally performed in the London theatres—were adapted and published for toy-theatre performance in what came to be called the "Juvenile Drama," and a hundred small printsellers were engaged in publishing the plays and the theatrical portraits for tinseling that often went with them. It was always a home activity, never a professional entertainment, and

Toy theatres

By courtesy of the trustees of the British Museum



Chinese children playing with marionettes, detail from "The Hundred Children," a hand scroll of the 17th century. In the British Museum.



A scene from a 19th-century Sicilian puppet theatre enacting the Battle of Roncesvalles. The Sicilian puppets were moved from above by both strings and rods. In the Puppet Theatre Collection, Munich.

By courtesy of the Puppentheatersammlung, Munich

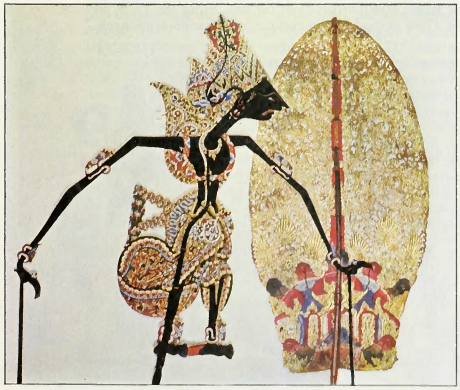
provided one of the most popular and creative fireside activities for Regency and Victorian families. Although few new plays adapted for the toy theatre were issued after the middle of the 19th century, a handful of publishers kept the old stock in print until the 20th century. After World War II this peculiarly English toy was revived. Toy theatres also flourished in other European countries during the 19th century: Germany published many plays; Austria published some extremely impressive model-theatre scenery; in France toy-theatre sheets were issued; in Denmark a line of plays for the toy theatre remains in print. The interest of these toy-theatre plays is largely social, as a form of domestic amusement, and theatrical, as a record of scenery, costume, and even dramatic gesture in a particular period of stage history.

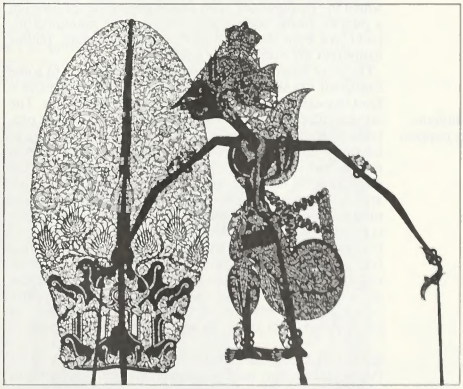
Shadow figures. These are a special type of flat figure, in which the shadow is seen through a translucent screen. They may be cut from leather or some other opaque material, as in the traditional theatres of Java, Bali, and Thailand, in the so-called ombres chinoises (French: literally "Chinese shadows") of 18th-century Europe, and in the art theatres of 19th-century Paris; or they may be cut from coloured fish skins or some other translucent material, as in the traditional theatres of China, India, Turkey, and Greece, and in the recent work of several European theatres. They may be operated by rods from below, as in the Javanese theatres; by rods held at right angles to the screen, as in the Chinese and Greek theatres; or by threads concealed behind the figures, as in the ombres chinoises and in its successor that came to be known as the English galanty show. Shadow figures need not be limited to two dimensions; rounded figures may also be used effectively. A particular type of shadow show that was conceived in terms of film is the silhouette films first made by the German filmmaker Lotte Reiniger in the 1920s; for these films, the screen was placed horizontally, like a tabletop, a light was placed beneath it, the camera was above it, looking downward, and the figures were moved by hand on the screen, being photographed by the stop-action technique. The shadow theatre is a medium of great delicacy, and the insubstantial character of shadow puppets exemplifies all the truest features of puppetry as an art form.

Other types. These five types by no means exhaust every kind of figure or every method of manipulation. There are, for instance, the puppets carried by their manipulators in full view of the audience. The most interesting of these are the Japanese bunraku puppets, which are named for a Japanese puppet master, Uemura Bunrakuken, of the 18th century. These figures, which are one-half to two-thirds life size, may be operated by as many as three manipulators: the chief manipulator controls head movements with one hand by means of strings inside the body, which may raise the eyebrows or swivel the eyes, while using the other hand to move the right arm of the puppet; the second manipulator moves the left arm of the puppet; and the third moves the legs; the coordination of movement between these three artists requires long and devoted training. The magnificent costumes and stylized carving of the bunraku puppets establish them as among the most striking figures of their kind in the world.

Bunraku puppets

Somewhat similar figures, though artistically altogether inferior, are the dummies used by ventriloquists; ventriloquism, as such, has no relation to puppetry, but the ventriloquists' figures, with their ingenious facial movements, are true puppets. The technique of the human actor carrying the puppet actor onto the stage and sometimes speaking for it is one that has been developed a great deal in some experimental puppet theatres in recent years. The human actor is sometimes invisible, through the lighting technique of "black theatre," but is sometimes fully visible. This represents a total rejection of much of





By courtesy of the Puppentheatersammlung, Munich

Indonesian wayang shadow puppet and decoration.
(Left) Spectators may sit on the same side of the screen as the performer, watching the performance as presented by colourful rod puppets. (Right) Connoisseurs of the wayang art usually prefer to sit on the other side of the screen, viewing the performance as a shadow play.



Japanese bunraku theatre; woodblock print by Utashige, 19th century. The puppeteers appear on stage with their puppets; the narrator is shown at the right.
By courtesy of the Puppentheatersammlung, Munich

the traditional thinking about the nature of puppetry, but it has become increasingly accepted.

Another minor form of puppet representation is provided by the jigging puppets, or *marionnettes à la planchette,* that were, during the 18th and 19th centuries, frequently performed at street corners throughout Europe. These small figures were made to dance, more or less accidentally, by the slight variations in the tension of a thread passing through their chests horizontally from the performer's knee to an upright post. Similar were puppets held by short rods projecting from the figures' backs, which were made to dance by bouncing them on a springy board on the end of which the performer sat. The unrehearsed movements of figures like these, when loosely jointed, have a spontaneous vitality that more sophisticated puppets often miss. Another interesting, if elemental, type of puppet, the "scarecrow puppets," or *lileki,* of Slovenia, is constructed from two crossed sticks draped with old clothes; two of these figures are held up on either side of a bench draped with a cloth, under which the manipulator lies. The puppets talk with each other and with a human musician who always joins in the proceedings. The playlets usually end with a fight between the two puppets.

Still another minor puppet form is the finger puppet, in which the manipulator's two fingers constitute the limbs of a puppet, whose body is attached over the manipulator's hand. An even simpler finger puppet is a small, hollow figure that fits over a single finger.

Humans as puppets

The giant figures that process through the streets of some European towns in traditional festivities are puppets of a kind, though they do not normally enact any plays. The same applies to the dragons that are a feature of street processions in China and are to be found in some places in Europe—as, for example, at Tarascon, Fr. Indeed, when a man hides himself within any external frame or mask, the result may be called a puppet. Many of the puppet theatres in Poland today also present plays acted by actors in masks; the Bread and Puppet Theatre in the United States is another example of the same tendency. The divisions between human actors and puppet actors are becoming increasingly blurred; if, in the past, many puppets tried to look and act like humans, today many human actors are trying to look and act like puppets. Clearly, puppetry is being recognized not merely as a particular form of dramatic craft but as one manifestation of total theatre.

STYLES OF PUPPET THEATRE

Puppet theatre has been presented in many diverse styles and for many different kinds of audience. Throughout history, the chief of these has been the performance of folk or traditional plays to popular audiences. The most familiar examples are the puppet shows that have grown up around a number of national or regional comic heroes who appear in a whole repertory of little plays. Pulcinella, for example, was a human character in the Italian commedia dell'arte who began to appear on the puppet stages early in the 17th century; he was carried around Europe by Italian puppet showmen and everywhere became adopted as a new character, hunchbacked and hook-nosed, in the native puppet plays. In France he became Polichinelle, in England Punch, in Russia Petrushka, and so on. In England alone did this wide repertory of plays based on popular legend become limited to the one basic pattern of the Punch-and-Judy show. At about the time of the French Revolution, at the end of the 18th century, a great many local puppet heroes displaced the descendants of Pulcinella throughout Europe: in France it was Guignol, in Germany Kasperl, in the Netherlands Jan Klaassen, in Spain Christovita, and so on. All these characters are glove puppets; many speak through a squeaker in the mouth of the performer that gives a piercing and unhuman timbre to their voices; and all indulge in the fights and other business typical of glove-puppet shows. It is a mistake, however, to regard them all as the same character; they are distinct national types. In Greece the comic puppet hero is Karagkiózis, a shadow puppet, who originally came from Turkey, where he is known as Karagöz.

Dramatic styles

The dramatic material in which these popular puppets play is sometimes biblical, sometimes based on folk tales, and sometimes from heroic sagas. A play on the Passion of Christ, for instance, is still presented by the Théâtre Toone in Brussels: the Faust legend has provided the classic theme for the German puppet theatre, and the Temptation of St. Anthony for the French; and the poems of the Italian Renaissance poet Ariosto, handed on through many popular sources, provide the themes of crusading chivalry for the puppet theatres of Sicily and Liège. More specifically dramatic or literary sources were used by the traveling marionette theatres of England and the United States in the 19th century, when popular plays such as *East Lynne* and *Uncle Tom's Cabin* were played to village audiences almost everywhere.

In Asia the same tradition of partly religious and partly legendary sources provides the repertory for the puppet theatres. The chief of these are the Hindu epics *Rāmāyaṇa* and *Mahābhārata,* which provide the basic plots for the puppet theatres of southern India and of Indonesia.

In distinction to these essentially popular shows, the puppet theatre has, at certain periods of history, provided a highly fashionable entertainment. In England, for instance, Punch's Theatre at Covent Garden, London, directed by Martin Powell from 1711 to 1713, was a popular attraction for high society and received many mentions in the letters and journalism of the day. From the 1770s to the 1790s several Italian companies attracted fashionable

Lalance—Ziolo



Marionnettes à la planchette, or jigging puppets, being operated by a young puppeteer who provides his own accompaniment on his drum and whistle. "Les Petites Marionnettes," engraving from *Le Bon Genre,* published in France in 1820.



Amusement with a simple finger puppet; lithograph by an unknown artist, *c.* 1850.
By courtesy of the Puppentheatersammlung, Munich

audiences and the commendation of Samuel Johnson. In Italy a magnificent puppet theatre was established in the Palace of the Chancellery in Rome in 1708, for which Alessandro Scarlatti, with other eminent composers, composed operas. In Austria-Hungary Josef Haydn was the resident composer of operas for a puppet theatre erected by Prince Esterházy about 1770. In France the ombres chinoises of François-Dominique Seraphin had been established at the Palais-Royal, in the heart of fashionable Paris, by 1781. The Italian scene designer Antonio Bibiena painted the scenery for a marionette theatre belonging to a young Bolognese prince, which performed in London in 1780. Exquisite Venetian marionette theatres preserved in the Bethnal Green Museum in London and the Cooper-Hewitt Museum in New York City indicate the elegance of these fashionable puppet theatres of the 18th century.

During the 18th century English writers began to turn to the puppet theatre as a medium, chiefly for satire. The novelist Henry Fielding presented a satiric puppet show, under the pseudonym of Madame de la Nash, in 1748. The caustic playwright and actor Samuel Foote used puppets to burlesque heroic tragedy in 1758 and sentimental comedy in 1773. In a similar vein, the dramatist Charles Dibdin presented a satiric puppet revue in 1775, and a group of Irish wits ran the Patagonian Theatre in London from 1776 to 1781 with a program of ballad operas and literary burlesques. In France there was a great vogue for the puppet theatre among literary men during the second half of the 19th century. This seems to have begun with the theatre created in 1847 at Nohant by George Sand and her son Maurice, who wrote the plays; well over a hundred plays were produced during a period of 30 years. These productions were purely for guests at the house; they are witty, graceful, and whimsical. Some years later another artistic dilettante conceived the idea of presenting a literary puppet show, but this time for the public; Louis Duranty opened his theatre in the Tuileries Gardens in Paris in 1861, but it lacked popular appeal and did not survive in its original form for very long. The next year Duranty's experiment inspired a group of literary and artistic friends to found the Theatron Erotikon, a tiny private puppet theatre, which only ran for two years, presenting seven plays to invited audiences. The moving spirit, however, was Lemercier de Neuville, who went on to create a personal puppet theatre that played in drawing rooms all over France until nearly the end of the century.

National preferences

All these literary puppet theatres in France had made use of hand puppets, while the English literary puppeteers of the previous century had used marionettes. In 1887 a French artist, Henri Rivière, created a shadow theatre that enjoyed considerable success for a decade at the Chat Noir café in Paris; Rivière was joined by Caran d'Ache and other artists, and the delicacy of the silhouettes was matched by especially composed music and a spoken commentary. Another type of puppet was introduced to Paris in 1888 when Henri Signoret founded the Little Theatre; this theatre used rod puppets mounted on a base that ran on rails below the stage, the movement of the limbs being controlled by strings attached to pedals. The plays presented were pieces by classic authors—Cervantes, Aristophanes, Shakespeare—and new plays by French poets. The Little Theatre, like all the 19th-century French literary puppet theatres, performed infrequently to small audiences in a bohemian milieu; as a movement, this literary enthusiasm for the puppet theatre had little popular influence, but it served as a witness to the potential qualities of puppet theatre.

The puppet theatre in Japan entered literature with the plays of Chikamatsu Monzaemon (1653–1725). This writer, known as the Shakespeare of Japan, took the form of the existing crude Japanese puppet dramas and developed it into a great art form with over a hundred pieces, many of which remain in the repertoire of the bunraku theatre today. In this form of theatre the text, or *jōruri*, is chanted by a *tayū* who is accompanied by a musician on a three-stringed instrument called a samisen.

20th-century styles

In Europe the art-puppet movement was continued into the 20th century by writers and artists associated with the Bauhaus, the highly influential German school of design, which advocated a "total" or "organic" theatre. One of its most illustrious teachers, the Swiss painter Paul Klee, created figures of great interest for a home puppet theatre, and others designed marionettes that reflected the ideas of Cubism. The eminent English man of the theatre Gordon Craig campaigned vigorously for the puppet as a medium for the thoughts of the artist. Between World Wars I and II and through the 1950s and '60s, a number of artists endeavoured in difficult economic conditions to demonstrate that puppets could present entertainment of high artistic quality for adult audiences. The marionettes of the Art Puppet Theatre in Munich, for instance, were striking exemplars of the German tradition in deeply cut wood carving. In Austria the Salzburg Marionette Theatre specializes in Mozart operas and has achieved a high degree of naturalism and technical expertise. In Czechoslovakia—a country with a fine puppet tradition—Josef Skupa's marionette theatre presented musical turns interspersed with witty satiric sketches introducing the two characters who gave their names to the theatre: Hurvínek, a precocious boy, and Špejbl, his slow-witted father. In France the prominent artists who designed for Les Comédiens de Bois included the painter Fernand Léger. Yves Joly stripped the art of the puppet to its bare essentials by performing hand puppet acts with his bare hands, without any puppets. The same effect was achieved by the Russian

By courtesy of Felix Klee, © Cosmopress, Geneva, and permission of S.P.A.D.E.M. 1971, by French Reproduction Rights, Inc.; photograph, Bil Baird Collection



Hand puppets made by Paul Klee (1879–1940); the centre puppet is a self-portrait. In the collection of Felix Klee.

puppeteer Sergey Obraztsov with a performance of charm and wit that was quite different from those of the great rod-puppet theatre that he founded. In England the fine craftsman Waldo Lanchester played an important part in the marionette revival; his productions included the early madrigal opera *L'Amfiparnaso.* Jan Bussell, with the Hogarth Puppets, achieved an international reputation with his marionette ballets and light operas. In London a permanent marionette theatre, the Little Angel, was opened by John Wright in 1961. Other permanent puppet theatres have been established in Birmingham and Norwich and at Biggar near Edinburgh.

In the United States the artistic puppet revival was largely inspired by Ellen Van Volkenburg at the Chicago Little Theatre with productions that included *A Midsummer Night's Dream* in 1916. She later directed plays for Tony Sarg, who became the most important influence in American puppetry, with such large-scale marionette plays as *Rip Van Winkle, The Rose and the Ring,* and *Alice in Wonderland.* A small group, the Yale Puppeteers, created a theatre in Hollywood, the Turnabout Theatre, that combined human and puppet stages at opposite ends of the auditorium and attracted fashionable audiences for its songs and sketches from 1941 to 1956. Bil Baird ran a puppet theatre in Greenwich Village, New York City, for some years from 1967 and made a great contribution to every aspect of puppetry. But the lack of the kind of state subsidy that is taken for granted in eastern Europe has made the development of large touring puppet theatres impossible in the United States. Professional puppetry there has developed in three main ways: in large, commercially supported productions for television (see below); in socially involved groups, such as the Bread and Puppet Theatre, which uses giant puppets to carry a political or idealistic message; and—at the other end of the scale—as a medium for intimate tabletop presentations by artists such as Bruce Schwartz, who makes no attempt to conceal himself as he handles a single figure with great delicacy.

Meanwhile, the puppet theatre was continuing on a less exalted plane to demonstrate that it could still provide enjoyable entertainment for popular audiences. From the 1870s a number of English marionette companies had developed the technique of their art to an extraordinarily high level, and their influence was widely spread through Europe, Asia, and America by a series of world tours. Their performances made a great feature of trick effects: there was the dissecting skeleton, whose limbs came apart and then came together again; the Grand Turk, whose arms and legs dropped off to turn into a brood of children while his body turned into their mother; the crinolined lady, who turned into a balloon; the Scaramouch, with three heads; and a host of jugglers and acrobats. The last of the great touring marionette theatres in this tradition was the Theatre of the Little Ones of Vittorio Podrecca, which introduced the marionette pianist and the soprano with heaving bosom that have been widely copied ever since.

During the 20th century there has been an increasing tendency to regard the puppet theatre as an entertainment for children. One of the first people to encourage this development was Count Franz Pocci, a Bavarian court official of the mid-19th century, who wrote a large number of children's plays for the traditional marionette theatre of Papa Schmid in Munich. Important also was Max Jacob, who developed the traditional folk repertoire of the German Kasperltheater, between the 1920s and '50s, into something more suited to modern ideas of what befits children's entertainment. Almost all contemporary puppeteers have created programs for audiences of children.

In this survey of the various styles of puppet theatre in different countries and in different cultures, there are certain features that are common to many otherwise differing forms. In many forms of puppet theatre, for instance, the dialogue is not conducted as if through the mouths of the puppets, but instead the story is recited or explained by a person who stands outside the puppet stage to serve as a link with the audience. This technique was certainly in use in England in Elizabethan times, when the "interpreter" of the puppets is frequently referred to; this character is well illustrated in Ben Jonson's *Bartholomew Fair,* in which one of the puppets leans out of the booth (they were hand puppets) and hits the interpreter on the head because it does not like the way he is telling the story. The same technique of the reciter is found in the Japanese bunraku theatre, in which the chanter contributes enormously to the full effect and is, indeed, regarded as one of the stars of the company. The technique is also found in the French shadow theatre at the Chat Noir, and its imitators and successors, which depended to a great extent upon the chansonnier. Many recent puppet productions utilize this technique as well. Elsewhere, such as in traditional puppet theatres of Java, Greece, and Sicily, all the speaking is done by the manipulator. The plays consist of a mixture of narration and dialogue, and, though the performer's voice will certainly vary for the different characters, the whole inevitably acquires a certain unity that is one of the most precious attributes of the puppet theatre.

The "interpreter"

Musical accompaniment is an important feature of many puppet shows. The gamelan gong and cymbal orchestra that accompanies a Javanese wayang performance is an essential part of the show; it establishes the mood, provides the cadence of the puppets' movements, and gives respite between major actions. Similarly, the Japanese samisen supports and complements the chanter. In the operatic puppet theatre of 18th-century Rome, the refined musical scores of Scarlatti and the stilted conventions and long-held gestures of the opera of that time must have been admirably matched by the slow, contrived but strangely impressive movements of the rod puppets. When in 1662 Samuel Pepys visited the first theatre to present Punch in England, he noted in his famous diary that "here among the fiddlers I first saw a dulcimer played on with sticks knocking of the strings, and it is very pretty." Even an old-fashioned Punch-and-Judy show had a drum and panpipes as an overture. Puppets without music can seem rather bald. At one time the gramophone was used extensively by puppeteers, and more recently the tape recorder has provided a more adaptable means of accompanying a puppet performance with music and other sound effects.

Music and lighting

Lighting effects can also play an important part in a puppet production. The flickering oil lamp of the Javanese wayang enhances the shadows of the figures on the screen; as long ago as 1781, the scene painter Philip James de Loutherbourg used a large model theatre called the Eidophusikon to demonstrate the range of lighting effects that could be achieved with lamps. Modern methods using ultraviolet lighting have enabled directors of puppet productions to achieve astonishing and spectacular effects.

PUPPETRY IN THE CONTEMPORARY WORLD

The puppet theatre in the contemporary world faces great difficulties and great opportunities. The audiences for the traditional folk theatres have almost disappeared. Punch and Judy on the English beaches and Guignol in the parks of Paris still draw a crowd, but the indoor theatres that once attracted humble audiences survive with difficulty, usually with the aid of a sympathetic town council or a local museum. Puppets are increasingly regarded as an entertainment only for children. They certainly do provide a kind of theatre to which children respond with enthusiasm, and, in the general development of children's theatre, the puppet theatre has a part to play. Some puppeteers are happy to play only for children. But others are eager to play also on an adult level; and, for these, audiences are few. No professional puppet theatre can exist in the West on a purely adult repertoire. Even those theatres that do play for children face great economic difficulties from the small size of audience to which puppets can play and from the modest admission fees that can be charged to children. If a few companies do continue to present performances of quality, this is a tribute to their dedication to their art.

Opportunities and problems

There are some possible means of performance beyond the children's theatre. There are cruise ships and nightclubs, which provide an opportunity for short turns but obviously no scope for serious drama. And there is television. At first sight, television would seem an ideal medium for puppetry, and many puppet shows have in fact appeared on it, but initially the great possibilities that it seemed to offer were not fully realized. A straight transference of

a puppet production to the television screen proved not to be effective, and puppet acts on television were often limited to short presentations on variety shows. Several programs designed for television, sometimes combining puppets with human performers, did, however, gain great success. In England, for instance, Muffin the Mule and his animal friends, manipulated by Ann Hogarth, appeared from 1946 on the top of a piano at which Annette Mills played and sang. In the United States a series featuring the Kuklapolitans, created by Burr Tillstrom, began airing in 1947; Kukla, a small boy, had a host of friends, including Ollie the Dragon, who exchanged repartee with Fran, a human actress standing outside the booth. In 1969, puppets were introduced on the educational program "Sesame Street"; these were created by Jim Henson and represented a type of figure that reached its full potential in "The Muppet Show," which attracted enormous audiences in more than a hundred countries between 1976 and 1981. Henson went on to create puppet films in which fantastic puppet characters were manipulated by radio-controlled mechanisms of extraordinary ingenuity. Another type of television puppetry could be seen in "Spitting Image," a program introduced in 1984 with caricatured puppets designed by Roger Law and Peter Fluck. It consisted of satiric sketches, originally of English politicians and personalities, and represented a revival of the 18th-century tradition of adult satiric puppet theatre.



Jim Henson (centre) and other puppeteers on the set of "The Muppet Show" watching their performance on video monitors as the television program is taped.

© Henson Associates, 1978

State subsidies

The economic difficulties facing puppet companies in western Europe and the United States have been lifted in eastern Europe and China, where the state provides generous subsidies for puppet theatres. Whereas in the West a puppet theatre is lucky if it can afford to pay a company of 5 or 6 performers, it is not unusual for a puppet theatre in the East to employ 50 or 60 performers, artists, and technicians. Interest in the puppet theatre has surged in eastern Europe since World War II, and, while the state supports these theatres, there is very little sign of any direct political propaganda in their programs. The results of all this aid have often been impressive in the sheer weight of numbers and scenic effects, and the productions have often been experimental and imaginative. Mere size, however, does not necessarily guarantee artistic success, and some of the best of these theatres would seem to feel a lack of confidence in their medium by their restless searching for new methods of presentation through "black theatre," mask theatre, and other techniques.

Puppetry in schools

A great feature of education during the 20th century has been the introduction of puppet making into schools as a craft activity. The difficulties facing professional puppet theatre are entirely absent here, and a puppet performance can synthesize many of the arts and skills of a group of children in making, costuming, and manipulating puppets, in writing plays for them, and in acting them. When this activity was first introduced, undue importance was often placed upon the mere construction of figures according to certain set methods and upon the painstaking preparation of a showing, so that the creative release of the performance was long delayed and sometimes never reached. Today the tendency is to create puppets quickly from scrap materials or from natural objects and to perform them impromptu, without rehearsal, as a form of dramatic self-expression. It is from such activities that the therapeutic potentialities of puppets have been utilized by psychiatrists working with disturbed children.

The future of the puppet theatre will certainly be greatly influenced by the cross-fertilization between different traditions in puppetry that will result from puppeteers meeting each other and seeing each other's performances at international festivals of the puppet theatre. These festivals now take place almost every year and are usually sponsored by UNIMA, the Union Internationale de la Marionnette, an international society of puppeteers. Originally founded in 1929 and reconstituted in 1957, UNIMA has members in more than 50 countries and provides a common meeting ground for professional and amateur performers, critics, and enthusiasts. In the meantime traditional styles of puppetry will not be neglected. Many countries now boast national organizations—the Puppeteers of America in the United States and Canada or The Puppet Centre in Great Britain, for example—which promote the differing local traditions of this minor but fascinating art.

By courtesy of WTTW-TV, Chicago—Public Broadcasting Service



Fran Allison with Kukla and Ollie, two puppets created by Burr Tillstrom for the series "Kukla, Fran and Ollie."

BIBLIOGRAPHY. A.R. PHILPOTT, *Dictionary of Puppetry* (1969), a brief but comprehensive guide to every aspect of the subject; CHARLES MAGNIN, *Histoire des marionnettes en Europe: depuis l'antiquité jusqu'à nos jours*, rev. ed. (1862, reprinted 1981), the classic history, not yet superseded; BIL BAIRD, *The Art of the Puppet* (1965, reprinted 1973), a magnificently illustrated general survey; MARGARETA NICULESCU (ed.), *The Puppet Theatre of the Modern World* (1967; originally published in German, 1965), an international presentation sponsored by UNIMA; GEORGE SPEAIGHT, *The History of the English Puppet Theatre*, rev. ed. (1990), an exploration of European puppets up to the 17th century, and *The History of the English Toy Theatre*, rev. ed. (1969); PAUL MCPHARLIN, *The Puppet Theatre in America: A History, 1524–1948*, rev. ed. (1969), with a supplement covering developments since 1948 by MARJORIE BATCHELDER MCPHARLIN, including a selected bibliography; JOHN WRIGHT, *Rod, Shadow, and Glove: Puppets from the Little Angel Theatre* (1986), a practical guide on puppet making; DAVID CURRELL, *The Complete Book of Puppet Theatre*, rev. ed. (1985), with special emphasis on educational uses of puppetry; and ANN HOGARTH and JAN BUSSELL, *Fanfare for the Puppets* (1985), an account of a lifetime of experience as performers.

(G.St.)

Radar

Radar is an electromagnetic sensor used for detecting, locating, tracking, and identifying objects of various kinds at considerable distances. It operates by transmitting electromagnetic energy toward objects, commonly referred to as targets, and observing the echoes returned from them. The targets may be aircraft, ships, spacecraft, automotive vehicles, and astronomical bodies, or even birds, insects, and raindrops. Radar can not only determine the presence, location, and velocity of such objects but can sometimes obtain their size and shape as well. What distinguishes radar from optical and infrared sensing devices is its ability to detect faraway objects under all weather conditions and to determine their range with precision.

Radar is an "active" sensing device in that it has its own source of illumination (a transmitter) for locating targets. In certain respects, it resembles active sonar, which is used chiefly for detecting submarines and other objects underwater; however, the acoustic waves of sonar propagate differently from electromagnetic waves and have different properties. Radar typically operates in the microwave region of the electromagnetic spectrum—namely, at frequencies extending from about 400 megahertz (MHz) to 40 gigahertz (GHz). It has, however, been used at lower frequencies for long-range applications (frequencies as low as several megahertz, which is the HF, or short-wave, band) and at optical and infrared frequencies (those of laser radar, or lidar). The circuit components and other hardware of radar systems vary with the frequency used, and systems range in size from those small enough to fit in the palm of the hand to those so enormous as to take up several football fields. These differences notwithstanding, the basic principles of operation of all radar systems remain the same.

Radar underwent rapid development during the 1930s and '40s to meet the needs of the military. It is still widely employed by the armed forces, and many advances in radar technology have in fact been subsidized by the military. At the same time, radar has found an increasing number of important civilian applications, notably air traffic control, remote sensing of the environment, aircraft and ship navigation, speed measurement for industrial applications and for law enforcement, space surveillance, and planetary observation.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 735, 736, and 738, and the *Index.*

This article is divided into the following sections:

Fundamentals of radar 458
- Basic principle 458
- Pulse radar 458
- Component parts of a radar system 459
- Target information obtained by radar 459
- Target recognition 461

Development of radar 461
- Early experiments 461
- First military radars 461
- Advances during World War II 462
- Radar technology since the mid-1940s 462

Radar subsystems 462
- Antennas 462
- Transmitters 463
- Receivers 463
- Signal and data processors 464
- Displays 464

Types of radar 464

Factors affecting radar performance 465
- Transmitter power and antenna size 465
- Receiver noise 466
- Target size 466
- Clutter 466
- Atmospheric effects 466
- Interference 466
- Electronic countermeasures 466

Major applications of radar 467
- Areas of application 467
- Radar applications by frequency 467

Examples of radar systems 468
- Airport surveillance radar 468
- Doppler weather radar 468
- Airborne combat radar 469
- Ballistic missile detection and satellite surveillance radar 469
- Ground-probing radar 470
- Over-the-horizon radar 470

Bibliography 470

FUNDAMENTALS OF RADAR

Basic principle. A typical radar operates by radiating a narrow beam of electromagnetic energy into space from an antenna (Figure 1). The narrow antenna beam is scanned to search a region where targets are expected. When a target is illuminated by the beam, it intercepts some of the radiated energy and reflects a portion back toward the radar system. Since most radar systems do not transmit and receive at the same time, a single antenna can be used on a time-shared basis for both transmitting and receiving.

A receiver attached to the output element of the antenna extracts the desired reflected signals and (ideally) rejects those that are of no interest. For example, a signal of interest might be the echo from an aircraft. Signals that are not of interest might be echoes from the ground or rain, which can mask and interfere with the detection of the desired echo from the aircraft. The radar measures the location of the target in range and angular direction. Range is determined by measuring the total time it takes for the radar signal to make the round trip to the target and back (see below). The angular direction of a target is usually found from the direction in which the antenna points at the time the echo signal is received. Through measurement of the location of a target at successive instants of time, its track can be determined. Once this information has been established, the target's location at a time in the future can be predicted. In many surveillance radar applications, the target is not considered to be "detected" until its track has been established.

Measuring target location in terms of range and angle

Pulse radar. The most common type of radar signal consists of a repetitive train of short-duration pulses. Figure 2 is a simple representation of a sine-wave pulse that might be generated by the transmitter of a medium-range radar designed for aircraft detection. The sine wave in the figure represents the variation with time of the output voltage of the transmitter. The numbers given in brackets in the figure are only meant to be illustrative and are not necessarily those of any particular radar. They are, however, similar to what might be expected for a ground-based radar system with a range of about 50 to 60 nautical miles (or 90 to 110 kilometres), such as the kind used for air traffic control at airports. The pulse width is given in the figure as one millionth of a second (one microsecond). It should be noted that the pulse is shown as containing only a few cycles of the sine wave; however, in a radar system having the values indicated, there would be 1,000 cycles within the pulse. In Figure 2 the time between successive pulses is given as one thousandth of a second (one millisecond), which corresponds to a pulse repetition frequency of 1,000 hertz (Hz; cycles per second). The

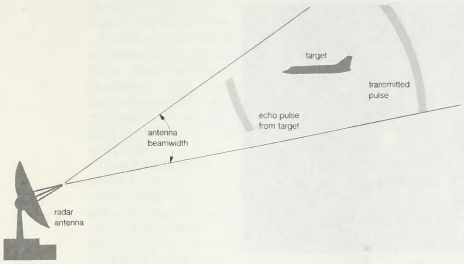


Figure 1: *Principle of radar operation.*
The transmitted pulse has already passed the target, which has reflected a portion of the radiated energy back toward the radar unit.

power of the pulse, called the peak power, is taken here to be 1,000,000 watts (1 megawatt). Since pulse radar does not radiate continually, the average power is much less than the peak power. In this example, the average power is 1,000 watts (1 kilowatt). The average power, rather than the peak power, is the measure of the capability of a radar system. Radars have average powers from a few milliwatts to as much as one or more megawatts, depending on the application.

A weak echo signal from a target might be as low as one trillionth of a watt (10^{-12} watt). In short, the power levels in a radar system can be very large (at the transmitter) and very small (at the receiver).

Another example of the extremes encountered in a radar system is the timing. An air-surveillance radar (one that is used to search for aircraft) might scan its antenna 360 degrees in azimuth in a few seconds, but the pulse width might be about one microsecond in duration. (Some radar pulse widths are 1,000 times smaller—*i.e.*, of nanosecond duration.)

The range to a target is determined by measuring the time that a radar signal takes to travel out to the target and back. Radar waves travel at the same speed as light—roughly 300,000,000 metres per second (or 186,000 miles per second). The range to the target is equal to $cT/2$, where c = velocity of propagation of radar energy, and T = round-trip time as measured by the radar. From this expression, the round-trip travel of the radar signal is at a rate of 150 metres per microsecond. For example, if the time that it takes the signal to travel out to the target and back were measured by the radar to be 600 microseconds (0.0006 second), then the range of the target would be 90 kilometres.

Component parts of a radar system. Figure 3 shows the basic parts of a typical radar system. The transmitter generates the high-power signal that is radiated by the antenna. The antenna is often in the shape of a parabolic reflector, similar in concept to an automobile headlight but much different in construction and size. It also might consist of a collection of individual antennas operating together as a phased-array antenna (see below *Radar subsystems: Antennas*). In a sense, an antenna acts as a "transducer" to couple electromagnetic energy from the transmission line to radiation in space, and vice versa. The duplexer permits alternate transmission and reception with the same antenna; in effect, it is a fast-acting switch that protects the sensitive receiver from the high power of the transmitter.

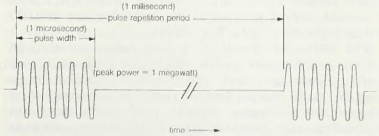


Figure 2: A typical pulse waveform transmitted by radar.

The receiver selects and amplifies the weak radar echoes so that they can be displayed on a television-like screen for the human operator or be processed by a computer. The signal processor separates the signals reflected by the target (*e.g.*, echoes from an aircraft) from unwanted echo signals (the clutter from land, sea, rain, etc.). It is not unusual for these undesired reflections to be much larger than desired target echoes, in some cases more than one million times larger. Large clutter echoes from stationary objects can be differentiated from small echoes from a moving target by noting the shift in the observed frequency produced by the moving target. This phenomenon is called the Doppler frequency shift (see below).

Signal processor

At the output of the receiver a decision is made (either by the human operator or automatically by a computer circuit) as to whether or not a target echo is present. If the output of the receiver is larger than a predetermined value, a target is assumed to be present.

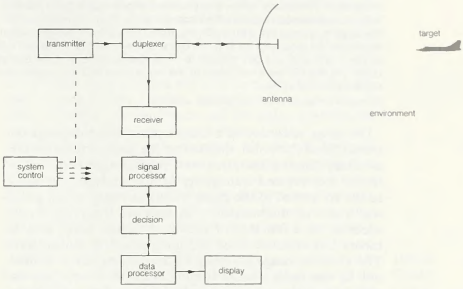


Figure 3: Basic parts of a radar system.

Once it has been decided that a target is present and its location (in range and angle) has been determined, the track of the target can be obtained by measuring the target location at different times. During the early days of radar, target tracking was performed by an operator marking the location of the target "blip" on the face of a cathode-ray tube (CRT) display with a grease pencil. Manual tracking has been largely replaced by automatic electronic tracking, which can process a much greater number of target tracks (many hundreds or even thousands) than can an operator, who can handle only a few simultaneous tracks. Automatic tracking is an example of an operation performed by a data processor.

An example of a radar display is shown in Figure 4. At the left is the unprocessed output of a conventional air-surveillance radar system. This is sometimes called "raw video." At the right is the processed output with the clutter eliminated and only the moving targets displayed.

The type of signal waveform transmitted and the associated received-signal processing in a radar system might be different depending on the type of target involved and the environment in which it is located. An operator can select the parameters of the radar to maximize performance in a particular environment. Alternatively, electronic circuitry in the radar system can automatically analyze the environment (determine which portions are land, sea, or rain) and select the proper transmitted signal, signal processing, and other radar parameters to optimize performance. The box labeled "system control" in Figure 3 is intended to represent this function. The system control also can provide the timing and reference signals needed to permit the various parts of the radar to operate effectively as an integrated system. (Further descriptions of the major parts of a radar system are given below in the section *Radar subsystems.*)

Target information obtained by radar. The ability to measure the range to a target accurately at long distances and to operate under adverse weather conditions are radar's most distinctive attributes. There are no other devices that can compete with radar in the measurement of range.

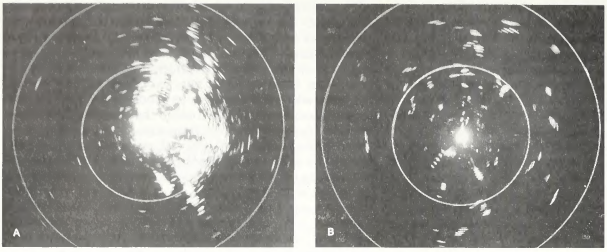


Figure 4: *Plan position indicator (PPI) display for an aircraft-surveillance radar system.* The radar is located at the centre. The radial direction represents range. The inner ring has a radius of 20 nautical miles, and the outer ring a radius of 40 nautical miles. On the left (A) is the unprocessed output of the radar showing large unwanted clutter echoes as well as the targets beyond the clutter. On the right (B) is the processed output, in which clutter is eliminated by means of an MTI. In (B) the camera shutter remained open for five rotations of the antenna (about one minute) to show the movement of the aircraft detected by the radar. (A) and (B) were not taken at the same time, and so targets seen on one are not the same as on the other.

Courtesy of George Linde, U.S. Naval Research Laboratory

The range accuracy of a simple pulse radar depends on the width of the pulse: the shorter the pulse, the better the accuracy. Short pulses, however, require wide bandwidths in the receiver and transmitter (since bandwidth is equal to the reciprocal of the pulse width). A radar with a pulse width of one microsecond can measure the range to an accuracy of a few tens of metres or better. Some special radars can measure to an accuracy of a few centimetres. The ultimate range accuracy of the best radars is limited not by the radar system itself, but rather by the known accuracy of the velocity at which electromagnetic waves travel. (The calculation of range involves the velocity of the electromagnetic energy transmitted as well as the round-trip time.)

Almost all radars use a directive antenna—*i.e.,* one that directs its energy in a narrow beam. The direction of a target can be found from the direction in which the antenna is pointing when the received echo is at a maximum. (There are other more precise means for determining the direction of a target, of which the monopulse method is probably the most important.) A dedicated tracking radar—one that follows automatically a single target so as to determine its trajectory—generally has a narrow symmetrical "pencil" beam. (A typical beamwidth might be about 1 degree.) Such a radar system can determine the location of the target in both azimuth angle and elevation angle. An aircraft-surveillance radar generally employs an antenna that radiates a "fan" beam, one that is narrow in azimuth (about 1 or 2 degrees) and broad in elevation (elevation beamwidths of from 20 to 40 degrees, or more). A fan beam allows only the measurement of the azimuth angle.

Doppler frequency shift

Radar can extract the Doppler frequency shift of the echo produced by a moving target by noting how much the frequency of the received signal differs from the frequency of the signal that was transmitted. (The Doppler frequency shift in radar is similar to the change in audible pitch experienced when listening to a train whistle or the siren of an emergency vehicle when the train or emergency vehicle is moving either toward or away from the listener.) A moving target will cause the frequency of the echo signal to increase if it is approaching the radar or to decrease if it is receding from the radar. For example, if a radar system operates at a frequency of 3,000 megahertz and an aircraft is moving toward it at a speed of 400 knots (740 kilometres per hour), the frequency of the received echo signal will be greater than that of the transmitted signal by about 4.1 kilohertz. The Doppler frequency shift in hertz is equal to $3.4 f_0 v_r$, where f_0 is the radar frequency in gigahertz and v_r is the radial velocity (the rate of change of range) in knots.

Since the Doppler frequency shift is proportional to radial velocity, a radar system that measures such a shift in frequency can provide the radial velocity of a target. The Doppler frequency shift also is used to separate moving targets (such as aircraft) from stationary ones (land or sea clutter) even when the undesired clutter power might be much greater than the power of the echo from the targets. A form of pulse radar that uses the Doppler frequency shift to eliminate stationary clutter is called either a moving-target indication (MTI) radar or a pulse Doppler radar, depending on the particular parameters of the signal waveform.

The above measurements of range, angle, and radial velocity assume that the target is like a point. Actual targets, however, are of finite size and can have distinctive shapes. The range profile of a finite-sized target can be determined if the range resolution of the radar is small compared to the target's size in the range dimension. (The range resolution of a radar, given in units of distance, is a measure of the ability of a radar to separate two closely spaced echoes.) Some radars can have resolutions smaller than one metre, which is quite suitable for determining the radial size and profile of many targets of interest.

The resolution in angle that can be obtained with conventional antennas is poor compared to that which can be obtained in range. It is possible, however, to achieve good resolution in angle, or cross range, by resolving in Doppler frequency (*i.e.,* separating one Doppler frequency from another). If the radar is moving relative to the target (as when the radar unit is on an aircraft and the target is the ground), the Doppler frequency shift will be different for different parts of the target. Thus the Doppler frequency shift can allow the various parts of the target to be resolved. The resolution in cross range derived from the Doppler frequency shift is far better than that achieved with a narrow-beam antenna. It is not unusual for the cross-range resolution obtained from Doppler frequency to be comparable to that obtained in the range dimension.

Synthetic aperture radar

Cross-range resolution obtained from Doppler frequency, along with range resolution, is the basis for synthetic aperture radar (SAR). SAR produces an image of a scene that is similar to, but not identical with, an optical photograph. One should not expect the image seen by radar "eyes" to be the same as that observed by optical ones. Each provides different information. Radar and optical images differ because of the large difference in the frequencies involved; optical frequencies are approximately 100,000 times higher than radar frequencies.

The SAR can operate from long range and through clouds or other atmospheric effects that limit optical and infrared imaging sensors. The resolution of a SAR image can be made independent of range, an advantage over passive optical imaging, where the resolution worsens with increasing range. Synthetic aperture radars that map areas of the Earth's surface with resolutions of a few metres can

provide information about the nature of the terrain and what is on the surface.

A SAR operates on a moving vehicle, such as an aircraft or spacecraft, to image stationary objects or planetary surfaces. Since relative motion is the basis for the Doppler resolution, high resolution (in cross range) also can be accomplished if the radar is stationary and the target is moving. This is called inverse synthetic aperture radar (ISAR). Both the target and the radar can be in motion with ISAR.

Target recognition. Radar can distinguish one kind of target from another (such as a bird from an aircraft), and some systems are able to recognize specific classes of targets (for example, a commercial airliner as opposed to a military jet fighter). Target recognition is accomplished by measuring the size and speed of the target and by observing the target with high resolution in one or more dimensions. Propeller or jet engines modify the radar echo from aircraft and can assist in target recognition. The flapping of the wings of a bird in flight produces a characteristic modulation which can be used to recognize that a bird is present or even to identify one type of bird from another.

DEVELOPMENT OF RADAR

Early experiments. Serious developmental work on radar began in the 1930s, but the basic idea of radar had its origins in the classical experiments on electromagnetic radiation conducted by the German physicist Heinrich Hertz during the late 1880s. Hertz set out to verify experimentally the earlier theoretical work of the Scottish physicist James Clerk Maxwell. Maxwell had formulated the general equations of the electromagnetic field, determining that both light and radio waves are examples of electromagnetic waves governed by the same fundamental laws but having widely different frequencies. Maxwell's work led to the conclusion that radio waves can be reflected from metallic objects and refracted by a dielectric medium just like light waves. Hertz demonstrated these properties in 1888, using radio waves at a wavelength of 66 centimetres (which corresponds to a frequency of about 455 MHz).

Contributions of Maxwell and Hertz

The potential utility of Hertz's work as the basis for the detection of targets of practical interest did not go unnoticed at the time. In 1904 a patent for "an obstacle detector and ship navigation device," based on the principles demonstrated by Hertz, was issued in several countries to Christian Hülsmeyer, a German engineer. Hülsmeyer built his invention and demonstrated it to the German navy, but failed to arouse any interest. There was simply no economic, societal, or military need for radar until the early 1930s, when a long-range military bomber capable of carrying large payloads was developed. This prompted the major countries of the world to look for a means with which to detect the approach of hostile aircraft.

Most of the countries that developed radar prior to World War II first experimented with other methods of aircraft detection. These included listening for the acoustic noise of aircraft engines and detecting the electrical noise from their ignition. Researchers also experimented with infrared sensors. None of these, however, proved effective.

First military radars. During the 1930s, efforts to use radio echo for aircraft detection were initiated independently and almost simultaneously in several countries that were concerned with the prevailing military situation and that already had practical experience with radio technology. The United States, Great Britain, Germany, France, the Soviet Union, Italy, and Japan all began experimenting with radar within about two years of one another and embarked, with varying degrees of motivation and success, on its development for military purposes. Most of these countries had some form of operational radar equipment in military service at the start of World War II in 1939.

The first observation of the radar effect at the U.S. Naval Research Laboratory (NRL) in Washington, D.C., was made in 1922. NRL researchers positioned a radio transmitter on one shore of the Potomac River and a receiver on the other. A ship sailing on the river caused fluctuations in the intensity of the received signals when it passed between the transmitter and receiver. (Today, such a configuration would be called bistatic radar.) In spite of the promising results of this experiment, U.S. Navy officials were unwilling to sponsor further work.

The principle of radar was "rediscovered" at the NRL in 1930 when L.A. Hyland observed that an aircraft flying through the beam of a transmitting antenna caused a fluctuation in the received signal. Although Hyland and his associates at the NRL were enthusiastic about the prospect of detecting targets by radio means and were anxious to pursue its development in earnest, little interest was shown by higher authorities in the navy. Not until it was learned how to use a single antenna for both transmitting and receiving (now termed monostatic radar) was the value of radar for detecting and tracking aircraft and ships fully recognized. Such a system was demonstrated at sea on the battleship USS *New York* in early 1939.

The first radars developed by the U.S. Army were the SCR-268 (at a frequency of 205 MHz) for controlling antiaircraft gunfire and the SCR-270 (at a frequency of 100 MHz) for detecting aircraft. Both of these radars were available at the start of World War II, as was the navy's CXAM shipboard surveillance radar (at a frequency of 200 MHz). It was an SCR-270, one of six available in Hawaii at the time, that detected the approach of Japanese warplanes toward Pearl Harbor, near Honolulu, on Dec. 7, 1941; however, the significance of the radar observations was not appreciated until bombs began to fall.

Britain commenced radar research for aircraft detection in 1935. The British government encouraged engineers to proceed rapidly because they were quite concerned about the growing possibility of war. By September 1938, the first British radar system, the Chain Home, went into 24-hour operation and remained operational throughout the war. The Chain Home radars allowed Britain to successfully deploy its limited air defenses against the heavy German air attacks conducted during the early part of the war. They operated at about 30 MHz—in what is called the short-wave, or high-frequency (HF), band—which is actually quite a low frequency for radar. It might not have been the optimum solution, but the inventor of British radar, Sir Robert Watson-Watt, believed that something that worked and was available was better than an ideal solution that was only a promise or might arrive too late.

British Chain Home radar system

The Soviet Union also started working on radar during the 1930s. At the time of the German attack on their country in June 1941, the Soviets had developed several different types of radars and had in production an aircraft-detection radar that operated at 75 MHz (in the very-high-frequency [VHF] band). Their development and manufacture of radar equipment was disrupted by the German invasion, and the work had to be relocated.

At the beginning of World War II, Germany had progressed further in the development of radar than any other country. The Germans employed radar on the ground and in the air for defense against Allied bombers. Radar was installed on a German pocket battleship as early as 1936. Radar development was halted by the Germans in late 1940 because they believed the war was almost over. The United States and Britain, however, accelerated their efforts. By the time the Germans realized their mistake, it was too late to catch up.

Except for some German radars that operated at 375 and 560 MHz, all of the successful radar systems developed prior to the start of World War II were in the VHF band, below about 200 MHz. The use of VHF frequencies posed several problems. First, beamwidths are broad. (Narrow beamwidths yield greater accuracy, better resolution, and the exclusion of unwanted echoes from the ground or other clutter.) Second, the VHF portion of the electromagnetic spectrum does not permit the wide bandwidths required for the short pulses that allow for greater accuracy in range determination (see above *Fundamentals of radar: Target information obtained by radar*). Third, VHF frequencies are subject to atmospheric noise, which limits receiver sensitivity. In spite of these drawbacks, VHF represented the frontier of radio technology in the 1930s, and radar development at this frequency range constituted a genuine pioneering accomplishment. It was well understood by the early developers of radar that operation at

even higher frequencies was desirable, particularly since narrow beamwidths could be achieved without excessively large antennas. (The beamwidth of an antenna of fixed size is inversely proportional to the radar frequency.)

Invention of the magnetron

Advances during World War II. The opening of higher frequencies (those of the microwave region) to radar, with its attendant advantages, came about in late 1939 when the cavity magnetron oscillator was invented by British physicists at the University of Birmingham. In 1940 the British generously disclosed to the United States the concept of the magnetron, which then became the basis for the work undertaken by the newly formed Massachusetts Institute of Technology (MIT) Radiation Laboratory at Cambridge, Mass. It was the magnetron that made microwave radar a reality in World War II. (For a description of the magnetron, see the article ELECTRONICS: *Principal devices and components: Electron tubes*.)

The successful development of innovative and important microwave radars at the MIT Radiation Laboratory has been attributed to the urgency for meeting new military capabilities as well as to the enlightened and effective scientific management of the laboratory and the recruiting of talented, dedicated scientists. Approximately 150 different radar systems were developed as a result of the laboratory's program during the five years of its existence (1940–45).

One of the most notable microwave radars developed by the MIT Radiation Laboratory was the SCR-584, a widely used gunfire-control system. It employed conical scan tracking, and, with its four-degree beamwidth, it had sufficient angular accuracy to place antiaircraft guns on target without the need for searchlights or optics, as was required with the older VHF SCR-268 gun-laying radar, which had very wide beamwidths. The SCR-584 operated in the frequency range from 2.7 to 2.9 GHz (in the S band) and had a parabolic reflector antenna with a diameter of nearly two metres (six feet). It was first used in combat early in 1944 on the Anzio beachhead in Italy. Its introduction was timely, since the Germans by that time had learned how to jam its predecessor, the SCR-268. The introduction of the SCR-584 microwave radar caught the Germans unprepared.

Radar technology since the mid-1940s. After the war, progress in radar technology slowed considerably. The last half of the 1940s was devoted principally to developments initiated during the war. Two of these were the monopulse tracking radar and the MTI radar (see above). It required many more years of developmental work to bring these two radar techniques to full capability.

New and better radar systems emerged during the 1950s. One of these was a highly accurate monopulse tracking radar designated the AN/FPS-16, which was capable of an angular accuracy of about 0.1 milliradian (roughly 0.006 degree). There also appeared large, high-powered radars designed to operate at 220 MHz (VHF) and 450 MHz (UHF [ultrahigh frequency]). These systems, equipped with large mechanically rotating antennas (more than 120 feet [36 metres] in horizontal dimension), could reliably detect aircraft at very long ranges. Another notable development was the klystron amplifier, which provided a source of stable high power for very long-range radars. Synthetic aperture radar first appeared in the early 1950s, but took almost 30 more years to reach a high state of development with the introduction of digital processing and other advances. The airborne pulse Doppler radar also was introduced in the late 1950s in the Bomarc air-to-air missile.

The klystron

The decade of the 1950s also saw the publication of important theoretical concepts that helped put radar design on a more quantitative basis. These included the statistical theory of detection of signals in noise; the so-called matched filter theory, which showed how to configure a radar receiver to maximize detection of weak signals; the Woodward ambiguity diagram, which made clear the trade-offs in waveform design for good range and radial velocity measurement and resolution; and the basic methods for Doppler filtering in MTI radars, which later became important when digital technology allowed the theoretical concepts to become a practical reality.

The Doppler frequency shift and its utility for radar were known before World War II, but it took years of developmental work to achieve the technology necessary for wide-scale adoption. Serious application of the Doppler principle to radar began in the 1950s, and today the principle has become vital in the operation of many radar systems. As previously explained, the Doppler frequency shift of the reflected signal results from the relative motion between the target and the radar. Doppler frequency is indispensible in continuous wave (CW; see below), MTI, and pulse Doppler radars, which all must detect moving targets in the presence of large clutter echoes. The detection of the Doppler frequency shift is the basis for the police speed meter. SAR and ISAR imaging radars make use of Doppler frequency to generate high-resolution images of terrain and targets. The Doppler frequency shift also has been used to measure the velocity of the aircraft carrying the radar system. The extraction of the Doppler shift in weather radars, moreover, allows the identification of severe storms not possible by other techniques.

The first large electronically steered phased-array radars were put into operation in the 1960s. Airborne MTI radar for aircraft detection was developed for the U.S. Navy's Grumman E-2A airborne-early-warning (AEW) aircraft at this time. Many of the attributes of HF over-the-horizon radar (see below *Examples of radar systems*) were demonstrated during the 1960s, as were the first radars designed for detecting ballistic missiles and satellites.

Impact of digital technology

During the 1970s digital technology underwent a tremendous advance, which made practical the signal and data processing required for modern radar. Digital processing allowed radar to achieve what was known to be theoretically possible but difficult to realize by other methods. Significant advances also were made in airborne pulse Doppler radar, greatly enhancing its ability to detect aircraft in the midst of heavy ground clutter. The U.S. Air Force's airborne warning and control system (AWACS) radar and military airborne intercept radar depend on the pulse Doppler principle. It might be noted too that radar began to be used in spacecraft for remote sensing of the environment during the 1970s.

Over the next decade radar methods evolved to a point where they were able to distinguish one type of target from another. Serial production of phased-array radars for air defense (the Patriot and Aegis systems), airborne bomber radar (B-1B aircraft), and ballistic missile detection (Pave Paws) also became feasible during the 1980s. Advances in remote sensing made it possible to measure winds at sea, the geoid (the Earth's figure that corresponds to mean sea level), ocean roughness, ice conditions, and other environmental effects. Solid-state technology, including very large-scale integration (VLSI) and integrated microwave circuitry, permitted new radar capabilities that were only academic curiosities a decade or two earlier.

Continued advances in computer technology in the 1990s allowed increased information about the nature of targets and the environment to be obtained from radar echoes. The introduction of Doppler weather radar systems (as, for example, Nexrad), which measure the radial component of wind speed as well as the rate of precipitation, provided new hazardous weather warning capability. Unattended radar operation with little downtime for repairs was demanded of manufacturers for such applications as air traffic control. HF over-the-horizon radar systems were operated by several countries, primarily for the detection of aircraft at very long ranges over large areas of the oceans. Space-based radars continued to gather information about the Earth's land and sea surfaces on a global basis. Improved imaging radar systems were carried by space probes to obtain higher-resolution pictures of the surface of Venus.

Satellite-borne radar systems

RADAR SUBSYSTEMS

Figure 3 (see above) shows the major subsystems that make up a typical radar system. These subsystems are described in greater detail here.

Antennas. A widely used form of radar antenna is the parabolic reflector, the principle of which is shown in cross section in Figure 5 (A). A horn antenna or other small antenna is placed at the focus of the parabola to illuminate

the parabolic surface of the reflector. After being reflected by this surface, the electromagnetic energy is radiated as a narrow beam. A paraboloid, which is generated by rotating a parabola about its axis, forms a symmetrical beam called a pencil beam. A fan beam, one with a narrow beamwidth in azimuth and a broad beamwidth in elevation, can be obtained by illuminating an asymmetrical section of the paraboloid. An example of an antenna that produces a fan beam is shown below in Figure 6.

The half-wave dipole (Figure 5 [B]), whose dimension is one-half of the radar wavelength, is the classic type of electromagnetic antenna. A single dipole is not of much use for radar, since it produces a beamwidth too wide for most applications. Radar requires a narrow beam (a beamwidth of only a few degrees) in order to concentrate its energy on the target and to determine the target location with accuracy. Such narrow beams can be formed by combining many individual dipole antennas so that the signals radiated or received by each elemental dipole are in unison, or in step. (The radar engineer would say that the signals are "in phase" with one another or that they are coherently added together.) This is called a phased-array antenna, the basic principle of which is shown in Figure 5 (C).

The phase shifters at each radiating antenna-element shift the phase of the signal, so that all signals received from a particular direction will be in step with one another. (As a result of this, the power radiated from the elements adds together.) Similarly, all signals radiated by the individual elements of the antenna will be in step with one another in some specific direction. Changing the phase shift at each element alters the direction of the antenna beam.

Electronically steered phased-array

An antenna of this kind is called an electronically steered phased-array. It allows rapid changes in the position of the beam without moving large mechanical structures. In some systems, the beam can be changed from one direction to another within microseconds.

The individual radiating elements of a phased-array antenna need not be dipoles; various other types of antenna elements also can be used. For example, slots cut in the side of a waveguide are common, especially at the higher microwave frequencies. In a radar that requires a one-degree, pencil-beam antenna, there might be about 5,000 individual radiating elements (the actual number depends on the particular design). The phased-array radar is more complex than radar systems that employ reflector antennas, but it provides capabilities not otherwise available.

Since there are many control points (each individual antenna element) in a phased-array, the radiated beam can be shaped to give a desired pattern to the beam. Controlling the shape of the radiated beam is important when the beam has to illuminate the air space where aircraft are found but not illuminate the ground, where clutter echoes are produced. Another example is when the stray radiation (called antenna sidelobes) outside the main beam of the antenna pattern must be minimized.

The electronically steered phased-array is attractive for applications that require large antennas or when the beam must be rapidly changed from one direction to another. Satellite surveillance radars and ballistic missile detection radars (such as the system shown below in Figure 8) are examples that usually require phased-arrays. The U.S. Army's Patriot battlefield air-defense system and the U.S. Navy's Aegis system for ship air-defense also depend on the electronically steered phased-array antenna.

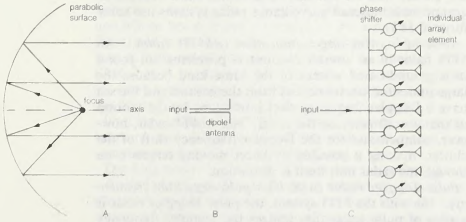


Figure 5: *Radar antennas.*
(A) A parabolic reflector antenna in which the energy radiated from the focus is reflected from the parabolic surface as a narrow beam. (B) A dipole antenna. (C) A phased-array antenna composed of many individual radiating elements.

The phased-array antenna is also used without the phase shifters in Figure 5 (C). The beam is steered by the mechanical movement of the entire antenna. Antennas of this sort are preferred over the parabolic reflector for airborne applications (see Figure 7), in land-based air-surveillance radars requiring multiple beams (as in the so-called 3D radars, which measure elevation angle in addition to azimuth and range), and in applications that require ultralow antenna sidelobe radiation.

Transmitters. The transmitter of a radar system must be efficient, reliable, not too large in size and weight, and easily maintained, as well as have the wide bandwidths and high power that are characteristic of radar applications. In MTI, pulse Doppler, and CW applications, the transmitter must generate noise-free, stable transmissions so that extraneous (unwanted) signals from the transmitter do not interfere with the detection of the small Doppler frequency shift produced by weak moving targets.

It was observed earlier that the invention of the magnetron transmitter in the late 1930s resulted in radar systems that could operate at the higher frequencies known as microwaves. The magnetron transmitter has certain limitations, but it continues to be widely used—generally in low-average-power applications such as ship navigation radar and airborne weather-avoidance radar. The magnetron is a power oscillator in that it self-oscillates (*i.e.*, generates microwave energy) when voltage is applied. Other radar transmitters usually are power amplifiers in that they take low-power signals at the input and amplify them to high power at the output. This provides stable high-power signals, as the signals to be radiated can be generated with precision at low power.

The klystron amplifier is capable of some of the highest power levels used in radar. It has good efficiency and good stability. The disadvantages of the klystron are that it is usually large and it requires high voltages (*e.g.*, about 90 kilovolts for one megawatt of peak power). At low power the instantaneous bandwidth of the klystron is small, but the klystron is capable of large bandwidth at high peak powers of a few megawatts.

The traveling-wave tube (TWT) is related to the klystron. It has very wide bandwidths at low peak power, but, as the peak power levels are increased to those needed for radar, its bandwidth decreases. As peak power increases, the bandwidths of the TWT and the klystron approach one another.

Solid-state transmitters

Solid-state transmitters, such as the silicon bipolar transistor, are attractive because of their potential for long life, ease of maintenance, and relatively wide bandwidth. An individual solid-state device generates relatively low power and can be used only when the radar application can be accomplished with low power (as in short-range applications or in the radar altimeter). High power can be achieved, however, by combining the outputs of many individual solid-state devices.

While the solid-state transmitter is easy to maintain and is capable of wide-band operation, it has certain disadvantages. It is much better suited for long pulses (milliseconds) than for the short pulses (microseconds). Long pulses can complicate radar operation because signal processing (such as pulse compression) is needed to achieve the desired range resolution. Furthermore, a long-pulse radar generally requires several different pulse widths: a long pulse for long range and one or more shorter pulses to observe targets at the ranges masked when the long pulse is transmitting. (A one-millisecond pulse, for example, masks echoes from 0 to about 80 nautical miles.)

Every kind of transmitter has its disadvantages as well as advantages. In any particular application, the radar engineer must continually search for compromises that give the results desired without too many negative effects that cannot be adequately accommodated.

Receivers. Like most other receivers, the radar receiver is a classic superheterodyne. It has to filter the desired

echo signals from unwanted clutter signals and receiver noise that interfere with detection. It also must amplify the weak received signals to a level where the receiver output is large enough to actuate a display or a computer. The technology of the radar receiver is well established and seldom sets a limit on radar performance.

The receiver must have a large dynamic range in situations where it is necessary to detect weak signals in the presence of very large clutter echoes by recognizing the Doppler frequency shift of the desired moving targets. Dynamic range can be loosely described as the ratio of the largest to the smallest signals that can be handled adequately by a receiver without distortion. A radar receiver might be required to detect signals that vary in power by a million to one—and sometimes much more.

In most cases, the sensitivity of a radar receiver is determined by the noise generated internally at its input. Because it does not generate much noise of its own, a transistor is usually used as the first stage of a receiver.

Signal and data processors. The signal processor is the part of the receiver that extracts the desired signal and rejects clutter. Doppler filtering in an MTI radar or in a pulse Doppler system is an example. Most signal processing is performed digitally with computer technology. Digital processing has significant capabilities in signal processing not previously available with analog methods. Without digital methods many of the signal processing techniques found in today's high-performance radars would not be possible. Digital processing also has made practical data processing, such as required for automatic tracking.

Digital processing

Pulse compression (described below) is sometimes included under signal processing. It too benefits from digital technology, but analog processors (*e.g.,* surface acoustic wave delay-lines) are used rather than digital methods when pulse compression must achieve resolutions of a few metres or less.

Displays. The cathode-ray tube has been the traditional means of displaying the output of a radar system. Although it has its limitations, the CRT has been the preferred technology ever since the early days of radar. The CRT has undergone continual improvement that has made it even more versatile.

The radar display shown in Figure 4 is an example of a plan position indicator, or PPI. It is a maplike presentation in polar coordinates of range and angle. The CRT screen is dark (other than for slight noise background) except when echo signals are present. The PPI is called an intensity-modulated display because the intensity of the electron beam of the CRT is increased sufficiently to excite the phosphor of the screen whenever an echo signal is present. The PPI is the most common form of display in use with radar. Another variety, the B-scope, is also an intensity-modulated display that presents the same information and the same coordinates as the PPI but in rectangular rather than polar format. In still another format, the A-scope, the received signal amplitude is displayed as the vertical coordinate and the range as the horizontal coordinate. The A-scope is called an amplitude-modulated display because echo signals are indicated by the increased amplitude (the vertical coordinate) on the CRT. The A-scope is not a suitable display for a surveillance radar that must search 360 degrees in azimuth, but it is used for tracking radars and in experimental radars when examining the nature of the echo signal is important.

All practical radar displays have been two dimensional (*i.e.,* they use a flat screen), yet most radars provide more information than can be displayed on the two coordinates of a flat screen. Colour coding of the intensity-modulated signal on the PPI is sometimes used to provide additional information about the echo signal. Colour has been employed, for example, to indicate the strength of the echo. Doppler weather radars make good use of colour coding to indicate on a two-dimensional display the rain intensity associated with each echo shown. They also utilize colour to indicate the radial speed of the wind, the wind shear, and other information relating to severe storms. The PPI displays targets as if seen in a horizontal plane. On the other hand, a range-height indicator, or RHI, is an intensity-modulated display that presents the echoes that appear in a vertical plane—*e.g.,* a vertical cut through the cloud of a severe storm.

Colour coding

The radar display has benefited from the availability of digital technology. Digital memory allows the radar to store data from an entire scan period (usually one rotation of the radar antenna) and present the information to the operator all at once (as in the case of a television-type monitor) rather than display targets only when they are actually within the antenna beam. This allows the operator to view the entire scene all the time and to manipulate the output to display the type of target information of most interest.

Modern surveillance radars rarely display the output of a radar receiver without further processing (raw video). When automatic detection of targets is employed in a radar system, the rejection of unwanted echoes such as land or sea clutter, the addition of the radar pulses received from a target, and the decision as to whether a target is present or not are all performed electronically without assistance from a human operator. The display then shows only detected targets without the background noise. This has been called a "cleaned-up" display or processed video. When automatic tracking is performed electronically (in a digital data processor), only processed target tracks are displayed and no individual target detections are indicated. The speed of a target and its direction of travel can be indicated on the CRT by the length of the line defining the track and its orientation. Near each target track on the display, alphanumeric information can be entered automatically to indicate information that is known about the target. For example, when the air-traffic-control radar-beacon system (ATCRBS) is used in conjunction with an air-surveillance radar, the alphanumeric data on the display can indicate the flight number of the aircraft and its altitude.

TYPES OF RADAR

Radar systems may be categorized according to the function they perform—*e.g.,* aircraft surveillance, surface (ground or sea) surveillance, space surveillance, tracking, weapon control, missile guidance, instrumentation, remote sensing of the environment, intruder detection, or underground probing. They also may be classified, as in the listing below, on the basis of the particular radar technique they employ. It is difficult to give in only a few words precise and readily understandable descriptions of the many types of radar available. The following survey is necessarily brief and qualitative. Additional information about each radar type can be found in the books listed in the *Bibliography.*

Simple pulse radar. This is by far the most widely used technique and constitutes what might be termed "conventional" radar. (For a discussion of its fundamentals, see above *Fundamentals of radar: Pulse radar.*) All but the last two techniques outlined below employ a pulse waveform; however, they have additional features that give an enhanced performance as compared to simple pulse radar.

Moving-target indication (MTI) radar. This is a form of pulse radar that uses the Doppler frequency shift of the received signal to detect moving targets, such as aircraft, and to reject the large unwanted echoes from stationary clutter that do not have a Doppler shift. Almost all ground-based aircraft surveillance radar systems use some form of MTI.

Airborne moving-target indication (AMTI) radar. An MTI radar in an aircraft encounters problems not found in a ground-based system of the same kind because the large undesired clutter echoes from the ground and the sea have a Doppler frequency shift introduced by the motion of the aircraft carrying the radar. The AMTI radar, however, compensates for the Doppler frequency shift of the clutter, making it possible to detect moving targets even though the radar unit itself is in motion.

Pulse Doppler radar (with high pulse-repetition frequency). As with the MTI system, the pulse Doppler radar is a type of pulse radar that utilizes the Doppler frequency shift of the echo signal to reject clutter and detect moving aircraft. However, it operates with a much higher pulse-repetition frequency (prf) than the MTI radar. (A high-prf pulse Doppler radar, for example, might have a prf of 100

kHz, as compared to an MTI radar with prf of perhaps 300 Hz.) The difference of prfs gives rise to distinctly different behaviour. The MTI radar uses a low prf in order to obtain an unambiguous range measurement. This causes the measurement of the target's radial velocity (as derived from the Doppler frequency shift) to be highly ambiguous and can result in missing some target detections. On the other hand, the pulse Doppler radar operates with a high prf so as to have no ambiguities in the measurement of radial velocity. A high prf, however, causes a highly ambiguous range measurement. The true range is resolved by transmitting multiple waveforms with different prfs.

Pulse Doppler radar (with medium pulse-repetition frequency). A modified form of pulse Doppler radar that operates at a lower prf (10 kHz, for example) than the above-mentioned high-prf pulse Doppler system has both range and Doppler shift ambiguities. It is, however, better for detecting aircraft with low closing speeds than high-prf pulse Doppler radar (which is better for detecting aircraft with high closing speeds). An airborne medium-prf pulse Doppler radar might have to use seven or eight different prfs in order to extract the target information without ambiguities.

High-range-resolution radar. This type of radar uses a very short pulse with range resolution from several metres to a fraction of a metre. Such a radar can profile a target and measure its projected length in the range dimension.

Pulse-compression radar. The ability to generate very short pulses with high peak power (and high energy) is limited for practical reasons by voltage breakdown, or arcing. Thus, conventional high-range-resolution radars with short pulses often are limited in peak power and are not capable of operating at long ranges. Pulse compression overcomes this limitation by obtaining the resolution of a short pulse but with the energy of a long pulse. It does this by modulating either the frequency or the phase of a long, high-energy pulse. The frequency or phase modulation allows the long pulse to be compressed in the receiver by an amount equal to the reciprocal of the signal bandwidth.

Radars with high resolution

Synthetic aperture radar (SAR). The SAR was described above as utilizing resolution in Doppler frequency to provide the equivalent of cross-range (or angle) resolution. More often, it is regarded as a synthetic antenna generated by a moving radar. The effect of a large antenna is obtained by storing the echo signals in a storage medium, or memory, and processing a substantial number of the previously received echoes just as if they were received by a large antenna. This kind of radar is primarily used for mapping the Earth's surface. Although it is not obvious, the two different models used for describing a SAR—a synthetic antenna and Doppler-frequency resolution—are equivalent and produce the same results.

Inverse synthetic aperture radar (ISAR). As previously noted, an ISAR depends on target motion to provide the Doppler frequency shift between various parts of the target and the radar unit so as to obtain high resolution in cross range. A two-dimensional high-resolution image of a target can be obtained by using ISAR for cross-range resolution in conjunction with either a short pulse or pulse-compression radar for resolution in the range dimension.

Side-looking airborne radar (SLAR). This variety of airborne radar employs a large side-looking antenna (*i.e.,* one whose beam is perpendicular to the aircraft's line of flight) and is capable of high range resolution. (The resolution in cross range is not as good as can be obtained with SAR, but it is simpler than the latter and is acceptable for some applications.) SLAR generates maplike images of the ground and permits detection of ground targets.

Imaging radar. Synthetic aperture, inverse synthetic aperture, and side-looking airborne radar techniques are sometimes referred to as imaging radars.

Tracking radar. This kind of radar continuously follows a single target in angle and range to determine its path, or trajectory, and to predict its future position. There are two classes of tracking radars: conical scan and monopulse. The conical scan tracker is simpler but not as accurate as the monopulse variety. Furthermore, monopulse tracking is not as susceptible to some forms of electronic countermeasure as the conical scan. The single-target tracking radar provides target location almost continuously. A typical tracking radar might measure the target location at a rate of 10 times per second.

Track-while-scan radar. This form of surveillance radar can provide tracks of all targets within its area of coverage by measuring the location of targets on each rotation of the antenna. Though called track-while-scan radar, it is more often known as automatic detection and tracking, or ADT. The output on a visual display from such a radar usually consists of the tracks of the targets (vectors showing direction and speed) rather than individual detections (blips). This type of tracking is suitable for surveillance radars, while continuous tracking is more appropriate for weapon control and instrumentation-radar applications (see below *Major applications of radar*). ADT

3D radar. Conventional air-surveillance radar measures the location of a target in two dimensions—range and azimuth. The elevation angle, from which target height can be derived, also can be determined. The so-called 3D radar is an air-surveillance radar that measures range in a conventional manner but that has an antenna which is mechanically rotated about a vertical axis to obtain the azimuth angle of a target and which has either fixed multiple beams in elevation or a scanned pencil beam to measure its elevation angle. There are other types of radar (such as electronically scanned phased arrays and tracking radars) that measure the target location in three dimensions, but a radar that is properly called 3D is an air-surveillance system that measures the azimuth and elevation angles as just described.

Electronically scanned phased-array radar. An electronically scanned phased-array antenna can position its beam rapidly from one direction to another without mechanical movement of large antenna structures. Agile, rapid beam switching permits the radar to track many targets simultaneously and to perform other functions as required.

Continuous-wave (CW) radar. Since a CW radar transmits and receives at the same time, it must depend on the Doppler frequency shift produced by a moving target to separate the weak echo signal from the strong transmitted signal. A simple CW radar can detect targets, measure their radial velocity (from the Doppler frequency shift), and determine the direction of arrival of the received signal. However, a more complicated waveform is required for finding the range of the target.

Frequency-modulated continuous-wave (FM-CW) radar. If the frequency of a CW radar is continually changed with time, the frequency of the echo signal will differ from that transmitted and the difference will be proportional to the range of the target. Accordingly, measuring the difference between the transmitted and received frequencies gives the range to the target. In such a frequency-modulated continuous-wave radar, the frequency is generally changed in a linear fashion, so that there is an up-and-down alternation in frequency. The most common form of FM-CW radar is the radar altimeter used on aircraft to determine height above the ground. Phase modulation, rather than frequency modulation, of the CW signal has also been used to obtain range measurement.

FACTORS AFFECTING RADAR PERFORMANCE

The performance of a radar system can be judged by the following: (1) the maximum range at which it can see a target of a specified size, (2) the accuracy of its measurement of target location in range and angle, (3) its ability to distinguish one target from another, (4) its ability to detect the desired target echo when masked by large clutter echoes, unintentional interfering signals from other "friendly" transmitters, or intentional radiation from hostile jamming (if a military radar), (5) its ability to recognize the type of target, and (6) its availability (ability to operate when needed), reliability, and maintainability. Some of the major factors that affect performance are discussed in this section.

Transmitter power and antenna size. The maximum range of a radar system depends in large part on the average power of its transmitter and the physical size of its antenna. (In technical terms, this is the power-aperture product.) There are practical limits to each. As noted

before; some radar systems have an average power of roughly one megawatt. Phased-array radars about 100 feet in diameter are not uncommon; some are much larger. Likewise, mechanically scanned reflector antennas about 100 feet or larger in size can be found. There are specialized radars with (fixed) antennas, such as some HF over-the-horizon radars and the U.S. Space Surveillance System (SPASUR), that extend more than one mile.

Receiver noise. The sensitivity of a radar receiver is determined by the unavoidable noise that appears at its input. At microwave radar frequencies, the noise that limits detectability is usually generated by the receiver itself (*i.e.*, by the random motion of electrons at the input of the receiver) rather than by external noise that enters the receiver via the antenna. The radar engineer often employs a transistor amplifier as the first stage of the receiver even though lower noise can be obtained with more sophisticated devices. This is an example of the application of the basic engineering principle that the "best" performance that can be obtained might not necessarily be the solution that best meets the needs of the user.

The receiver is designed to enhance the desired signals and to reduce the noise and other undesired signals that interfere with detection. The designer attempts to maximize the detectability of weak signals by using what radar engineers call a "matched filter," which is a filter that maximizes the signal-to-noise ratio at the receiver output. The matched filter has a precise mathematical formulation that depends on the shape of the input signal and the character of the receiver noise. A suitable approximation to the matched filter for the ordinary pulse radar, however, is one whose bandwidth in hertz is the reciprocal of the pulse width in seconds.

Target size. The size of a target as "seen" by radar is not always related to the physical size of the object. The measure of the target size as observed by radar is called the radar cross section and is given in units of area (square metres). It is possible for two targets with the same physical cross sectional area to differ considerably in radar size, or radar cross section. For example, a flat plate one square metre in area will produce a radar cross section of about 1,000 square metres at a frequency of 3,000 megahertz (S band; see below) when viewed perpendicular to the surface. A cone-sphere (an object resembling an ice-cream cone) when viewed in the direction of the cone rather than the sphere could have a radar cross section one thousandth of a square metre even though its projected area is also one square metre. In theory, this value does not depend to a great extent on the size of the cone or the cone angle. Thus the flat plate and the cone-sphere can have radar cross sections that differ by a million to one even though their physical projected areas are the same.

Radar cross section

The sphere is an unusual target in that its radar cross section is the same as its physical cross section area (when its circumference is large compared to the radar wavelength). That is to say, a sphere with a projected area of one square metre has a radar cross section of one square metre.

Commercial aircraft might have radar cross sections from about 10 to 100 square metres, except when viewed broadside, where it is much larger. (This is an aspect that is seldom of interest, however.) Most air-traffic-control radars are required to detect aircraft with a radar cross section as low as two square metres, since some small general-aviation aircraft can be of this value. For comparison, the radar cross section of a man has been measured at microwave frequencies to be about one square metre. A bird can have a cross section of 0.01 square metre. Although this is a small value, a bird can be readily detected at ranges of several tens of miles by long-range radar. In general, many birds can be picked up by radar so that special measures must usually be taken to insure that echoes from birds do not interfere with the detection of desired targets.

The radar cross section of an aircraft and most other targets of practical interest is not a constant but, rather, fluctuates rapidly as the aspect of the target changes with respect to the radar unit. It would not be unusual for a slight change in aspect to cause the radar cross section to change by a factor of 10 to 1,000. (Radar engineers have to take this fluctuation in the radar cross section of targets into account in their design.)

Clutter. Echoes from land, sea, rain, snow, hail, birds, insects, auroras, and meteors are of interest to those who observe and study the environment, but they are a nuisance to those who want to detect and follow aircraft, ships, missiles, or other similar targets. Clutter echoes can seriously limit the capability of a radar system; thus a significant part of radar design is devoted to minimizing the effects of clutter without reducing the echoes from desired targets. The Doppler frequency shift is the usual means by which moving targets are distinguished from the clutter of stationary objects. Detection of targets in rain is less of a problem at the lower frequencies, since the radar echo from rain decreases rapidly with decreasing frequency and the average cross section of aircraft is relatively independent of frequency in the microwave region. Because raindrops are more or less spherical (symmetrical) and aircraft are asymmetrical, the use of circular polarization can enhance the detection of aircraft in rain. With circular polarization, the electric field rotates at the radar frequency. Because of this, the electromagnetic energy reflected by the rain and the aircraft will be affected differently, thereby making it easier to distinguish between the two. (In fair weather, most radars use linear polarization—*i.e.*, the direction of the field is fixed.)

Detecting targets during rainfall

Atmospheric effects. As was mentioned, rain and other forms of precipitation can cause echo signals that mask the desired target echoes. There are other atmospheric phenomena that can affect radar performance as well. The decrease in density of the Earth's atmosphere with increasing altitude causes radar waves to bend as they propagate through the atmosphere. This usually increases the detection range at low angles to a slight extent. The atmosphere can form "ducts" that trap and guide radar energy around the curvature of the Earth and allow detection at ranges beyond the normal horizon. Ducting over water is more likely to occur in tropical climates than in colder regions. Ducts can sometimes extend the range of an airborne radar, but on other occasions they may cause the radar energy to be diverted and not illuminate regions below the ducts. This results in the formation of what are called radar holes in the coverage. Since it is not predictable or reliable, ducting can in some instances be more of a nuisance than a help.

Loss of radar energy, when propagation is through the clear atmosphere or rain, is usually insignificant for systems operating at microwave frequencies.

Interference. Signals from nearby radars and other transmitters can be strong enough to enter a radar receiver and produce spurious responses. Well-trained operators are not often deceived by interference, though they may find it a nuisance. Interference is not as easily ignored by automatic detection and tracking systems, however, and so some method is usually needed to recognize and remove interference pulses before they enter the automatic detector and tracker of a radar.

Electronic countermeasures. The purpose of hostile electronic countermeasures (ECM) is to deliberately degrade the effectiveness of military radar. ECM can consist of (1) noise jamming that enters the receiver via the antenna and increases the noise level at the input of the receiver, (2) false target generation, or repeater jamming, by which hostile jammers introduce additional signals into the radar receiver in an attempt to confuse the receiver into thinking they are real target echoes, (3) chaff, which is an artificial cloud consisting of a large number of tiny metallic reflecting strips that create strong echoes over a large area to mask the presence of real target echoes or to create confusion, and (4) decoys, which are small, inexpensive air vehicles or other objects designed to appear to the radar as if they were real targets. Military radars are also subject to direct attack by conventional weapons or by antiradiation missiles (ARMs) that use radar transmissions to find the target and home on it.

Military radar engineers have developed various ways of countering hostile ECM and maintaining the ability of a radar system to perform its mission. It might be noted that a military radar system can often accomplish its

mission satisfactorily even though its performance in the presence of ECM is not what it would be if such measures were absent.

MAJOR APPLICATIONS OF RADAR

Areas of application. Over the years, radar has found many and varied uses for both civilian and military purposes. A sampling of some of the more significant applications is given here.

Military. Radar originally was developed to meet the needs of the military, and it continues to have significant application for military purposes. It is used to detect aircraft, missiles, artillery and mortar projectiles, ships, land vehicles, and satellites. In addition, radar controls, guides, and fuzes weapons; allows one class of target to be distinguished from another; aids in the navigation of aircraft and ships; performs reconnaissance; and determines the damage caused by weapons to targets. The importance of radar in modern warfare is borne out by the many measures designed to negate its effectiveness (in addition to direct attack, which is an option for any military target of value). Attempts to degrade military radar capability include electronic warfare (jamming, deception, chaff, decoys, and interception of radar signals), antiradiation missiles that home on radar transmissions, reduced radar cross-section targets to make detection more difficult (stealth), and high-power microwave energy transmissions to degrade or burn out sensitive receivers. A major objective of military radar development is to insure that a radar system can continue to perform its mission in spite of the various measures that attempt to degrade it.

Promoting civilian air safety

Air traffic control. Radar supports air traffic control by providing surveillance of aircraft and weather in the vicinity of airports as well as en route between airports. In the United States and elsewhere, airport surveillance radar (ASR) is employed at most major airports. It is designed to detect both commercial aircraft and general aviation aircraft, as well as precipitation, in the area around an air terminal. A larger system, the air route surveillance radar (ARSR), tracks aircraft en route. It has a range of about 200 nautical miles. Many major airports also employ airport surface detection equipment (ASDE), which is a high-resolution radar that provides the airport controller with the location and movement of ground targets within the airport, including service vehicles and taxiing aircraft. The location of dangerous weather phenomena such as "downbursts" (downward blasts of air associated with storms that have been identified as a major cause of fatal weather-related aircraft accidents) can be pinpointed with a specially configured terminal Doppler weather radar (TDWR) located near airports. Radar also has been used to "talk down" pilots to safe landings in adverse weather conditions. This is called ground-controlled approach (GCA) by the military.

Scientific applications of radar

Remote sensing. One of the early applications of remote sensing involved the observation of rainfall. The radar measurement of the radial velocity of precipitation (from the Doppler frequency shift) in conjunction with the strength of the reflected signals (reflectivity) can indicate the severity of storms, as well as provide other important information for reliable weather forecasting (see also CLIMATE AND WEATHER: *Meteorological measurement and weather forecasting: History of weather forecasting: Modern trends and developments: Application of radar*).

Astronomers have made radar observations of meteors, auroras, and certain planets. Synthetic aperture radars on orbiting spacecraft have mapped the surface of Venus beneath the ever-present cloud cover that blocks observation at optical wavelengths. Space-based radar systems have measured the Earth's geoid and ocean roughness. An important application of imaging radar from either aircraft or spacecraft is the surveillance of sea ice; information about pack ice distribution and concentration is used to route shipping in cold-weather regions.

Radar has even been used to study the movement of birds and insects at distances and under conditions where visual observation would not be possible.

Aircraft navigation. The radar altimeter measures the height of an airplane above the local terrain, Doppler navigation radar determines the plane's own speed and direction, and high-resolution radar mapping of the ground contributes to its navigation. Radars carried aboard aircraft also provide information about the location of dangerous weather so that it can be avoided. Military aircraft can fly at low altitudes with the aid of terrain-avoidance and terrain-following radars that warn of obstacles.

Ship safety. Small, relatively simple radar systems on board ships aid in piloting and collision avoidance. Similar radars on land provide harbour surveillance.

Space applications. Radars have been used in space for rendezvous, docking, and landing of spacecraft. Since size and weight are important in space, the same equipment is used on a time-shared basis aboard the U.S. Space Shuttle both as radar to allow rendezvous with (and sometimes retrieve) other spacecraft and as a two-way data link to relay satellites that communicate with ground stations. Besides providing remote sensing of the Earth's surface (see above), radar carried by orbiting spacecraft is able to monitor rainfall over the oceans. Large land-based radar systems permit the detection and tracking of satellites and space debris.

Police radar

Law enforcement. The familiar police radar is a relatively simple, low-power continuous-wave system that measures the speed of vehicles by detecting the Doppler frequency shift introduced in the echo signal by a moving vehicle. The Doppler shift is directly proportional to the radial speed of the vehicle. (A similar kind of CW radar is used to measure the speed of a baseball to determine how fast a pitcher can throw.) Radar also has been used in security systems for intrusion detection; it can "sense" the movement of people attempting to penetrate a protected area.

Instrumentation. Surveyors may make use of special radars to measure distances. CW radars are used to measure speed in certain industrial applications; the sensor does not make contact with the object whose speed is to be determined. Instrumentation radars are employed at missile test ranges for precision tracking of targets.

Radar applications by frequency. Each radar application seems to have a particular frequency band to which it is best suited. The various types of application found at the different radar frequency bands are surveyed below. The frequency letter-band nomenclature used here is that approved by the Institute of Electrical and Electronic Engineers (IEEE Standard 521–1984). These letter bands also are recognized by the U.S. Department of Defense and are listed in its Index of Specifications and Standards.

HF (3 to 30 MHz). Although the first British radar system, Chain Home (see above), operated in the HF band, it is ordinarily not a good frequency region for radar. Antenna beamwidths are very wide, the available bandwidths are narrow, the spectrum is crowded with other users, and the external noise (both natural noise and noise due to other transmitters) is high. There is, nevertheless, an important application for radar in this band—namely, long-range radar, which takes advantage of refraction by the ionosphere to extend ranges by an order of magnitude greater than can be obtained by ground-based microwave aircraft-detection radars. The ionosphere is a region of ionized gases produced by solar radiation at altitudes from about 80 to 400 kilometres or higher. The ions bend radio waves enough to return to Earth at considerable distances (see below *Examples of radar systems: Over-the-horizon radar*).

VHF (30 to 300 MHz). For reasons similar to those cited above, this frequency band is not too popular for radar. However, very long-range radars for either aircraft or satellite detection can be built at the VHF band more economically than at higher frequencies. Radar operations at such frequencies are not bothered by rain clutter or insects, but auroras and meteors produce large echoes that can interfere with target detection.

Frequency band for AEW radar

UHF (300 to 1,000 MHz). Military airborne early warning (AEW) radars operate in the UHF band to detect aircraft in the midst of clutter. This is a good frequency range for detecting extraterrestrial targets (*e.g.*, satellites and missiles), since large antennas and high power are readily obtained for this application.

L band (1,000 to 2,000 MHz). This is the preferred frequency band for long-range (200 nautical miles) air surveillance radar, such as the air-traffic control systems used to track aircraft en route between airports. It also is a band of interest for military space surveillance and missile detection because it is not as susceptible to nuclear blackout effects as radar systems that operate at the lower frequencies.

S band (2 to 4 GHz). Medium-range (50 to 60 nautical miles) airport surveillance radars are well suited for this band. It is the preferred frequency band for long-range weather observation radars. Military 3D radars that determine elevation angle as well as range and azimuth angle are often in S band, but they may also be at L band. Frequencies lower than S band are good for long-range surveillance, since large power, large antennas, and good moving-target detection are better there than at high frequencies. Frequencies greater than S band are preferred for extracting target information, as in tracking radars and weapon-control systems. Therefore, when a single frequency must be used for both surveillance and information extraction (as is necessary when only a single-frequency phased-array antenna is employed), S band can be a compromise.

C band (4 to 8 GHz). Single-frequency phased-array radars that must perform both surveillance and weapon control for air defense operate at these frequencies as well as at S band. This frequency region is well suited for long-range, precision-tracking radars.

X band (8 to 12 GHz). This is a band frequently used for shipboard civil marine radar, tracking radar, airborne weather-avoidance radar, systems for detecting mortar and artillery projectiles, and police speed meters. Most synthetic aperture radars operate at X band; the exceptions are some remote-sensing SARs that are designed for lower frequencies.

K band (12 to 40 GHz). Radars at this frequency band are usually of short range, because it is difficult to obtain the large antennas and large power necessary for long-range applications. This band has been used for airborne radar and for short-range airport surface detection (ASDE).

Millimetre waves (40 to 300 GHz). Although there has been much interest in exploring the potential of radars at millimetre wavelengths, it has not been practical for most applications because of high attenuation even in the "clear" atmosphere. It is difficult to use millimetre-wave radar for anything other than short range (a few kilometres) within the atmosphere. For deployment in outer space where there is no atmosphere to attenuate these frequencies, millimetre-wave radar, however, can be considered.

Laser radar. Laser radars, which operate at infrared and optical frequencies, also suffer from attenuation by the atmosphere, especially in bad weather, and therefore are of limited utility. Laser radar systems, however, have been used for precision range-finding in weapon control and for distance measuring in surveying. They also have been considered for use on board spacecraft to probe the nature of the atmosphere.

EXAMPLES OF RADAR SYSTEMS

Airport surveillance radar. This is a medium-range radar system capable of reliably detecting and tracking aircraft at altitudes below 25,000 feet and within 40 to 60 nautical miles of the airport where it is located. Systems of this type have been installed at more than 100 major airports throughout the United States. The ASR-9 is designed to be operable at least 99.9 percent of the time, which means that the system is down less than 10 hours per year. This high availability is attributable to reliable electronic components, a "built-in test" to search for failures, remote monitoring, and redundancy (*i.e.,* the system has two complete channels except for the antenna; when one channel must be shut down for repair, the other continues to operate). The ASR-9 is designed to operate unattended with no maintenance personnel at the radar site. A number of radar units can be monitored and controlled from a single location. When trouble occurs, the fault is identified and a maintenance person dispatched for repair.

Echoes from rain that mask the detection of aircraft are reduced by the use of Doppler filtering and other techniques devised to separate moving aircraft from undesired clutter. It is important for air-traffic controllers to recognize areas of severe weather so that they can direct aircraft safely around, rather than through, rough or hazardous conditions. The ASR-9 has a separate receiving channel that recognizes weather echoes and provides their location to air traffic controllers. Six different levels of precipitation intensity can be displayed, either with or without the aircraft targets superimposed.

The ASR-9 system operates within S band from 2.7 to 2.9 GHz. Its klystron transmitter has a peak power of 1.3 megawatts, a pulse width of 1 microsecond, and an antenna with a horizontal beamwidth of 1.4 degrees that rotates at 12.5 revolutions per minute (4.8-second rotation period).

The reflector antenna shown in Figure 6 is a section of a paraboloid. It is 16.5 feet wide and 9 feet high. Atop the radar (riding piggyback) is a lightweight planar-array antenna for the air-traffic-control radar-beacon system. Its dimensions are 26 feet by 5.2 feet. ATCRBS is the primary means for detecting and identifying aircraft equipped with a transponder that can reply to the ATCRBS interrogation. The ATCRBS transmitter, which is independent of the radar system and operates at a different frequency, radiates a coded interrogation signal. Aircraft equipped with a suitable transponder can recognize the interrogation and send a coded reply at a frequency different from the interrogation frequency. The interrogator might then ask the aircraft, by means of other coded signals, to automatically identify itself and to report its altitude. ATCRBS only works with cooperative targets (*i.e.,* those with an operational transponder).

By courtesy of Westinghouse Electric Corporation

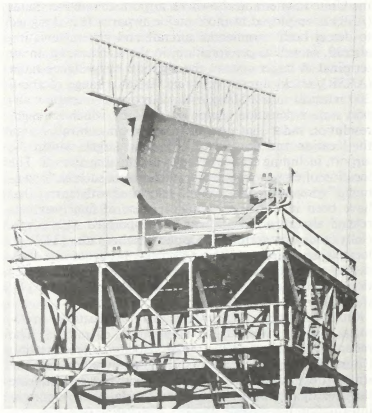


Figure 6: Reflector antenna for the ASR-9 airport surveillance radar, with an air-traffic-control radar-beacon system (ATCRBS), or Mode S, antenna mounted on top.

Doppler weather radar. For many years radar has been used to provide information about the intensity and extent of rain and other forms of precipitation. This application of radar is well known in the United States from the familiar television weather reports of precipitation observed by the radars of the National Weather Service. A major improvement in the capability of weather radar came about when engineers developed new radars that could measure the Doppler frequency shift in addition to the magnitude of the echo signal reflected from precipitation. The Doppler frequency shift is important because it is related to the radial velocity of the precipitation blown by wind (the component of the wind moving either toward or away from the radar installation). Since tornadoes, mesocyclones (which spawn tornadoes), hurricanes, and

other hazardous weather phenomena tend to rotate, the radial wind speed, as a function of angle (as is measured by Doppler radar), will identify rotating weather patterns. (Rotation is indicated when the measurement of the Doppler frequency shift shows that the wind is coming toward the radar at one angle and away from it at a nearby angle.)

Nexrad

The pulse Doppler weather radars employed by the National Weather Service, known as Nexrad, make quantitative measurements of precipitation, warn of potential flooding or dangerous hail, provide wind speed and direction, indicate the presence of wind shear and gust fronts, track storms, predict thunderstorms, and provide other meteorologic information. In addition to measuring precipitation (from the intensity of the echo signal) and radial velocity (from the Doppler frequency shift), Nexrad can also measure the spread in radial speed (difference between the maximum and the minimum speeds) of the precipitation particles within each radar resolution cell. The spread in radial speed is an indication of wind turbulence.

Another improvement in the weather information provided by Nexrad is the digital processing of radar data, a procedure that renders the information in a form that can be interpreted by an observer who is not necessarily a fully trained meteorologist. The computer automatically identifies severe weather effects and indicates their nature on a CRT display viewed by the observer. High-speed communication lines integrated with the Nexrad system allow timely weather information to be transmitted for display to various users.

The Nexrad radar operates at frequencies from 2.7 to 3.0 GHz (S band) and is equipped with a 25-foot-diameter antenna. It takes five minutes to scan its 1 degree beamwidth through 360 degrees in azimuth and from 0 to 20 degrees in elevation. The Nexrad system can measure rainfall up to a distance of 460 kilometres and determine its radial velocity as far as 230 kilometres.

A serious weather hazard to aircraft in the process of landing or taking off from an airport is the downburst, or microburst. This is the above-mentioned strong downdraft that causes wind shear capable of forcing aircraft to the ground. Terminal Doppler weather radar is the name of the type of system at airports that is specially designed to detect dangerous microbursts. It is similar in principle to Nexrad, but is a shorter-range system since it only has to observe dangerous weather phenomena in the vicinity of an airport. It also operates at a higher frequency (C band) to avoid interference with the Nexrad and ASR systems (which operate at S band).

Airborne combat radar. A modern combat aircraft is generally required not only to intercept hostile aircraft but also to attack surface targets on the ground or on the sea. The radar that serves such an aircraft must have the capabilities to perform these distinct military missions. This is not easy because each mission has different requirements. The different ranges, accuracies, and rates at which the radar data is required, the effect of the environment (land or sea clutter), and the type of target (land features or moving aircraft) call for different kinds of radar waveforms (different pulse widths and pulse repetition frequencies). In addition, an appropriate form of signal processing is required to extract the particular information needed for each military function. Radar for combat aircraft must therefore be multimode—*i.e.*, operate with different waveforms, signal processing, and antenna scanning. It would not be unusual for an airborne combat radar to have from 8 to 10 air-to-air modes and 6 to 10 air-to-surface modes. Furthermore, the radar system might be required to assist in rendezvous with a companion combat craft or with a refueling aircraft, provide guidance as to air-to-air missiles, and counter hostile electronic jamming. The problem of achieving effectiveness with these many modes is a challenge for the radar designer and is made more difficult by the limited size and weight available on combat aircraft.

Multimode systems

The AN/APG-66 radar built for the U.S. F-16 fighter is shown in Figure 7. This is a pulse Doppler radar system that operates in the X-band region of the spectrum. The version of this radar for the British Hawk 200 aircraft occupies a volume of less than three cubic feet, weighs less than 237 pounds, and requires an input power of 2.25 kilowatts. It can search 120 degrees in azimuth and elevation and is supposed to have a range of 35 nautical miles in the "look-up" mode and 27.5 nautical miles in the "look-down" mode. The look-up mode is a more or less conventional radar mode with a low pulse-repetition-frequency that is used when the target is at medium or high altitude and no ground clutter echoes are present to mask target detection. The look-down mode uses a medium-prf pulse Doppler waveform and signal processing that provide target detection in the presence of heavy clutter. (A low prf for an X-band combat radar might be from 250 Hz to 5 kHz, a medium prf from 5 to 20 kHz, and a high prf from 100 to 300 kHz.) Radars for larger combat aircraft can have greater capability but are, accordingly, bigger and heavier than the system just described.

Ballistic missile detection and satellite surveillance radar. The systems for detecting and tracking ballistic missiles and orbiting satellites are much larger than those for aircraft detection because the ranges are longer. Such radars might be required to have maximum ranges of 2,000 to 3,000 nautical miles, as compared with 200 nautical miles for a typical long-range aircraft detection system. The average power of the transmitter might be from several hundred kilowatts to one megawatt or more, which is about 100

By courtesy of Westinghouse Electric Corporation

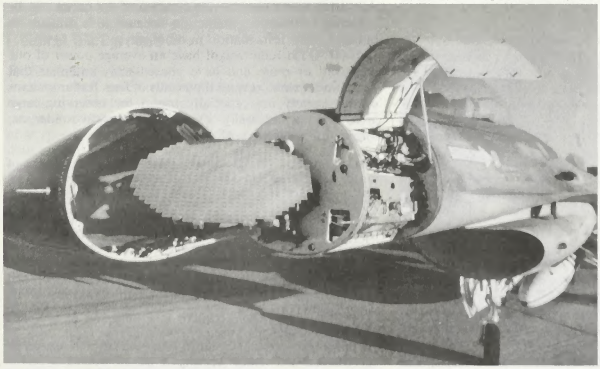


Figure 7: *The AN/APG-66 radar in an F-16 fighter aircraft.*
The mechanically scanned planar phased-array antenna with radiating horizontal slots is 29 inches wide by 19 inches high.

times greater than the average power of radars designed for aircraft detection. Antennas for this application have dimensions on the order of tens of metres to a hundred metres or more and are electronically scanned phased-array antennas capable of steering the radar beam without moving large mechanical structures. Radar systems so equipped are commonly found at the lower frequencies (typically at frequency bands of 420–450 MHz and 1,215–1,400 MHz).

The Pave Paws radar (or AN/FPS-115) is a UHF (420–450 MHz) phased-array system for detecting submarine-launched ballistic missiles. It is supposed to detect targets with a radar cross section of 10 square metres at a range of 3,000 nautical miles. The array antenna contains 1,792 active elements within a diameter of 72.5 feet. It can be expanded to 102 feet. Each active element is a module with its own solid-state transmitter, receiver, duplexer, and phase shifter. The total average power per antenna is about 145 kilowatts. Two antennas, such as shown in Figure 8, make up a system, with each capable of covering a sector 120 degrees in azimuth. Vertical coverage is from 3 to 85 degrees. An upgraded variant of this type of radar is used in the Ballistic Missile Early Warning System (BMEWS) network, with installations in Alaska, Greenland, and England. BMEWS is designed to provide warning of intercontinental ballistic missiles. Each array antenna measures almost 84 feet across and has 2,560 active elements identical to those of the Pave Paws system. Both the BMEWS and Pave Paws radars detect and track satellites and other space objects in addition to warning of the approach of ballistic missiles.

Ground-probing radar. Radar waves are usually thought of as being reflected from the surface of the ground. At the lower frequencies (below several hundred megahertz) radar energy, however, can propagate into the ground and be reflected from buried objects. The loss in propagating in the ground is very high at these frequencies, but low enough to permit ranges of about 1 to 10 metres or more. This is sufficient for detecting buried utility pipes and cables, probing the subsurface soil, detecting underground tunnels, and monitoring the subsurface conditions of highways and bridge roadways. (The ranges in ice can be much greater because the propagation loss is less in ice than in most soils.) The short ranges require that the radar system be able to resolve closely spaced objects, which means wide-bandwidth signals must be radiated. Normally, wide bandwidth is not available at the lower frequencies (especially when a 30-centimetre range resolution requires a 500 MHz bandwidth). However, since the energy is directed into the ground rather than radiated into space, the large frequency band needed for high resolution can be obtained without interference to other users of the radio spectrum.

Subsurface detection and monitoring

By courtesy of Raytheon Company

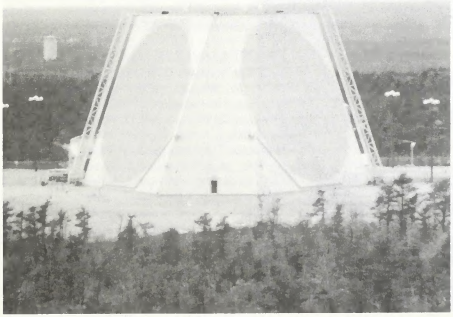


Figure 8: The AN/FPS-115 (Pave Paws) UHF, all solid-state electronically steered phased-array radar for ballistic missile and satellite detection.

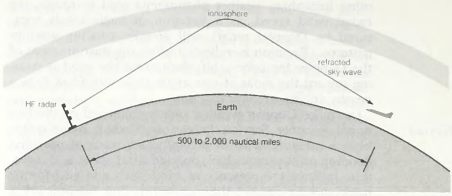


Figure 9: Refraction of HF radar radiation by the ionosphere (see text).

A ground-probing radar might radiate over frequencies ranging from 5 to 500 MHz in order to obtain good penetration (which requires low frequencies) with high resolution (which requires wide bandwidth). The antenna can be placed directly on the ground. The radar unit is small in size and so is portable.

Over-the-horizon radar. Radars at frequencies lower than 100 or 200 MHz are not desirable for radar application except in special cases. The ground-probing radar mentioned above is one such special case; here a radar at the lower frequencies can provide a capability not available with other types of radar or other sensors. Another example where lower frequencies can provide a unique and important capability is in the short wave, or HF, portion of the radio band (from 3 to 30 MHz). The advantage of the HF band is that radio waves of these frequencies are refracted (bent) by the ionosphere so that the waves return to the Earth's surface at long distances beyond the horizon, as shown in Figure 9. This permits target detection at distances from about 500 to 2,000 nautical miles. Thus, an HF OTH radar can detect aircraft at distances up to 10 times that of a ground-based microwave air-surveillance radar whose range is limited by the curvature of the Earth. Besides detection and tracking of aircraft at long ranges, an HF OTH radar can be designed to detect ballistic missiles (particularly the disturbance caused by ballistic missiles as they travel through the ionosphere), ships, and weather effects over the ocean. Winds over the ocean generate waves on the water that can be detected by OTH radar. From the Doppler frequency spectrum produced by echoes from the water waves, one can determine the direction of the waves generated by the wind and, hence, the direction of the wind itself. The strength of the waves (which indicates the state of the sea, or roughness) also can be ascertained. Timely information about the winds that drive waves over a wide expanse of the ocean can be valuable for weather prediction because it is difficult to obtain similar information in other ways.

Determining the direction and strength of ocean waves

An HF OTH radar might have an average power of one megawatt or more and have phased-array antennas that sometimes extend several thousands of feet. Radar systems of this variety are especially useful for observing large areas that are not easily covered by microwave radar, as, for example, expanses of oceans.

BIBLIOGRAPHY. Complete basic information on all forms of radar is presented in MERRILL I. SKOLNIK, *Introduction to Radar Systems*, 2nd ed. (1980); and MERRILL I. SKOLNIK (ed.), *Radar Handbook*, 2nd ed. (1990), a comprehensive reference work written for radar engineers, covering the various aspects of radar technology. S.S. SWORDS, *Technical History of the Beginnings of Radar* (1986), surveys the developments prior to World War II, focusing on early European achievements. HENRY E. GUERLAC, *Radar in World War II*, 2 vol. (1987), is an excellent, authoritative history written by a professional historian with a background in science. MERRILL I. SKOLNIK (ed.), *Radar Applications* (1988), is a collection of technical papers describing the various applications of radar, such as air traffic control, military air defense, aircraft and spacecraft surveillance, and remote sensing of the environment. DAVID K. BARTON, *Modern Radar System Analysis* (1988), provides a fine technical presentation of many aspects of radar design.

(M.I.S.)

Radiation

Radiation consists of a flow of atomic and subatomic particles and of waves, such as those that characterize heat rays, light rays, and X rays. All matter is constantly bombarded with radiation of both types from cosmic and terrestrial sources. This article delineates the properties and behaviour of radiation and the matter with which it interacts and describes how energy is transferred from radiation to its surroundings. Considerable attention is devoted to the consequences of such an energy transfer to living matter, including the normal effects on many life processes (*e.g.*, photosynthesis in plants and vision in animals) and the abnormal or injurious effects that result from the exposure of organisms to unusual types of radiation or to increased amounts of the radiations commonly encountered in nature. The applications of various forms of radiation in medicine and technological fields are touched upon as well.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 111, 112, and 128.

The article is divided into the following sections:

General background 471
  Types of radiation 471
    Electromagnetic rays and neutrinos
    Matter rays
  The structure and properties of matter 472
  The effects of radiation 473
Fundamental processes involved in the interaction of radiation with matter 473
  The passage of electromagnetic rays 473
    The field concept
    Frequency range
    Properties of light
    Wave aspects of light
    Electromagnetic waves and atomic structure
    Particle aspects of light
  The passage of matter rays 477
    Heavy charged particles
    Electrons
    Neutrons
Secondary effects of radiation 480
  Purely physical effects 480
  Molecular activation 481
    Luminescence
    Ionization phenomena
    Excitation states
    Energy transfer
  Ionization and chemical change 482
    Photochemistry
    Radiation chemistry
Tertiary effects of radiation on materials 484
  Heating effects 484
  Crystal-lattice effects 485
  Surface effects 486
Biologic effects of ionizing radiation 486
  Historical background 487
  Units for measuring ionizing radiation 487
  Sources and levels of radiation in the environment 487
    Natural sources
    Artificial sources
  Mechanism of biologic action 488
  Radionuclides and radioactive fallout 489
    Accumulation in critical organs
    The hazards of long-lived radioisotopes
  Major types of radiation injury 490
    Effects on the cell
    Effects on organs of the body (somatic effects)
    Effects on the growth and development of the embryo
    Effects on the incidence of cancer
    Shortening of the life span
  Protection against external radiation 494
  Control of radiation risks 495
Biologic effects of non-ionizing radiation 495
  Effects of Hertzian waves and infrared rays 495
    Hertzian waves
    Infrared rays
  Effects of visible and ultraviolet light 496
    Intrinsic action
    Photodynamic action
    Effects on development and biologic rhythms
    Effects on the eyes
Applications of radiation 497
  Medical applications 497
    Imaging techniques
    Other radiation-based medical procedures
  Applications in science and industry 498
    Photochemistry
    High-energy radiation
    Lasers
Bibliography 499

General background

TYPES OF RADIATION

Basic characteristics

Radiation may be thought of as energy in motion either at speeds equal to the speed of light in free space—approximately 3×10^{10} centimetres (186,000 miles) per second—or at speeds less than that of light but appreciably greater than thermal velocities (*e.g.*, the velocities of molecules forming a sample of air). The first type constitutes the spectrum of electromagnetic radiation that includes radio waves, microwaves, infrared rays, visible light, ultraviolet rays, X rays, and gamma rays, as well as the neutrino (see below). These are all characterized by zero mass when (theoretically) at rest. The second type includes such particles as electrons, protons, and neutrons. In a state of rest, these particles have mass and are the constituents of atoms and atomic nuclei. When such forms of particulate matter travel at high velocities, they are regarded as radiation. In short, the two broad classes of radiation are unambiguously differentiated by their speed of propagation and corresponding presence or absence of rest mass. In the discussion that follows, those of the first category are referred to as "electromagnetic rays" (plus the neutrino) and those of the second as "matter rays."

Wave–particle duality

At one time, electromagnetic rays were thought to be inherently wavelike in character—namely, that they spread out in space and are able to exhibit interference when they come together from two or more sources. (Such behaviour is typified by water waves in the way they propagate and periodically reinforce and cancel one another.) Matter rays, on the other hand, were considered to be inherently particle-like in character—*i.e.*, localized in space and incapable of interference. During the early 1900s, however, major experiments and attendant theories revealed that all forms of radiation, under appropriate conditions, can exhibit both particle-like and wavelike behaviour. This is referred to as the wave–particle duality and provides in large part the foundation for the modern quantum theory of matter and radiation. The wave behaviour of radiation is apparent in its propagation through space, while the particle behaviour is revealed by the nature of interactions with matter. Because of this, care must be exercised to use the terms waves and particles only when appropriate.

Electromagnetic rays and neutrinos. *Visible light and the other components of the electromagnetic spectrum.* According to the theory of relativity, the velocity of light is a fixed quantity independent of the velocity of the emitter, the absorber, or a presumably independent observer, all

three of which do affect the velocities of common wave-like disturbances such as sound. In an extended definition, the term light embraces the totality of electromagnetic radiation. It includes the following: the long electromagnetic waves predicted by the British physicist James Clerk Maxwell in 1864 and discovered by the German physicist Heinrich Hertz in 1887 (now called radio waves); infrared and ultraviolet rays; the X rays discovered in 1895 by Wilhelm Conrad Röntgen of Germany; the gamma rays that accompany many radioactive-decay processes; and some even more energetic (with higher energy) X rays and gamma rays produced as the normal accompaniment of the operations of ultrahigh-energy machines (*i.e.,* particle accelerators such as the Van de Graaff generator, the cyclotron and its variants, and the linear accelerator).

The behaviour of light seems to have interested ancient philosophers but without stimulating them to experiment, though all of them were impressed by vision. The first meaningful optical experiments on light were performed by the English physicist and mathematician Isaac Newton (beginning in 1666), who showed (1) that white light diffracted by a prism into its various colours can be reconstituted into white light by a prism oppositely arranged and (2) that light of a particular colour selected from the diffracted spectrum of a prism cannot be further diffracted into beams of other colour by an additional prism. Newton hypothesized that light is corpuscular in its nature, each colour represented by a different particle speed, an erroneous assumption. Furthermore, in order to account for the refraction of light, the corpuscular theory required, contrary to the wave theory of the Dutch scientist Christiaan Huygens (developed at about the same time), that light corpuscles travel with greater velocity in the denser medium. Support for the wave theory came in the electromagnetic theory of Maxwell (1864) and the subsequent discoveries of Hertz and of Röntgen of both the very long and the very short waves Maxwell had included in his theory. The German physicist Max Planck proposed a quantum theory of radiation to counter some of the difficulties associated with the wave theory of light, and in 1905 Einstein proposed that light is composed of quanta (later called photons). Thus, experiment and theory had led around from particles (of Newton) that behave like waves (Huygens) to waves (Maxwell) that behave like particles (Einstein), the apparent velocity of which is unaffected by the velocity of the source or the velocity of the receiver. Furthermore it was found, in 1922, that the shorter-wavelength electromagnetic radiations (*e.g.,* X rays) have momentum such as may be expected of particles, part of which can be transferred to electrons with which they collide (*i.e.,* the Compton effect).

Quantum theory of light

Neutrinos and antineutrinos. Neutrinos and their antiparticles are forms of radiation similar to electromagnetic rays in that they travel at the speed of light and have little or no rest mass and zero charge. They too are produced by ultrahigh-energy particle accelerators and certain types of radioactive decay (see below).

Matter rays. Unlike X rays and gamma rays, some high-energy radiations travel at less than the speed of light. Some of these were identified initially by their particulate nature and only later were shown to travel with wavelike character. One example of this kind of radiation is the electron, first established as a negatively charged particle in 1897 by the English physicist Joseph John Thomson and later as the component of beta rays emitted by radioactive elements. The electron was shown by the American physicist Robert Millikan in 1910 to have a fixed charge and by George Paget Thomson, an English physicist, and the American physicists Clinton J. Davisson and Lester H. Germer (1927) to have wavelike as well as particulate character. Electrons classified as radiation have velocities that range from as low as 10^8 centimetres per second to almost the speed of light. The negative electron, still commonly called an electron, is identified more precisely as a negatron. In 1932 the American physicist Carl Anderson demonstrated the existence of a positive electron, generally called a positron and identified as one of the antiparticles of matter. The collision of a positron and an electron results in the intermediate production of a short-lived atomlike system called positronium, which decays in about 10^{-7} second into two gamma rays. Other entities commonly classified as matter when traveling with high velocity include the positively charged nucleus of the hydrogen atom, or proton; the nucleus of deuterium (*i.e.,* heavy hydrogen, the nucleus of which has double the mass of normal hydrogen's nucleus), or deuteron, also positively charged; and the nucleus of the helium atom, or alpha particle, which has a double positive charge. The more-massive positive nuclei of other atoms show similar wavelike properties when sufficiently accelerated in an electric field. All charged matter rays have a charge exactly equal to that of the negative or positive electron or to some integral multiple of that charge.

Major types

The neutron also is a matter ray. It is emitted in certain radioactive-decay processes and in fission, the process in which a nucleus splits into two smaller nuclei. The neutron decays in free space with a 12- to 13-minute half-life—*i.e.,* one-half of any given number of neutrons decay within 12–13 minutes, each into a proton and a negatron plus an antineutrino (see above). The mass of the neutron approximates that of the hydrogen atom, about 1,850 times the mass of the electron.

Another class of the so-called elementary particles is the meson, which comes both positively and negatively charged (*i.e.,* with the same charge as that of an electron), as well as electrically neutral. The masses of mesons are always greater than those of electrons, and most have a mass less than that of the proton; a few have slightly greater mass. Although all mesons are classified as matter rays when traveling at high velocities, they are so few that their chemical effects are not presently studied. Because they are part of the constant bombardment from free space to which all matter is constantly exposed, however, they may have considerable effects, such as contributing to the processes of aging and evolution.

THE STRUCTURE AND PROPERTIES OF MATTER

Matter in bulk comprises particles that, compared to radiation, may be said to be at rest, but the motion of the molecules that compose matter, which is attributable to its temperature, is equivalent to travel at the rate of hundreds of metres per second. Although matter is commonly considered to exist in three forms, solid, liquid, and gas, a review of the effects of radiation on matter must also include mention of the interactions of radiation with glasses, attenuated (low-pressure) gases, plasmas, and matter in states of extraordinarily high density. A glass appears to be solid but is actually a liquid of extraordinarily high viscosity, or a mixture of such a liquid and embedded microcrystalline material, which unlike a true solid remains essentially disorganized at temperatures much below its normal freezing point. Low-pressure gases are represented by the situation that exists in free space, in which the nearest neighbour molecules, atoms, or ions may be literally centimetres apart. Plasmas, by contrast, are regions of high density and temperature in which all atoms are dissociated into their positive nuclei and electrons. (For a detailed description of the various states of matter, see MATTER.)

The capability of analyzing and understanding matter depends on the details that can be observed and to an important extent on the instruments that are used. Bulk, or macroscopic, matter is detectable directly by the senses supplemented by the more common scientific instruments, such as microscopes, telescopes, and balances. It can be characterized by measurement of its mass and, more commonly, its weight, by magnetic effects, and by a variety of more sophisticated techniques, but most commonly by optical phenomena—by the visible or invisible light (*i.e.,* photons) that it absorbs, reflects, or emits or by which its observable character is modified. Energy absorption, which always involves some kind of excitation, and the opposed process of energy emission depend on the existence of ground-state and higher energy levels of molecules and atoms. A simplified system of energy states, or levels, is shown schematically in Figure 1. Such a system is exactly fixed for each atomic and molecular system by the laws of quantum mechanics; the "allowed,"

Absorption and quantum laws

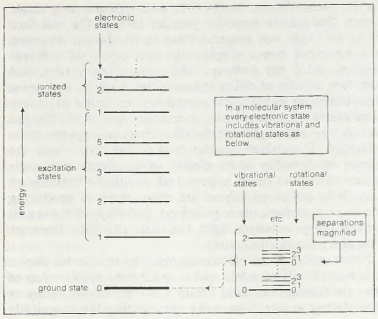


Figure 1: Energy states in molecular systems (see text).

or "permitted," transitions between levels, which may involve energy gain or loss, are also established by those same laws of nature. Excitation to energy levels above those of the energetically stable molecules or atoms may result in dissociation or ionization: molecules can dissociate into product molecules and free radicals, and, if the energy absorption is great enough, atoms as well as molecules can yield ions and electrons (*i.e.,* ionization occurs). Atomic nuclei themselves may exist in various states in which they absorb and emit gamma rays under certain conditions, and, if the nuclei are raised to, or by some process left in, energy states that are sufficiently high, they may themselves emit positrons, negatrons, alpha particles, or neutrons (and neutrinos) or dissociate into the nuclei of two or more lighter atoms. The resulting atoms may be similarly short-lived and unstable, or they may be extremely long-lived and quite stable.

THE EFFECTS OF RADIATION

The interaction of radiation with matter can be considered the most important process in the universe. When the universe began to cool down at an early stage in its evolution, stars, like the Sun, and planets appeared, and elements such as hydrogen (H), oxygen (O), nitrogen (N), and carbon (C) combined into simple molecules such as water (H_2O), ammonia (NH_3), and methane (CH_4). The larger hydrocarbons, alcohols, aldehydes, acids, and amino acids were ultimately built as a result of the action (1) of far-ultraviolet light (wavelength less than 185 nanometres) before oxygen appeared in the atmosphere, (2) of penetrating alpha, beta, and gamma radiations, and (3) of electric discharges from lightning storms when the temperature dropped and water began to condense. These simple compounds interacted and eventually developed into living matter. To what degree—if at all—the radiations from radioactive decay contributed to the synthesis of living matter is not known, but the occurrence of high-energy-irradiation effects on matter at very early times in the history of this world is recorded in certain micas as microscopic, concentric rings, called pleochroic halos, produced as the result of the decay of tiny specks of radioactive material that emitted penetrating products, such as alpha particles. At the termini of their paths, particles of this kind produced chemical changes, which can be seen microscopically as dark rings. From the diameters of the rings and the known penetrating powers of alpha particles from various radioactive elements, the nature of the specks of radioactive matter can be established. In some cases, alpha particles could not have been responsible for the effects observed; in other cases, the elementary specks that occupied the centres of the halos were not those of any presently known elements.

It can be readily surmised that some of the elements that participated in the evolution of the world were not originally present but were produced as the result of external high-energy bombardment, that some disappeared as the result of such processes, and that many compounds required for the living processes of organisms evolved as a consequence of the high-energy irradiation to which all matter is subjected. Hence, radiation is believed to have played a major role in the evolution of the universe and is ultimately responsible not only for the existence of life but also for the variety of its forms.

Fundamental processes involved in the interaction of radiation with matter

THE PASSAGE OF ELECTROMAGNETIC RAYS

The field concept. A discussion of this subject requires preliminary definition of a few of the more common terms. Around every particle, whether it be at rest or in motion, whether it be charged or uncharged, there are potential fields of various kinds. As one example, a gravitational field exists around the Earth and indeed around every particle of mass that moves with it. At every point in space, the field has direction in respect to the particle. The strength of the gravitational field around a specific particle of mass, m, at any distance, r, is given by the product of g, the universal gravitational constant, and m divided by the square of r, or gm/r^2. The field extends indefinitely in space, moves with the particle when it moves, and is propagated to any observer with the velocity of light. Newton showed that the mass of a homogeneous spherical object can be assumed to be concentrated at its centre and that all distances can be measured from it. Similarly, electric fields exist around electric charges and move with them. Magnetic fields exist around electric charges in motion and change in intensity with all changes in the accompanying electric field, with the magnetic field at any point being perpendicular to the electric field in free space. Any regular oscillation is time-dependent, as is any change in field strength with time.

Time-dependent electric and magnetic fields occur jointly; together they propagate as what are called electromagnetic waves. In an assumed ideal free space (without intrusion from other fields or forces of any kind, devoid of matter, and, thus, in effect without any intrusions, demarcations, or boundaries), such waves propagate with the speed of light in the so-called transverse electromagnetic mode—one in which the directions of the electric field, the magnetic field, and the propagation of the wave are mutually perpendicular. They constitute a right-handed coordinate system; *i.e.,* with the thumb and first two fingers of the right hand perpendicular to each other, the thumb points in the direction of the electric field, the forefinger in that of the magnetic field, and the middle finger in that of propagation. A boundary may be put on the space by appropriate physical means (bound space), or the medium may be something other than a vacuum (material medium). In either case, other forces and other fields come into the picture, and propagation of the wave is no longer exclusively in the transverse electromagnetic mode. Either the electric field or the magnetic field (a matter of arbitrary choice) may be considered to have a component parallel to the direction of propagation itself. It is this parallel component that is responsible for attenuation of energy of the waves as they propagate.

The transverse electromagnetic mode

Frequency range. Electromagnetic waves span an enormous range of frequencies (number of oscillations per second), only a small part of which fall in the visible region. Indeed, it is doubtful that lower or upper limits of frequency exist, except in regard to the applicability of present-day instrumentation. Figure 2 indicates the usual terminology employed for electromagnetic waves of different frequency or wavelength. Customarily, scientists designate electromagnetic waves by fields, waves, and particles in increasing order of the frequency ranges to which they belong. Traditional demarcations into fields, waves, and particles (*e.g.,* gamma-ray photons) are shown in the figure. The distinctions are largely of classical (*i.e.,* nonquantum) origin; in quantum theory there is no need for such distinctions. They are preserved, however, for common usage. The term field is used in a situation in which the wavelength of the electromagnetic waves is larger than the physical size of the experimental setup. For wave designation, the wavelength is comparable to or

Classification of electromagnetic regions

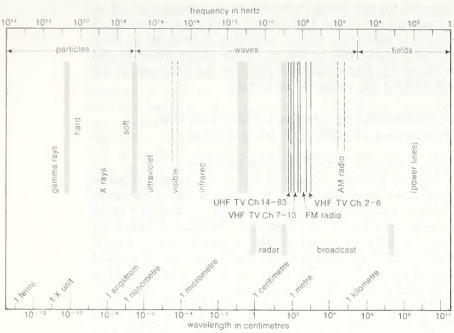


Figure 2: *Electromagnetic spectrum.*
Two scales are shown: the frequency scale, expressed in hertz, or cycles per second, and the wavelength scale, expressed in centimetres. Other units of wavelength customarily used to define certain regions of the spectrum are also given; *e.g.*, the nanometre (nm; 1/1,000,000,000 metre), the unit useful in the visible region, with an approximate range from 760 nm (red) to 380 nm (violet). Demarcations between the various classifications of radiations (*e.g.*, gamma rays and X rays) are not sharp.

smaller than the physical extent of the setup, and at the same time the energy of the photon is low. The particle description is useful when wavelength is small and photon energy is high.

Properties of light. The ordinary properties of light, such as straight-line propagation, reflection and refraction (bending) at a boundary or interface between two mediums, and image formation by mirrors or lenses, can be understood by simply knowing how light propagates, without inquiring into its nature. This area of study essentially is geometrical optics. On the other hand, the extraordinary properties of light do require answers to questions regarding its nature (physical optics). Thus, interference, diffraction, and polarization relate to the wave aspect, while photoelectric effect, Compton scattering, and pair production relate to the particle aspect of light. As noted above, light has dual character. It is the duality in the nature of light, as well as that of matter, that led to quantum theory.

Wave aspects of light. In general, radiation interacts with matter; it does not simply act on nor is it merely acted upon. Understanding of what radiation does to matter requires also an appreciation of what matter does to radiation.

When a ray of light is incident upon a plane surface separating two mediums (*e.g.*, air and glass), it is partly reflected (thrown back into the original medium) and partly refracted (transmitted into the other medium). The laws of reflection and refraction state that all the rays (incident, reflected, and refracted) and the normal (a perpendicular line) to the surface lie in the same plane, called the plane of incidence. Angles of incidence and reflection are equal; for any two mediums the sines of the angles of incidence and refraction have a constant ratio, called the mutual refractive index. All these relations can be derived from the electromagnetic theory of Maxwell, which constitutes the most important wave theory of light. The electromagnetic theory, however, is not necessary to demonstrate these laws.

Maxwell's theory of radiation

Double refraction. In double refraction, light enters a crystal the optical properties of which differ along two or more of the crystal axes. What is observed depends on the angle of the beam with respect to the entrant face. Double refraction was first observed in 1669 by Erasmus Bartholin in experiments with Iceland spar crystal and elucidated in 1690 by Huygens.

If a beam of light is made to enter an Iceland spar crystal at right angles to a face, it persists in the crystal as a single beam perpendicular to the face and emerges as a single beam through an opposite parallel face. If the exit face is at an angle not perpendicular to the beam, however, the emergent beam is split into two beams at different angles, called the ordinary and extraordinary rays, and they are usually of different intensities. Clearly, any beam that enters an Iceland spar crystal perpendicular to its face and emerges perpendicular to another face is of changed character—although superficially it may not appear to be changed. Dependent on the relative intensities and the phase relationship of its electric components (*i.e.*, their phase shift), the beam is described as either elliptically or circularly polarized. There are other ways of producing partially polarized, plane-polarized, and elliptically (as well as circularly) polarized light, but these examples illustrate the phenomena adequately.

Polarization of an electromagnetic wave can be shown mathematically to relate to the space-time relationship of the electromagnetic-field vector (conventionally taken as the electric vector, a quantity representing the magnitude and direction of the electric field) as the wave travels. If the field vector maintains a fixed direction, the wave is said to be plane-polarized, the plane of polarization being the one that contains the propagation direction and the electric vector. In the case of elliptic polarization, the field vector generates an ellipse in a plane perpendicular to the propagation direction as the wave proceeds. Circular polarization is a special case of elliptic polarization in which the so-described ellipse degenerates into a circle.

Kinds of polarization

An easy way to produce circularly polarized light is by passage of the light perpendicularly through a thin crystal, as, for example, mica. The mica sample is so selected that the path difference for the ordinary and the extraordinary rays is one-quarter the wavelength of the single-wavelength, or monochromatic, light used. Such a crystal is called a quarter-wave plate, and the reality of the circular polarization is shown by the fact that, when the quarter-wave plate is suitably suspended and irradiated, a small torque—that is, twisting force—can be shown to be exerted on it. Thus, the action of the crystal on the light wave is to polarize it; the related action of the light on the crystal is to produce a torque about its axis.

The ratio of the intensity of the reflected light to that of the incident light is called the reflection coefficient. This quantitative measure of reflection depends on the angles of incidence and refraction, or the refractive index, and also on the nature of polarization.

It can be shown that the reflection coefficient at any angle of incidence is greater for polarization perpendicular to the plane of incidence than for polarization in the plane of incidence. As a result, if unpolarized light is incident at a plane surface separating two media, reflected light will be partially polarized perpendicular to the plane of incidence, and refracted light will be partially polarized in the plane of incidence. An exceptional case is the Brewster angle, which is such that the sum of the angles of incidence and refraction is 90°. When that happens, the reflection coefficient for polarization in the plane of incidence equals zero. Thus, at the Brewster angle, the reflected light is wholly polarized perpendicular to the plane of incidence. At an air-glass interface, the Brewster angle is approximately 56°, for which the reflection coefficient for perpendicular polarization is 14 percent. Another extremely important angle for refraction is the critical angle of incidence when light passes from a denser medium to a rarer one. It is that angle for which the angle of refraction is 90° (in this case the angle of refraction is greater than the angle of incidence). For angles of incidence greater than the critical angle there is no refracted ray; the light is totally reflected internally. For a glass-to-air interface the critical angle has a value 41°48′.

Brewster angle

Dispersion. The variation of the refractive index with frequency is called dispersion. It is this property of a prism that effects the colour separation, or dispersion, of white light. An equation that connects the refractive index with frequency is called a dispersion relation. For visible light the index of refraction increases slightly with frequency, a phenomenon termed normal dispersion. The degree of refraction depends on the refractive index. The increased

bending of violet light over red by a glass prism is therefore the result of normal dispersion. If experiments are done, however, with light having a frequency close to the natural electron frequency, some strange effects appear. When the radiation frequency is slightly greater, for example, the index of refraction becomes less than unity (<1) and decreases with increasing frequency; the latter phenomenon is called anomalous dispersion. A refractive index less than unity refers correctly to the fact that the speed of light in the medium at that frequency is greater than the speed of light in vacuum. The velocity referred to, however, is the phase velocity or the velocity with which the sine-wave peaks are propagated. The propagation velocity of an actual signal or the group velocity is always less than the speed of light in vacuum. Therefore, relativity theory is not violated. An example is shown in Figure 3, in which a light source is initially pointed in the direction *A*. The source rotates in such a way that the velocity of the light image moves from *D* to *E* with a velocity *v* approximating *c*. Thus, the phase velocity with which the image moves from *A* to *B* is greater than *c*, but the relativity principle is not violated because the velocity of transmission of matter or energy does not exceed the velocity of light.

Propagation velocity of electromagnetic energy

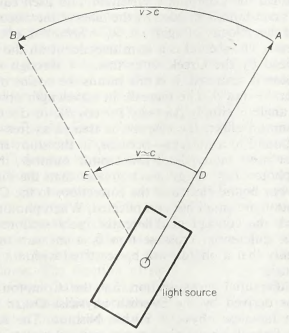


Figure 3: Contrast of phase velocity, *v*, and wave velocity, *c*. As the light source turns counterclockwise on its axis, the velocity with which the signal moves (*e.g.*, from *A* to *B*) can exceed the velocity of light without violation of the relativity principle (see text).

Electromagnetic waves and atomic structure. *Quantum concepts.* Quantum mechanics includes such concepts as "allowed states"—*i.e.*, stationary states of energy content exactly stipulated by its laws. The energy states shown in Figure 1 are of that kind. A transition between such states depends not only on the availability (*e.g.*, as radiation) of the precise amount of energy required but also on the quantum-mechanical probability of such a transition. That probability, the oscillator strength, involves so-called selection rules that, in general terms, state the degree to which a transition between two states (which are described in quantum-mechanical terms) is allowed. As an illustration of allowed transition in Figure 1, the only electronic transitions permitted are those in which the change in vibrational quantum number accompanying a change in electronic excitation is plus or minus one or zero, except that a $0 \leftrightarrow 0$ (zero-to-zero) change is not permitted. All electronic states include vibrational and rotational levels, so that the probability of a specific electronic transition includes the probabilities of transition between all the vibrational and rotational states that can conceivably be involved. Figure 1 is, of course, a simplified picture of a compendium of energy states available to a molecule (polyatomic structure)—and the selection rules are accordingly more involved in such a case. The selection rules are worked out by scientists in a process of discovery; the attempt is to state them systematically so that the applicable rules in an experimentally unstudied case may be stated on the basis of general principle.

Oscillator strength

Absorption and emission. In transit through matter, the intensity of light decreases exponentially with distance; in effect, the fractional loss is the same for equal distances of penetration. The energy loss from the light appears as energy added to the medium, or what is known as absorption. A medium can be weakly absorbing at one region of the electromagnetic spectrum and strongly absorbing at another. If a medium is weakly absorbing, its dispersion and absorption can be measured directly from the intensity of refracted or transmitted light. If it is strongly absorbing, on the other hand, the light does not survive even a few wavelengths of penetration. The refracted or transmitted light is then so weak that measurements are at best difficult. The absorption and dispersion in such cases, nevertheless, may still be determined by studying the reflected light only. This procedure is possible because the intensity of the reflected light has a refractive index that separates mathematically into contributions from dispersion and absorption. In the far ultraviolet it is the only practical means of studying absorption, a study that has revealed valuable information about electronic energy levels and collective energy losses (see below *Molecular activation*) in condensed material.

Experimental studies of the chemical effects of radiation on matter can be greatly forwarded by the use of beams of high intensity and very short duration. Such studies are made possible by employment of the laser, a light source developed by the American physicists Arthur L. Schawlow and Charles H. Townes (1958) from the application of one of the Einstein equations. Einstein suggested (on the basis of a principle of detailed balancing, or microscopic reversibility) that, just as the amount of light absorbed by a molecular system in a light field must depend on the intensity of the light, the amount of light emitted from excited states of the same system must also exhibit such dependency. In this fundamentally important idea of microscopic reversibility can be seen one of the most dramatic illustrations of the physical effects of radiation.

Principle of microscopic reversibility

Under any circumstance, the absorption probability in the ground state is given by the number of molecules (or atoms), N_g, in that state multiplied both by the probability, B_{ij}, for transition from state i to state j and by the light intensity, $I(\nu)$, at frequency symbolized by the Greek letter nu, ν; *i.e.*, $N_i B_{ij} I(\nu)$. Light emission from an excited state to the ground state depends on the number of molecules (or atoms) in the upper state, N_e, multiplied by the probability of spontaneous emission, A_{ji}, to the ground state plus the additional induced emission term, $N_j B_{ji} I(\nu)$, in which B_{ji} is a term that Einstein showed to be equal to B_{ij} and that relates the probability of such induced emission, so that in the general case in any steady-state situation (in which light absorption and emission are occurring at equal rates):

$$N_i B_{ij} I(\nu) = N_j [A_{ji} + B_{ji} I(\nu)] .$$

There is a well-developed theoretical relationship (not here presented) of a quantum-mechanical nature between A_{ji} and B_{ji}. Ordinarily, the light intensity, $I(\nu)$, is so low that the second term on the right can be neglected. At sufficiently high light intensities, however, that term can become important. In fact, if the light intensity is high, as in a laser, the induced-emission probability can easily exceed that of spontaneous emission.

Spontaneous emission of light is random in direction and phase. Induced emission has the same direction of polarization and propagation as that of the incident light. If by some means a greater population is created in the upper level than in the lower one, then, under the stimulus of an incident light of appropriate frequency, the light intensity actually increases with path length provided that there is enough stimulated emission to compensate for absorption and scattering. Such stimulated emission is the basis of laser light. Practical lasers such as the ruby or the helium-neon lasers work, however, on a three-level principle.

Particle aspects of light. The energy required to remove an orbital electron from an atom (or molecule) is called its binding energy in a given state. When light of photon energy greater than the minimum binding energy is incident upon an atom or solid, part or all of its energy

may be transformed through the photoelectric effect, the Compton effect, or pair production—in increasing order of importance with increase of photon energy. In the Compton effect, the photon is scattered from an electron, resulting in a longer wavelength, thus imparting the residual energy to the electron. In the other two cases the photon is completely absorbed or destroyed. In the pair-production phenomenon, an electron–positron pair is created from the photon as it passes close to an atomic nucleus. A minimum energy (1,020,000 electron volts [eV]) is required for this process because the energy of the electron–positron pair at rest—the total mass, $2m$, times the velocity of light squared ($2mc^2$)—must be provided. If the photon energy ($h\nu$) is greater than the rest mass, the difference ($h\nu - 2mc^2$), called the residual energy, is distributed between the kinetic energies of the pair with only a small fraction going to the nuclear recoil.

Mechanism of pair production

The photoelectric effect. The photoelectric effect is caused by the absorption of electromagnetic radiation and consists of electron ejection from a solid (or liquid) surface, usually of a metal, though nonmetals have also been studied. In the case of a gas, the term photoionization is more common, though there is basically little difference between these processes. In spite of experimental difficulties connected with surface-adsorbed gas and energy loss of ejected electrons in penetrating a layer of the solid into vacuum, early experimenters established two important features about the photoelectric effect. These are: (1) although the photoelectric current (*i.e.*, the number of photoelectrons) is proportional to the incident-light intensity, the energy of the individual photoelectrons is independent of light intensity; and (2) the maximum energy of the ejected electron is roughly proportional to the frequency of light. These observations cannot be explained in terms of wave theory. Einstein argued that the light is absorbed in quanta of energy equal to Planck's constant (h) times light frequency, $h\nu$, by electrons, one at a time. A minimum energy symbolized by the Greek letter psi, ψ, called the photoelectric work function of the surface, must be supplied before the electron can be ejected. When a quantum of energy is greater than the work function, photoelectric emission is possible with the maximum energy symbolized by the Greek letter epsilon, ε, of the photoelectron (ε_{max}) being stated by Einstein's photoelectric equation as equaling the difference between the photon energy and the work function; *i.e.*, $\varepsilon_{max} = h\nu - \psi$. Einstein's interpretation gave strong support for the quantum theory of radiation. Early experiments determined Planck's constant, h, independently through the above equation and also established the fact that an immeasurably small time delay is involved between absorption of a quantum of light and the ejection of an electron. The latter is clearly indicative of particle-like interaction.

Accurate and reliable values of the work function and ejection energy are now available for most solids; the chief obstacles to the development of such data were the difficulty of preparing clean surfaces and the energy loss of electrons in penetration into vacuum. The photoelectric threshold frequency, symbolized by the Greek letter nu with subscript zero, ν_0, is that frequency at which the effect is barely possible; it is given by the ratio of the work function symbolized by the Greek letter psi, ψ, to Planck's constant ($\nu_0 = \psi/h$). The photoelectric yield, defined as the ratio of the number of photoelectrons to that of incident photons, serves as a measure of the efficiency of the process. Photoelectric yield starts from a zero value at threshold, reaches a maximum value (about 1/1,000) at about twice the threshold frequency, and falls again when frequency is further increased. Some unusual alloys exhibit yields up to 100 times greater than normal (*i.e.*, about 0.1). Normally the yield depends also on polarization and angle of incidence of the radiation. Parallel polarization (polarization in the plane of incidence) gives higher yield than does perpendicular polarization, in some instances by almost 10 times.

Photoelectric threshold frequency

Cross section and Compton scattering. A useful concept in describing the absorption of radiation in matter is called cross section; it is a measure of the probability that photons interact with matter by a particular process. When the energy of each individual photon ($h\nu$) is much smaller than the rest energy of the electron (its mass times the velocity of light squared [mc^2]), the scattering of photons is described by a cross section derived by J.J. Thomson. This cross section is called the Thomson cross section, symbolized by the Greek letter sigma with subscript zero, σ_0, and is equal to a numerical factor times the square of the term, electric charge squared divided by electron rest energy, or $\sigma_0 = (8\pi/3)(e^2/mc^2)^2$. When the photon energy is equal to or greater than the electron's rest energy of ($h\nu \geq mc^2$), inelastic (*i.e.*, energy loss) scatterings begin to appear. One such is Compton scattering, in which an X ray or gamma ray (electromagnetic radiation from an atomic nucleus) experiences an increase in wavelength (reduction in energy) after being scattered through an angle. Arthur Holly Compton, an American physicist, correctly interpreted the effect by using the laws of classical relativistic mechanics. He showed that the increase in wavelength symbolized by the Greek letters delta and lambda, $\Delta\lambda$, is independent of the energy of the photon and is given by an expression in which the product of two terms appears. The first is a universal constant symbolized by the Greek letter lambda with subscript zero, λ_0, generally called the Compton wavelength, and itself equal to Planck's constant, h, divided by the mass of the electron at rest and the velocity of light; *i.e.*, $\lambda_0 = h/mc = 2.4 \times 10^{-10}$ centimetre. The second is a term dependent on the angle symbolized by the Greek letter theta, θ, through which the photon is scattered; it is one minus the cosine of that angle, or $1 - \cos\theta$. The increase in wavelength observed at that angle is simply $\Delta\lambda = \lambda_0(1 - \cos\theta)$. In discussing the Compton effect, the electron is treated as free—that is, not bound to a nucleus—because, in the study of that effect for most materials of low atomic number, the incident photon has energy much greater than the binding energy. For bound electrons, the corrections to the Compton relation are small but complicated. When photons are scattered, the concept of differential cross sections may be used; differential cross section is a measure of the probability that a photon will be scattered within a given small angle.

The differential cross section for the Compton process was derived by the Swedish physicist Oskar Klein and the Japanese physicist Yoshio Nishina. The Klein–Nishina formula shows almost symmetrical scattering for low-energy photons about 90° to the beam direction. As the photon energy increases, the scattering becomes predominantly peaked in the forward direction, and, for photons with energies that are greater than five times the rest energy of the electron, almost the entire scattering is confined within an angle of 30°. When averaged over the angle, the Klein–Nishina cross section shows variation with the incident photon energy. At low energy this cross section increases uniformly and approaches the classical Thomson value as energy is decreased; at high energy the cross section is inversely proportional to the energy. The energy distribution of Compton electrons (recoil or scattered electrons) and outgoing photons may also be derived from the Klein–Nishina theory. The result shows a wide distribution; for atoms of low atomic number and incident photon energies in the region of importance (*i.e.*, 1,000,000 to 100,000,000 eV), the probability of scattering per unit energy interval is fairly constant—except that, for the case of nearly total conversion of the photon energy into electron kinetic energy, a plot of energy versus angle shows a sharp, narrow peak. Thus, as a crude approximation, the average energy of a Compton electron is about half the incident photon energy.

Klein–Nishina formula

Compton scattering plays a key role in the interaction of matter with intermediate-energy gamma rays and high-energy X rays. For these radiations, it is almost the exclusive mechanism by which energy is transferred from the radiation and added to the matter. An example may be cited of the penetration of gamma rays from the radioactive substance cobalt-60 into a sample of water or aqueous solution. The electron density is approximately 3×10^{23} per millilitre. Taking the Compton cross section as approximately 3×10^{-25} square centimetre per electron, calculation yields a mean free path for Compton scatter-

ing of about 10 centimetres—that is to say, a photon will move about 10 centimetres between successive encounters with electrons. The dominant radiation effect produced by a gamma ray therefore is attributable to the recoil electron and the vast number of progeny (such as secondary and tertiary electrons) that are produced. These higher generation electrons are produced through electron-impact ionization (an electron is removed from an atom by the collision of another electron), a process that continues until barred either by energetic limitation or by low cross section. For cobalt-60 gamma rays the average Compton energy in a material of low atomic number, such as water, is approximately 600,000 eV.

Energy transfer from gamma rays

Pair production. Pair production is a process in which a gamma ray of sufficient energy is converted into an electron and a positron. A fundamental law of mechanics, given by Newton, is that in any process total linear (as well as angular) momentum remains unchanged. In the pair-production process a third body is required for momentum conservation. When that body is a heavy nucleus, it takes very little recoil energy, and therefore the threshold is just twice the rest energy of the electron; *i.e.,* twice its mass, m, times the square of the velocity of light, c^2, or $2mc^2$. Pair production also can occur in the field of an atomic electron, to which considerable recoil energy is thereby imparted. The threshold for such a process is four times the rest energy of an electron, or $4mc^2$. The total pair-production cross section is the sum of the two components, nuclear and electronic. These cross sections depend on the energy of the gamma ray and are usually calculated in an electron theory proposed by the British physicist P.A.M. Dirac through a method of approximation that is a simplification of a method (a "first approximation") devised by the German physicist Max Born (*i.e.,* a "first Born approximation"). The process is envisaged by Dirac as the transition of an electron from a negative to a positive energy state. Corrections are required for these cross sections at high energy, at high atomic number, and for atomic screening (the intrusion of the field of the electrons in an atom); these are normally made via numerical procedures. The fraction of residual energy, symbolized by the Greek letter alpha, unexpended in conversion of energy to mass, that appears in any one particle (*e.g.,* the electron) is thus given by the kinetic energy of that electron E_e minus its rest energy mc^2 divided by the energy of the gamma ray $h\nu$ (*i.e.,* the product of Planck's constant and the frequency of the gamma ray) minus twice the rest energy of the electron $2mc^2$, or $\alpha = (E_e - mc^2)/(h\nu - 2mc^2)$. Because the same equation applies to each of the two electrons that are formed, it must be symmetric about the condition that each of the particles has half the residual energy, symbolized by the Greek letter alpha, α (in excess of that conveyed to the "third body"); *i.e.,* that $\alpha = 0.5$. Below an energy of about 10,000,000 eV for the gamma ray, the probability for pair production (*i.e.,* the pair-production cross section) is almost independent of the atomic number of the material, and, up to about 100,000,000 eV of energy, it is also almost independent of the quantity α. Even at extremely high energies the probability that a certain fraction of the total available energy will appear in one particle is almost independent of the fraction as long as energy is comparably distributed between the two particles (excepting in cases in which almost all energy is dumped into one particle alone). Typical pair-production cross sections at 100 MeV (million electron volts) are approximately 10^{-24} to 10^{-22} square centimetre, increasing with atomic number. At high energies, approximately equal to or greater than 100 MeV, pair production is the dominant mechanism of radiation interaction with matter.

Born's "first approximation"

Clearly, as the photon energy increases, the dominant interaction mechanism shifts from photoelectric effect to Compton scattering to pair production. Rarely do photoelectric effect and pair production compete at a given energy. Compton scattering, however, at relatively low energy competes with the photoelectric effect and at high energy competes with pair production. Thus, in lead, interaction below 0.1 MeV is almost exclusively photoelectric; between 0.1 MeV and 2.5 MeV both photoelectric and Compton processes occur; and between 2.5 MeV and 100 MeV Compton scattering and pair production share the interaction. In the pair process the photon is annihilated, and an electron–positron pair is created. On the other hand, an electron or positron with energy approximately equal to or greater than 100 MeV loses its energy almost exclusively by production of high-energy bremsstrahlung (X rays produced by decelerating electric charges) as the result of interaction with the field of a nucleus. The cross section for bremsstrahlung production is nearly independent of energy at high energies, whereas at low energies the dominant energy-loss mechanism is by the creation of ionizations and excitations. A succession of bremsstrahlung and pair-production processes generates a cascade or shower in the absorber substance. This phenomenon can be triggered by an electron, a positron, or a photon, the triggering mechanism being unimportant as long as the starting energy is high. A photon generates a pair through pair production, and the charged particles generate photons through bremsstrahlung, and so on repeatedly as long as the energy is kept sufficiently high. With penetration into the substance, the shower increases in size at first, reaches a maximum, and then gradually decreases. Loss of particles by degradation to lower energies (in which the yield of bremsstrahlung is low), ionization loss, and production and absorption of low-energy photons eventually reduce the size of the cascade. The mathematical theory of cascades has been developed in great detail.

Bremsstrahlung production

X rays and gamma rays. When light of sufficiently high frequency (or energy equal to $h\nu$), independent of its source, is absorbed in a molecular system, the excited molecular state so produced, or some excited state resultant from it, may either interact with other molecules or decompose to produce intermediate or ultimate products; *i.e.,* chemical reactions ensue. Study of such processes is encompassed in the subject of photochemistry (see also below *Molecular activation*).

Electromagnetic waves of energy greater than those usually described as ultraviolet light (see Figure 2) are included in the classes of X rays or gamma rays. X-ray and gamma-ray photons may be distinguished by definition on the basis of source. They are indistinguishable on the basis of effects when their energy is absorbed in matter.

The total effect of X-ray or gamma-ray irradiation of matter, in the almost immediate time interval, is the production of high-energy electrons of energy related to that of the incident ray. Such electrons behave like beta rays (electrons emitted from atomic nuclei) or electrons from a machine source of the same energy. They lose energy by excitation and ionization of atoms and molecules of the systems they traverse. The amount of energy such an electron gives to an atom or molecule tends to exceed that deposited in photochemical processes, and the variety of initial physical (and consequent chemical) effects is more numerous and diverse. The situation is further complicated by the fact that the secondary electrons produced in ionization processes in which the input of energy is high may themselves initiate other ionization and excitation processes that can yield further chemistry, the totality of which is embraced in the title of radiation chemistry (see below *Molecular activation* and *Ionization and chemical change*).

THE PASSAGE OF MATTER RAYS

Heavy charged particles. Charged particles, such as atomic or molecular ions or molecular fragments, that travel in a material medium deposit energy along their paths, or tracks. If the medium is sufficiently thick, the velocity of the charged particle is reduced to near zero so that its energy is all but totally absorbed and is totally utilized in producing physical, chemical, and, in viable (living) matter, biologic changes. If the sample is sufficiently thin, the particle may ultimately emerge, but with reduced energy.

Ionization tracks

Linear energy transfer and track structure. The stopping power of a medium toward a charged particle refers to the energy loss of the particle per unit path length in the medium. It is specified by the differential $-dE/dx$, in which $-dE$ represents the energy loss and dx represents

the increment of path length. What is of interest to the radiation scientist is the spatial distribution of energy deposition in the particle track. In approximate terms, it is customary to refer to linear energy transfer (LET), the energy actually deposited per unit distance along the track (*i.e.*, $-dE/dx$). For not-so-fast particles, stopping power and LET are numerically equal; this situation covers all heavy particles studied so far in chemistry and biology but not electrons. In a refined study and redefinition of LET or restricted linear collision stopping power, a quantity symbolized by the letter L with subscript Greek letter delta, L_Δ, is defined as equal to the fractional energy lost ($-dE$) per unit distance traversed along the track (dl), or $L_\Delta = -(dE/dl)_\Delta$, in which the subscript delta (Δ) indicates that only collisions with energy transfer less than an amount Δ are included. The quantity L_Δ may be expressed in any convenient unit of energy per unit length. For Δ equal to 100 eV, even the most energetic secondary electrons (*i.e.*, electrons ejected by the penetrating particle) produce on average only about three subsequent ionizations. The latter, however, are closely spaced because of the low energy of the electron, and hence the corresponding energy density is high. It is higher yet for lower-energy secondary electrons. In contrast, for Δ much in excess of 100 eV, more subsequent ionizations are produced, but their spacing is increased significantly and the corresponding density of energy deposition is low. Since only the region of high energy density is of concern for many applications, the quantity L_{100} is often used to characterize LET.

Linear energy transfer

The bulk of energy deposition resulting from the passage of a fast-moving, charged particle is concentrated in the "infratrack," a very narrow region extending typically on the order of 10 interatomic distances perpendicular to the particle trajectory. The extent of the infratrack is dependent on the velocity of the particle, and it is defined as the distance over which the electric field of the particle is sufficiently strong and varies rapidly enough to produce electronic excitation. Inside the infratrack, electrons of the medium are attracted toward the trajectory of a positively charged particle. Many cross the trajectory, depositing energy on both sides. Consequently, the infratrack is characterized by an exceedingly high density of energy deposition and plays a vital role in determining the effects of ionizing radiation on the medium. (The magnitude of energy deposition in the infratrack is further increased by the preponderance of collective [plasma] excitations in that region.) The concept of the infratrack was developed by the American physicists Werner Brandt and Rufus H. Ritchie and independently by Myron Luntz. The region outside the infratrack is beyond the direct influence of the penetrating particle. Energy deposition in this outer region, or "ultratrack," is due primarily to electronic excitation and ionization by secondary electrons having sufficient energy to escape from the infratrack. In contrast to the infratrack, the ultratrack does not have well-defined physical bounds. Its spatial extent may reasonably be equated with the maximum range of secondary electrons transverse to the particle trajectory.

Concept of infratrack

For practical purposes, LET is associated with the main track, which may be thought of as including the infratrack and a portion of the ultratrack out to which energy density is still relatively high—*i.e.*, the region over which excitation is caused by secondary electrons of initial energy less than some value Δ, say 100 eV. Energy deposited in "blobs" or "short tracks" to the side of the main track, as described in the Mozumder–Magee theory of track effects (named for Asokendu Mozumder, an Indian-born physicist, and John L. Magee, an American chemist) is purposefully excluded. LET, so defined, characterizes energy deposition within a limited volume—*i.e.*, energy locally deposited about the particle trajectory.

Stopping power. By use of classical mechanics, Bohr developed an equation of stopping power, $-dE/dx$, given as the product of a kinematic factor and a stopping number. The kinematic factor includes such terms as the electronic charge and mass, the number of atoms per cubic centimetre of the medium, and the velocity of the incident charged particle. The stopping number includes the atomic number and the natural logarithm of a term that includes the velocity of the incident particle as well as its charge, a typical transition energy in the system (see Figure 1; a crude estimate is adequate because the quantity appears within the logarithm), and Planck's constant, h. Bohr's stopping-power formula does not require knowledge of the details of atomic binding. In terms of the stopping number, B, the full expression for stopping power is given by $-dE/dx = (4\pi Z_1^2 e^4 N/mv^2)B$, where Z_1 is the atomic number of the penetrating particle and N is the atomic density of the medium (in atoms/volume).

For a heavy incident charged particle in the nonrelativistic range (*e.g.*, an alpha particle, a helium nucleus with two positive charges), the stopping number B, according to the German-born American physicist Hans Bethe, is given by quantum mechanics as equal to the atomic number (Z) of the absorbing medium times the natural logarithm (ln) of two times the electronic mass times the velocity squared of the particle, divided by a mean excitation potential (I) of the atom; *i.e.*, $B = Z \ln (2mv^2/I)$.

Bethe's stopping number

Bethe's stopping number for a heavy particle may be modified by including corrections for particle speed in the relativistic range ($\beta^2 + \ln [1 - \beta^2]$), in which the Greek letter beta, β, represents the velocity of the particle divided by the velocity of light, and polarization screening (*i.e.*, reduction of interaction force by intervening charges, represented by the symbol $\delta/2$), as well as an atomic-shell correction (represented by the ratio of a constant C to the atomic number of the medium); *i.e.*, $B = Z (\ln 2mv^2/I - \beta^2 - \ln[1 - \beta^2] - C/Z - \delta/2)$.

The most important nontrivial quantity in the equation for stopping number is the mean excitation potential, I. Experimental values of this parameter, or quantity, are known for most atoms, but no single theory gives it over the whole range of atomic numbers because the calculation would require knowledge of the ground states and all excited states. Statistical models of the atom, however, come close to providing a theory. Calculations by the American physicist Felix Bloch in 1933 showed that the mean excitation potential in electron volts is about 14 times the atomic number of the element through which the charged particle is passing ($I = 14Z$). A later calculation gives the ratio of the potential to atomic number as equal to a constant (a) plus another constant (b) times the atomic number raised to the $-2/3$ power in which $a = 9.2$ and $b = 4.5$—*i.e.*, $I/Z = a + bZ^{-2/3}$. This formula is widely applicable. Other exact quantum-mechanical calculations for hydrogen give its mean excitation potential as equal to 15 eV.

Even though the basic stopping-power theory has been developed for atoms, it is readily applied to molecules by virtue of Bragg's rule (named for the British physicist William H. Bragg), which states that the stopping number of a molecule is the sum of the stopping numbers of all the atoms composing the molecule. For most molecules Bragg's rule applies impressively within a few percent, though hydrogen (H_2) and nitrous oxide (NO) are notable exceptions. The rule implies: (1) similarity of atomic binding in different molecules having one common atom or more, and (2) that the vacuum ultraviolet transitions, in which most electronic transitions are concentrated under such irradiation, involve energy losses much higher than the strengths of most chemical bonds.

Bragg's rule

The charge on a heavy positive ion fluctuates during penetration of a medium. In the beginning it captures an electron, which it quickly loses. As it slows down, however, the cross section of electron loss decreases relative to that for capture. Basically, the impinging ion undergoes charge-exchange cycles involving a single capture followed by a single loss. Ultimately, an electron is permanently bound when it becomes energetically impossible for the ion to lose it. A second charge-exchange cycle then occurs. This phenomenon continues repeatedly until the velocity of the heavy ion approximates the orbital velocity of the electron in Bohr's theory of the atom, when the ion spends part of its time as singly charged and another part as a neutral atom. The kinematic factor in the expression for stopping power is proportional to the square of the nuclear charge of the penetrating particle, and it is modified to account for electron capture as the particle slows

down. On slowing down further, the electronic energy-loss mechanism becomes ineffective, and energy loss by elastic scattering dominates. The mathematical expressions presented here apply strictly in the high-velocity, electronic excitation domain.

Range. The total path length traversed by a charged particle before it is stopped is called its range. Range is considered to be taken as the sum of the distance traversed over the crooked path (track), whereas the net projection measured along the initial direction of motion is known as the penetration. The difference between range and penetration distances results from scattering encountered by the particle along its path. For heavy charged particles with high initial velocities (those that are appreciable fractions of the speed of light), large-angle scatterings are rare. The corresponding trajectories are straight, and the difference between range and penetration distance is, for most purposes, negligible.

Computation of particle ranges

Particle ranges may be obtained by (numerical) integration of a suitable stopping-power formula. Experimentally, range is more easily measured than is stopping power. For heavy particles a critical incident energy in low-atomic-number mediums is 1,000,000 eV divided by the mass of the particle in atomic mass units (amu)—*i.e.*, 1 MeV/amu. For incident energies higher than this critical value, range is usually well-known, and computation agrees with experiment within about 5 percent. In the case of aluminum, which is the best studied material, the accuracy is within about 0.5 percent. For incident energies less than the critical value, however, range calculations are usually uncertain, and agreement with experiment is poor. The range–energy relation is often given adequately as a power law, that range (R) is proportional to energy (E) raised to some power (n); that is, $R \propto E^n$. Protons in the energy interval of a few hundred MeV conform to this kind of relation quite well with the exponent n equal to 1.75. Similar situations exist for other heavy particles. Measurements of range and stopping power are of great importance in particle identification and measurement of their energies. Many experimental data and computations are available for ranges of heavy particles as well as of electrons. The theory by which Bethe derived a stopping number is generally accepted as providing the framework for understanding the variation of range with energy, though in practice the mean excitation potential, I, must be obtained in many cases by experimental curve fitting.

Both stopping power and range should be understood as mean (or average) values over an ensemble of atoms or molecules, because energy loss is a statistical phenomenon. Fluctuations are to be expected. In general, these fluctuations are called straggling, and there are several kinds. Most important among them is the range straggling, which suggests that, for statistical reasons, particles in the same medium have varying path lengths between the same initial and final energies. Bohr showed that for long path lengths the range distribution is approximately Gaussian (a type of relationship between number of occurrences and some other variable). For short path lengths, such as those encountered in penetration of thin films, the emergent particles show a kind of energy straggling called Landau type (for the Soviet physicist Lev Landau). This energy straggling means that the distribution of energy losses is asymmetric when a plot is drawn, with a long tail on the high-energy-loss side. The intermediate case is given by a distribution according to Sergey Ivanovich Vavilov, a Soviet physicist, that must be evaluated numerically. There is evidence in support of all three distributions in their respective regions of validity.

Bragg peak

The ionization density (number of ions per unit of path length) produced by a fast charged particle along its track increases as the particle slows down. It eventually reaches a maximum called the Bragg peak close to the end of its trajectory. After that, the ionization density dwindles quickly to insignificance. In fact, the ionization density follows closely the LET. With slowing, the LET at first continues to increase because of the strong velocity denominator in the kinematic factor of the stopping-power formula. At low speeds, however, LET goes through a maximum because of: (1) progressive lowering of charge by electron capture, and (2) the effect of the logarithmic term in the stopping-power formula. In general, the maximum occurs at a few times the Bohr orbital velocity. A curve of ionization density (also called specific ionization or number of ion pairs—negative electron and associated positive ion—formed per unit path length) versus distance in a given medium is called a Bragg curve. The Bragg curve includes straggling within a beam of particles; thus, it differs somewhat from the specific ionization curve for an individual particle in that it has a long tail of low ionization density beyond the mean range. The mean range of radium-C′ alpha particles in air at normal temperature and pressure (NTP), for example, is 7.1 centimetres; the Bragg peak occurs at about 6.3 centimetres from the source with a specific ionization of about 60,000 ion pairs per centimetre.

Electrons. In the first Born approximation, inelastic cross section depends only on velocity and the magnitude of the charge on the incident particle. Hence, an electron and a positron at the same velocity should have identical stopping powers, which should be the same as that of a proton at that velocity. In practice, there is some difference in the case of an electron because of the indistinguishability of the incident and atomic electrons. In describing an ionization caused by an incident electron, the more energetic of the two emergent electrons is called, by convention, the primary. Thus, maximum energy loss (ignoring atomic binding) is half the incident energy. Incorporating this effect, the stopping number of an electron is given by a complicated expression that involves a different arrangement of the parameters found in the stopping number of heavy charged particles; *i.e.*,

Electron stopping number

$$B_e = \frac{Z}{2}\left[\ln mc^2\beta^2 E/2I^2(1-\beta^2) - (2\sqrt{1-\beta^2} - 1 + \beta^2)\ln 2 + (1-\beta^2) + \frac{1}{8}(1-\sqrt{1-\beta^2})^2 - 2\frac{C}{Z} - \delta\right].$$

This stopping-power formula has a wide range of validity, from approximately a few hundred electron volts to a few million electron volts in materials of low atomic number. For low velocities, the Born approximation gradually breaks down, and highly excited states begin to be inaccessible to transitions by virtue of small maximum energy transfer. Yet, with some corrections the electron-stopping-power formula may be extended down to about 50 eV. Below that value any stopping-power formula is of doubtful validity, even though it is certain that most of the energy is still being lost to electronic states down to a few eV of energy.

On the high-velocity side, relativistic effects increase electron-stopping power from about 1,000,000 eV upward. Except for the term δ attributable to polarization screening, the relativistic stopping power tends to infinity as the electron velocity approaches the speed of light ($v/c = \beta \rightarrow 1$). One-half of the stopping power, called the restricted stopping power, is numerically equal to the linear energy transfer and changes smoothly to a constant value, called the Fermi plateau, as the ratio β approaches unity. The other half, called the unrestricted stopping power, increases without limit, but its effect at extreme relativistic velocities (those very near the speed of light) becomes small compared with energy loss by nuclear encounters.

At extremely high velocities an electron loses a substantial part of its energy by radiative nuclear encounter. Lost energy is carried by energetic X rays (*i.e.*, bremsstrahlung). The ratio of energy loss by nuclear radiative encounter to collisional energy loss (excitation and ionization) is given approximately by the incident electron energy (E) in units of 1,000,000 eV times atomic number (Z) divided by 800; *i.e.*, $EZ/800$. For a large class of mediums (atomic number equal to or greater than 8; *i.e.*, that for oxygen), the electron stopping is dominated by bremsstrahlung radiation for energies greater than 100 MeV.

Cherenkov radiation. When the speed of a charged particle in a transparent medium (air, water, plastics) is so high that it is greater than the group velocity of light in that medium, then a part of the energy is emitted as Cherenkov radiation, first observed in 1934 by Pavel A. Cherenkov, a Soviet physicist. Such radiation rarely accounts for more

than a few percent of the total energy loss. Even so, it is invaluable for purposes of monitoring and spectroscopy. Cherenkov radiation is spread over the entire visible region and into the near ultraviolet and near infrared. The direction of its propagation is confined within a cone, the axis of which is the direction of electron motion.

Energy-transfer mechanism. At the low-velocity end of its path, an electron continues to excite electronic levels of atoms or molecules until its kinetic energy falls below the lowest (electronically) excited state (see Figure 1). After that it loses energy mainly by exciting vibrations in a molecule. Such a mechanism proceeds through the intermediary of temporary negative ion states, for direct momentum-transfer collisions are very inefficient. In a condensed medium (liquid, solid, or glass) very low-energy (less than 1 eV) electrons continue to lose energy by a process called phonon emission and by interaction with other low-frequency intermolecular motions of the medium.

Electron and heavy-particle straggling

An electron and a singly charged heavy particle with the same velocity have about equal stopping powers. Because of the small mass of the electron, however, the relative retardation (decrease in velocity per unit path length) is much more for it. This larger retardation for an electron means that, if an electron and a heavy particle start with the same velocity, the electron will have a much smaller range. Electron tracks show much more straggling and scattering compared with that of a heavy particle. The first effect results from the fact that the electron can lose a large fraction of its energy in a single encounter; the second is the result of small mass. A power law may be used to connect range and energy of electrons in a given medium—*i.e.,* the range is proportional to energy raised to a power n; as in the case of a heavy particle, the index n is slightly less than two at high energies. At low energies the relationship is such that the exponent is one or less. Many formulas and tables are available for stopping powers and for ranges of electrons as well as of heavy particles over a wide range of energies.

Neutrons. A neutron is an uncharged particle with the same spin as an electron and with mass slightly greater than a proton mass. In free space it decays into a proton, an electron, and an antineutrino and has a half-life of about 12–13 minutes, which is so large compared with lifetimes of interactions with nuclei that the particle disappears predominantly by such interactions.

Neutron beams may be produced in a variety of ways. A modern method is to extract a high-intensity beam from a nuclear reactor. A simpler but expensive device is one that employs a mixture of radium and beryllium. The reaction of the alpha (α) particles emitted by the radium with beryllium nuclei produces a copious output of neutrons. The neutron is a major nuclear constituent and is responsible for nuclear binding. A free neutron interacts with nuclei in a variety of ways, depending on its velocity and the nature of the target. Ordinary interactions include scattering (elastic and inelastic), absorption, and capture by nuclei to produce new elements. Unlike the electron, a neutron loses energy significantly through elastic collisions, because its mass is comparable to masses of atoms of low atomic number. (According to the laws of mechanics, in elastic collision, on the average, an object loses half its energy to another object of equal mass.)

The average fraction of energy transferred from a neutron per collision, symbolized by $(\Delta E/E)_{av}$, is twice the atomic mass number (A) of the struck atom divided by the square of the mass number plus one; *i.e.,*

$$(-\Delta E/E)_{av} = 2A/(A+1)^2.$$

Thus, only 18, 25, 42, 90, and 114 collisions are required to thermalize (reduce the energy of motion to that of the surrounding atoms) a fast neutron in hydrogen, deuterium, helium, beryllium, and carbon, respectively.

Pure absorption does not result in a new element, even though it is sometimes accompanied by emission of gamma rays. In certain cases of capture, radioactivity follows, often with production of beta (β) particles. In another class of interaction, a heavy charged particle is ejected (such as an α-particle or proton); the resultant nucleus is often but not always radioactive. As an example, the reaction of neutrons on boron to produce alpha particles provides the basis for alpha-particle welding. The principle of such welding, invented by the Soviet chemist V.I. Goldansky, is to deposit a thin layer of a boron (or lithium) compound in the interface between diverse materials, which is thereafter irradiated with neutrons. The high-energy α-particles produced from the nuclear reaction weld the materials together.

Neutron diffraction

Extraordinary interactions of the neutron are represented by diffraction, nuclear fission, and nuclear fusion. Diffraction, exhibited by low-energy neutrons (approximately equal to or less than 0.05 eV), demonstrates their wave nature and is consistent with de Broglie's hypothesis of the wave character of matter. Neutron diffraction complements X-ray technique in locating the positions of atoms in molecules and crystals, especially atoms of low atomic number such as hydrogen. Fission is the breakup of a heavy nucleus (either spontaneously or under the impact, for example, of a neutron) into two smaller ones with liberation of energy and neutrons. Spontaneous-fission rates and cross sections of fission induced by agencies other than the neutron are so small that in most applications only neutron-induced fission is important. Also, the neutron-induced-fission cross section depends on the particular isotope (species of an element with the same atomic number and similar chemical behaviour but different atomic mass) involved and the neutron energy. The fission process itself generates fast neutrons, which, when suitably slowed down by elastic scattering (a process called moderation), are again ready to induce more fission. The ratio of neutrons produced to neutrons absorbed is called the reproduction factor. When that factor exceeds unity, a chain reaction may be started, which is the basis of nuclear-power reactors and other fission devices. The chain is terminated by a combination of adventitious absorption, leakage, and other reactions that do not regenerate a neutron. At the power level at which a reactor operates, the loss rate always balances the generation rate through fission. The Hungarian-born American physicist Eugene P. Wigner, in the course of consideration of the possible effects of fast neutrons, suggested in 1942 that the process of energy transfer by collision from neutron to atom might result in important physical and chemical changes. The phenomenon, known as the Wigner effect and sometimes as a "knock on" process, was actually discovered in 1943 by the American chemists Milton Burton and T.J. Neubert and found to have profound influences on graphite and other materials.

The "knock on" process

Secondary effects of radiation

PURELY PHYSICAL EFFECTS

With respect to radiation effects the terms primary and secondary are used in a relative sense; the usage depends on the situation under study. Thus, ionization and excitation may be considered as primary with respect to some physical and chemical effects. For other chemical effects, production of free radicals (molecular fragments) may be considered as primary even though that process requires a much longer time for its accomplishment. Still longer times are involved in biologic processes, in which the end product of an earlier chemical reaction may be considered as primary.

Generally, an atomic solid (a material consisting of only one atomic species) exhibits little or no permanent chemical change upon irradiation. Important among the atomic solids are such materials as metals and graphite. Production of molecular carbon (C_2) or bigger clusters upon irradiation of carbon and graphite may, in a certain marginal sense, be considered a chemical change. Ionization of a condensed atomic medium followed by recombination regenerates the same atom, but its locale may be affected. For a molecular medium the situation is quite different. Excited electronic states are often dissociative for a molecule and yield chemically reactive radicals. Positive ions, similarly produced, can experience a variety of reactions even before neutralization occurs. Such an ion may fragment all by itself, or it may react with a neutral molecule in what is called an ion–molecule reaction. In

either case new chemical species are created. These transformed ions and radicals, as well as the electrons, parent ions, and excited states, are capable of reacting with themselves and with molecules of the medium, as well as with a solute (a dissolved substance) that may be present in homogeneous distribution. The end products of the reactions can be, on the one hand, new stable compounds or, on the other, regenerated molecules of the original species, as in the case of water irradiation.

Physical changes caused by irradiation

A variety of purely physical effects have been observed in different substances under irradiation. They may be broadly classified as: (1) structural change in the crystal, sometimes accompanied by change in the structural dimensions, (2) change in static mechanical properties, such as elasticity and hardness, (3) change in dynamic mechanical properties, such as internal friction and strain, and (4) changes in transport properties, such as heat conductivity and electrical resistivity. Such changes are considered below in *Tertiary effects of radiation on materials.*

MOLECULAR ACTIVATION

A molecule is considered activated when it absorbs energy by interaction with radiation. In this energy-rich state it may undergo a variety of unusual chemical reactions that are normally not available to it in thermal equilibrium. Of special importance is electronic activation—*i.e.,* production of an electronically excited state of the molecule (see Figure 1). This state can be reached (1) by direct excitation by photon absorption, (2) by impact of charged particles, either directly or indirectly through charge neutralization, or by excitation transfer from excited positive ions, and (3) by charge transfer in collision with (relatively) slow incident positive ions. Among the variety of ensuing processes is light emission, or luminescence.

Luminescence. The language of luminescence is clouded by history. Originally, fast luminescence was called fluorescence and slow (*i.e.,* delayed or protracted) luminescence was called phosphorescence. Present scientific practice is to define the terms on the basis of so-called quantum-mechanical selection rules: fluorescence is an allowed transition (*e.g.,* singlet–singlet) and occurs in a typical time of about 10^{-9} second; phosphorescence is a forbidden transition (*e.g.,* triplet–singlet) and may require 10^{-6} second or longer.

In the gas phase (gaseous state), an excited molecule either luminesces, undergoes a process called internal conversion, or undergoes dissociation. Luminescence is the rule for anthracene, whereas for water it is dissociation into hydrogen (H) and hydroxide (OH). As a rule, luminescence processes occur by default—that is to say, only if dissociation is energetically impossible or involves a complicated energy-transfer process or if internal conversion to a nonluminescing state is inefficient.

Fluorescence usually takes place from the lowest electronically excited state (see Figure 1); if higher states are excited they either dissociate or energetically cascade to the lowest excited state by one of several possible internal transition mechanisms before emission occurs. (A notable exception to this rule is afforded by azulene.)

A similar situation exists for triplet excited molecules. The rate of emission, however, is even slower, for in this case it is forbidden by selection rules. If the triplet excitation energy is insufficient for molecular-bond breakage (dissociation), the molecule may remain in a metastable state (one of apparent, not real, stability) for a long time until it either phosphoresces, undergoes internal conversion, or combines with other triplets. Such a combination produces a highly excited state, which has enough energy for dissociation. Some of the latter excited states are formed as singlets capable of light emission. This discussion relates to the more common, general features. There are also special cases, not discussed, that do not follow the general pattern.

Ionization phenomena. Ionization (see Figure 1) is that extreme form of excitation in which an electron is ejected, leaving behind a positive molecular ion. The minimum energy required for this process is called the ionization potential (IP). The actual energetics are described by the Franck–Condon principle, which simply recognizes that, during the extremely short time of an electronic transition, the nuclear configuration of a molecule experiences no significant change. As a consequence of this principle, in an optical process the ion is almost invariably formed in some kind of excited state by input of energy greater than the IP. Also, because of Franck–Condon restrictions, excitation of an inner electron may result in initial production of nonionized, superexcited molecules (suggested by R.L. Platzman, an American physicist) with energy exceeding the ionization potential. A superexcited molecule is short-lived and usually converts rapidly (in a time as short as 10^{-14} second) either to neutral products or to an ion plus a free electron with marked excess energy. The ion itself may fragment to give other species with excess kinetic or internal vibrational and rotational energy.

Franck–Condon principle

Excitation states. All the various kinds of excitation that occur in the gas phase may also take place in the condensed states of matter (liquid, glass, or solid), but their relative contributions may be affected. In addition, special activated states are produced for which there is no analogue in the gaseous state. They owe their existence to the collective behaviour of atoms and molecules in close proximity. The more important of them are the exciton state, the polaron state, the charge-transfer (or charge-separated) state, and the plasmon state.

The exciton state is a cooperative state of molecules in which the excitation energy belongs simultaneously to all.

In a polaron state an electron belongs to the association of molecules; but its motion is relatively slow so that it carries with it its own polarization field, which is described as "a cloud of virtual phonons." A solvated electron (an electron associated with a particular molecule or group of molecules) is an example of this.

The charge-transfer state is an excited state. In a certain sense, electronic excitation involves motion of an electron from a lower orbit to a higher one. Quantum mechanics notes that the electron does not revolve around an atomic nucleus in a precise classical orbit but rather that it occupies an orbital in which it is to be found with maximum probability in the location of the classical orbit. When a molecule in a condensed system is excited, the resulting electronic orbital may overlay one or more adjacent molecules, and, in that sense, the electron belongs to the group because its excitation level does not correspond to the electronic properties of a single, isolated molecule.

The plasmon state is a highly delocalized state formed collectively through Coulombian (electrostatic) interaction of weakly bound electrons. Energy losses, approximating 10–20 eV in most materials, resulting from formation of plasmon states are seen in the impact of electrons of a few tens of kilovolts energy on thin films. Both metals and nonmetals, including plastics, show plasma energy losses. The lost energy may reappear in the form of ultraviolet or visible radiation (Ferrell radiation, 1960); no chemical effect is known to have occurred from such losses.

Energy transfer. *Fluorescence and phosphorescence.* In general, a small, simple molecule luminesces in the ultraviolet, and a more complex one emits near the blue-violet end of the visible spectrum. Dye molecules, on the other hand, may emit throughout the visible region, including the red end. The ground electronic state of most molecules is a singlet state. Usually, therefore, the optically allowed emission, or fluorescence, is from the lowest excited singlet state to the ground state. The lowest triplet state of the molecule lies somewhat below the excited singlet. Light emission from this triplet state is forbidden by the quantum-mechanical selection rules, but it does occur by default when other processes are even less probable. Such emission is called phosphorescence. It is relatively weak, slow, and shifted toward longer wavelength. Triplet states may be produced from higher singlets by processes called internal conversion and intersystem crossing. The states may also be produced in excitation from the ground state by impact of relatively slow charged particles, such as electrons.

Much of the effect of optical radiation in a condensed system is not on the molecule in which the energy is initially absorbed but on a more remote molecule to which the energy is transferred in a variety of possible processes.

They include excitation transfer either directly between adjacent molecules, by a direct quantum-mechanical interaction of an excited molecule with a remote one at a distance of 40 angstroms (4×10^{-7} centimetre) or less, or by the so-called trivial process of fluorescence emission from one molecule and reabsorption by one at any distance. These processes are studied mostly in regard to fluorescence and phosphorescence phenomena.

Production and reactions of ions in solution

With high-energy radiation (such as that of electrons, X rays, and gamma rays), an additional mechanism involving ions is also available. In the case of a solute M in a solvent S, for example, a simplified description of some possible effects of radiation is represented by the following expressions, in which the symbol $\rightsquigarrow$ is read, "is acted upon by high-energy radiation to give" and e represents an ejected electron:

$$S \rightsquigarrow S^+ + e \qquad \text{(ionization of solvent)}$$

$$e + M \longrightarrow M^- \qquad \text{(electron attachment to solute)}$$

$$M^- + S^+ \longrightarrow \begin{cases} M^* + S \\ M + S^* \end{cases} \qquad \text{(neutralization)}$$

$$M^* \longrightarrow M + h\nu \qquad \text{(light emission)}$$

or

$$M^* \longrightarrow A + B \qquad \text{(product formation)}$$

Any actual process is considerably more complicated and involves a larger number of species.

Photographic process. One of the most important effects of radiation on matter is seen in photographic action. Apart from its various uses in art, commerce, and industry, photography is an invaluable scientific tool. It is used extensively in spectroscopy, in photometry, and in X-ray examinations. Also, photographic emulsion techniques have been widely used in the detection and characterization of high-energy charged particles. It is important to note that all speculation regarding the primary phenomena involves the notion that, in an energy absorption process, either direct or sensitized, a chloride (or other halide) ion in a silver halide lattice loses an electron. That electron is thereafter captured by a silver ion located at such a point in the lattice that under suitable conditions of exposure and development a silver grain grows to a size representative of the duration and intensity of the light exposure.

IONIZATION AND CHEMICAL CHANGE

Earlier in this section, the ionization phenomenon was briefly discussed as a special case of molecular activation. The ionization process, however, does have certain characteristic features. Most notably, the probabilities (or cross sections) for ionization by light (photoionization) and for ionization by charged-particle impact are different in magnitude and in lowest—radiation—energy of occurrence (*i.e.,* threshold behaviour) for the same atom or molecule. The photoionization cross section shows abrupt onset (*i.e.,* a step behaviour) to a high value at threshold, falling thereafter only gradually with increase of photon energy. Electron-impact ionization in simple atoms (such as hydrogen and helium) begins at the ionization potential, increases in direct proportion to the energy near the threshold, and shows a peak at an incident energy of about 100–200 eV. With molecules the behaviour is similar except that the peak is broad and much less pronounced. When the incident energy is high and the ejected electron has kinetic energy (energy of motion) largely in excess of its binding energy, the cross section for the process approaches a limit called the classical Rutherford value, after the British physicist Ernest Rutherford.

Electron-impact ionization

In general, the initial processes resulting from the action of high-energy radiation on matter involve the intermediate production and participation of positive ions (both stable and unstable), electrons, negative ions, excited species, and free radicals and atoms, which in turn may enter into the processes of classical reaction kinetics.

Ordinary low-energy (or optical) processes usually involve only excited species and free radicals and atoms—all formed by processes that do not involve outright transfer of electric charge (*i.e.,* electrons) between different atoms and molecules.

The important feature that characterizes the chemistry both of optical processes (photochemistry) and of high-energy radiation (radiation chemistry) is that they are conveniently employed and their kinetics studied at room temperature and lower.

Photochemistry. There are two "laws" of photochemistry. The first, the Grotthuss–Draper law (named for the chemists Christian J.D.T. von Grotthuss and John W. Draper), is simply: for light to produce an effect upon matter it must be absorbed. The second, or Stark–Einstein law (for the physicists Johannes Stark and Albert Einstein), in its most modern form is: one resultant primary physical or chemical act occurs per photon absorbed. The quantum yield of a particular species of product is the number of moles of that product divided by the number of einsteins of light (units of 6.02×10^{23} photons)—or the number of molecules of product per photon—absorbed. In the ideal case the quantum yield, frequently denoted by the Greek letters gamma, γ, or phi, Φ, is unity. In real cases, Φ may approach zero on the one hand—particularly if a back reaction is involved—or it may be of the order of 1,000,000, in which case the primary product may start a chain reaction, as in a clean, dry mixture of hydrogen (H) and chlorine (Cl). In the following chemical equations each symbol for an element stands for one atom, and the number of atoms bonded into a molecule is given as a subscript following the symbol, while the number of molecules precedes the formula; the arrow indicates the course of the reaction:

Grotthuss–Draper and Stark–Einstein laws

$$Cl_2 \xrightarrow{h\nu} 2Cl \qquad (1)$$

$$Cl + H_2 \longrightarrow HCl + H \qquad (2)$$

$$H + Cl_2 \longrightarrow HCl + Cl \qquad (3)$$

etc.,

in which reactions 2 and 3 reoccur repeatedly in a chain reaction. The symbol $\xrightarrow{h\nu}$ may be read "when a photon of light frequency, symbolized by the Greek letter nu, ν (which is always stipulated), is absorbed, gives." Because h is Planck's constant of action (approximately 6.6×10^{-27} erg second) and ν is expressed in reciprocal seconds (*i.e.,* second^{-1}), the product $h\nu$ indicates the energy absorbed per photon. Some reactions may give two primary products; *e.g.,*

$$CH_3CHO \xrightarrow{h\nu} CH_3 + HCO \text{ or } CH_4 + CO.$$

In that case, there are different quantum yields for each of the primary reactions, and the ratio of those yields varies with the frequency, ν, of the light absorbed.

Radiation chemistry. When a target is bombarded by a positive ion such as the hydrogen ion H^+ or the deuterium ion D^+ from a particle accelerator or the alpha particle $^4He^{2+}$ from nuclear decay, or indeed any high-energy heavy positive ion, the initial effects differ significantly from those of a high-energy electron. This situation results from the fact that, for the same kinetic energy, $\frac{1}{2}mv^2$, a particle of greater mass, m, travels with smaller velocity, v. The smaller the velocity of a particle of a particular charge in the domain of high (but not ultrarelativistic) velocities, the greater its probability of interaction with the medium traversed—that is to say, the greater the linear energy transfer. Thus, positive ions produce their initial effects close together in the ionization track in a condensed medium such as water (perhaps one or two angstroms, 1 or 2×10^{-8} centimetre, apart), whereas equally energetic electrons traveling through the same medium deposit energy in small collections called spurs, which may be 1,000 angstroms (10^{-5} centimetre) or so apart. The appearance of the excitation and ionization track has been likened to a rope (in the case of positive-ion bombardment), on the one hand, as compared with isolated beads on a string (in the

Nature of positive-ion tracks

case of electron bombardment), on the other. The dense track, as well as the isolated spurs, contains ions, excited molecules, and electrons; however, the distributions in the two essentially different types of track are so different that the ensuing chemical reactions (*i.e.*, the track effects) may be quite dissimilar. As an example, alpha-particle irradiation of pure water produces substantial yields of hydrogen and hydrogen peroxide (H_2O_2), whereas irradiation with beta particles, X rays, or gammas is essentially without effect. One of the reaction sequences suggested in overall considerations of the radiation chemistry of water is

$$H_2O \rightsquigarrow H + OH \quad (1)$$

$$2H \longrightarrow H_2 \quad (2)$$

$$2OH \longrightarrow H_2O_2; \textit{ i.e.}, HOOH, \quad (3)$$

in which reaction (1) summarizes the early chemical consequences both of ionization and of excitation. It has been suggested that reactions (2) and (3) occur with high probability in dense tracks (*e.g.*, of alpha particles) but that, in isolated spurs (as in fast-particle tracks), such reactions may occur only with low probability. In such a case, according to the American chemist A. Oliver Allen, the hydrogen atoms and OH radicals enter with somewhat greater probability into back-reaction chains with any $H_2 + H_2O_2$ already produced and existent in the body of the liquid:

$$H + HOOH \longrightarrow H_2O + OH \quad (4)$$

$$OH + H_2 \longrightarrow H_2O + H. \quad (5)$$

The H atom produced in reaction (5) thereupon enters into reaction (4), so that whatever small amounts of H_2 and H_2O_2 are actually produced in reactions (2) and (3) are consumed in reactions (5) and (4), respectively, and remain essentially undetectable no matter how long the reaction is run.

Mechanism of radiation chemical reactions

Radiation chemical reactions. In more detailed discussions of the mechanism of radiation chemical reactions, the roles of both excitation and ionization are considered. Information regarding the former is available from the extensive data of photochemistry; frequently, the initial excitation process leads to no significant chemical effect. By contrast, ionization may result in a large variety of chemical changes involving the positive ion, the outgoing electron, and the excited states resultant from charge neutralization, as well as (parent) positive-ion fragmentation and ion-molecule reactions. Some such consequences are summarized for a few cases.

Different channels of fragmentation from the same parent ion (*e.g.*, the propane ion $C_3H_8^+$), such as

$$C_3H_8^+ \longrightarrow C_3H_7^+ + H \qquad \text{atom elimination}$$

$$\left.\begin{aligned} C_3H_8^+ &\longrightarrow C_2H_5^+ + CH_3 \\ C_3H_8^+ &\longrightarrow C_2H_5 + CH_3^+ \end{aligned}\right\} \qquad \text{radical elimination}$$

$$C_3H_8^+ \longrightarrow C_3H_6^+ + H_2 \qquad \text{molecule elimination,}$$

compete unless barred by energetic considerations. Because ionization potentials of various possible fragments may differ greatly, charge localization may occur on only one of them. On the other hand, because the initial ionization rarely leads to the ground state of the positive ion, the energy is usually adequate for bond breakage.

Ion-molecule reactions such as that between a water ion and a molecule,

$$H_2O^+ + H_2O \longrightarrow H_3O^+ + OH,$$

are more important in the condensed phase, and fragmentation is more important in the gas phase. The parent ion in liquid water almost invariably undergoes ion–molecule reaction as indicated above. Many ion–molecule reactions have high cross sections. The same ion may undergo fragmentation or ion–molecule reaction, depending on circumstances. Thus, methane (CH_4), acted upon by high-energy gamma radiation, producing an electron, symbolized by

$$CH_4 \rightsquigarrow CH_4^+ + e^-,$$

may be followed by fragmentation,

$$CH_4^+ \longrightarrow CH_3^+ + H \text{ or } CH_2^+ + H_2,$$

as well as an ion–molecule reaction,

$$CH_4^+ + CH_4 \longrightarrow CH_5^+ + CH_3.$$

Electrons in chemical transformations

The electron ejected in an initial ionization process may further ionize and excite other molecules in its path, thus causing other chemical transformations. In addition, it may produce chemical changes of its own by dissociative attachment, as in carbon tetrachloride (CCl_4) and nitrous oxide (N_2O),

$$e^- + CCl_4 \longrightarrow CCl_3 + Cl^-$$

$$e^- + N_2O \longrightarrow N_2 + O^-,$$

and by formation of negative ions of either permanent or virtual (*i.e.*, very short-lived) nature. Many of the negative ions produced in a dissociation process are chemically reactive (H^-, O^-, etc.) as well. Virtual negative ions are almost invariably in a high vibrational state—*i.e.*, they are vibrationally hot.

The important point to note from this limited discussion of primary physical effects and their consequences in radiation chemistry is that in general each such effect is the progenitor of many ionizations and excitations, the distribution of which in space depends on the energy of the particle involved as well as on the system traversed. There is no single resultant primary process corresponding to the result of absorption of a single optical photon and thus no analogue to the concept of quantum yield in photochemistry.

In radiation chemistry, yields are conventionally reported on the purely empirical basis of the number of molecules of a particular kind produced (or destroyed) per 100 eV input of a particular type of radiation. In the radiolysis (radiation-induced decomposition) of cyclohexane, for example, by cobalt-60 gamma radiation or by electrons of about 2,000,000 eV of energy, the overall yield of hydrogen per 100 eV input is frequently given as approximately 5.6 or $G(H_2) \simeq 5.6$, in which the symbol G is read as "the 100-electron-volt yield of." Sometimes a small g is used to denote the 100-electron-volt yield of a postulated intermediate not directly determinable by measurement. Table 1 summarizes some typical G values.

Symbolism of radiation chemistry. The symbolism of radiation chemistry differs from that of photochemistry. The chemistry is somewhat more complicated, and the establishment of the variety of initial chemical processes is somewhat more of a chore. For the action of high-energy radiation on water, the variety of early products is typically indicated by the relation

$$H_2O \rightsquigarrow H, OH, H_2, H_2O_2, e_{aq}^-, \text{ etc.},$$

in which $\rightsquigarrow$ is read "acted upon by high-energy radiation, gives" and e_{aq}^- is the symbol for the hydrated electron. Particular note is addressed to the species e_{aq}^- (*i.e.*, an electron solvated by water) indicated in the same reaction. For many years there was an awareness in the radiation chemistry of water of the anomalous behaviour of the hydrogen atom, H, as compared with the same atom produced in established chemical processes. The anomaly was resolved on the one hand by John W. Boag and E.J. Hart, who spectroscopically observed the species e_{aq}^- in the spectral region predicted by Platzman, and on the other hand by Harold A. Schwarz and Gideon Czapski, who showed the existence of the ionic reducing species with charge of minus unity.

Time scales in radiation chemistry. The time scale characteristic of radiation chemistry ranges from the extraordinarily short time required for a fast electron to traverse a molecule (about 10^{-18} second) to the time required for es-

Table 1: Examples of Yield Studies at Room Temperature

state	system	radiation	yields
Gas	hydrogen and chlorine	alpha (radium)	$G(HCl) \sim 10^6$
	perfluoroethane	gamma (cobalt-60)	$G(CF_4) = 1.6$ $G(c\text{-}C_4F_8) = 0.30$ $G(C_2F_6) = 0.21$ $G(C_3F_{10}) = 0.14$
Liquid	ferrous sulfate and sulfuric acid in aerated water (Fricke solution)	gamma (cobalt-60)	$G(Fe^{3+}) = 15.6$
		alpha (from neutron irradiation of boron)	$G(Fe^{3+}) = 4.2$
		alpha (from neutron irradiation of lithium)	$G(Fe^{3+}) = 5.7$
	cyclohexane	1.8 MeV electrons	$G(H_2) = 5.6$ $G(C_6H_{10}) = 2.86$ $G[(C_6H_{11})_2] = 1.55$ $G(C_6H_{11} \cdot C_6H_{13}?) \sim 0.068$
		1.8 MeV or gamma (cobalt-60)	$G(C_6H_{10})/G[(C_6H_{11})_2] = 1.85$
		alpha (polonium-210)	$G(C_6H_{10})/G[(C_6H_{11})_2] = 3.0$
	benzene	1.5 MeV electrons	$G(H_2) = 0.022$ $G(C_2H_2) = 0.017$ $G(C_6H_6 \rightarrow$"polymer"$) \sim 0.8$
Solid	ferrocene	1 MeV electrons	$G(-\text{ferrocene}) = 0.0101$ $G(C_5 \text{ products}) = 0.0036$ $G(\text{inorganic Fe}) = 0.042$

Source: *Chemical and Engineering News.*

sential completion of some neutralization processes in very viscous media (about three hours). In between, there can be a variety of reactions involving intermediate formation and disappearance of the collections of the various species already discussed. The time-scale spread is so great that a *pt* scale (in which *pt* is defined as minus the logarithm, to the base 10, of the time [in seconds] *t* [*i.e.*, $-\log_{10} t$]) is conveniently employed. Actual observances in the long time-scale region follow fairly well-established chemical practice. The short-time region, on the other hand, presents interesting challenges. The Van de Graaff generator and the linear accelerator both made possible irradiations by electrons and X rays in the microsecond (10^{-6} second) region, and spectroscopic devices were quickly devised to make observations in that region. Improvements in irradiation technique (with X rays by Herbert Dreeskamp and Milton Burton, and with ultraviolet by Paul K. Ludwig and Juan d'Alessio) and in observation techniques in the study of luminescence extended precision of observation to 5×10^{-10} second in the work of William P. Helman.

Chemical studies with linac

John K. Thomas combined use of a fast linac (linear accelerator) with Cherenkov radiation as a marker to extend chemical studies into the same region. Use of the same radiation as a light source for spectroscopic observation of the chemistry produced by a traveling electron front (from a linac) made possible actual observations in the time range of (2 to 4) $\times 10^{-11}$ second. Table 2 summarizes an approximate time scale of events.

Table 2: Approximate Time Scale of Events in a Condensed System

$-\log t$ (seconds) $\equiv pt$	events	stage
18	fast electron traverses molecule	physical
17	1-MeV proton traverses molecule; time for energy loss to fast secondary electrons	physical
16	time for energy loss to electronic states (Franck–Condon process)	physical
14	fast ion–molecule reactions involving hydrogen-atom transfer; molecular vibration; fast dissociation	physicochemical
12	electron thermalizes; self-diffusion time scales for liquids of low molecular weight	physicochemical
11	dielectric relaxes in water; neutralization time for polar media	physicochemical
10	spur* formed	physicochemical
9	spur* reactions	physicochemical
8	intratrack reactions completed	physicochemical
7	neutralization times in media of low viscosity and low dielectric constant	physicochemical
6		physicochemical
4	escape time for electrons in media of low viscosity and low dielectric constant	chemical
3		chemical
0	radiative lifetime of triplet excited states	chemical
−2	neutralization times for media of high viscosity and low dielectric constant	chemical
−4		chemical

*A spur is a region about 2×10^{-7} centimetre in diameter in a liquid and of high ionic and excitation density.

Source: *Advances in Radiation Chemistry*, Wiley-Interscience.

Tertiary effects of radiation on materials

The electrons liberated by high-energy irradiation that have sufficient energy cause further ionizations in which additional electrons are produced. Some of these second generation electrons also cause additional ionizations, and this process continues until their remaining energy becomes inadequate. Even though this process goes through several generations of events, it actually takes little time and thus appears as an impact phenomenon as far as radiation-induced chemical changes are concerned. For this purpose, then, they may be considered as primary. Fast chemical changes induced by radiation may take time on the order of nanoseconds (a nanosecond is 10^{-9} second) or less to complete. Slower reactions involving relatively less reactive scavengers (reagents that eliminate residues) in dilute concentrations may require a time span of approximately 10^{-4} second.

This section is concerned with radiation effects measurable on much longer time scales, arbitrarily greater than about one minute. Attention is here addressed to physical changes in the solid state, about which there is a wealth of experimental information. It should again be emphasized that little chemical change is expected in an atomic medium in which the absorption of ionizing radiation also results ultimately in structural changes and induced imperfections. With neutron irradiation, in addition to specific nuclear interactions, one gets "knocked off" atoms or ions (note the discussion of the Wigner effect in *Neutrons* above). These ions quickly capture electrons and the resulting neutral atoms then travel on. Even though a small effect occurring in ionization and electronic excitation attributable to knocked off ions cannot be denied, it is believed that this effect is small compared with that brought about by the neutral knock offs in the form of structural changes.

HEATING EFFECTS

The simplest ultimate effect of absorption of radiation is heating. It can be argued that, for ionizing radiation of low linear energy transfer, the heating effect is negligible. A spur created by such low-LET radiation is a small spherical region in which the energy deposit is localized in isolation. The temperature rise, ΔT, of the spur above the surrounding temperature has a space-time dependence that, by hypothesis, has a statistical distribution, called Gaussian, because of random superposition of events leading to the heating process. The temperature rise at a point located a distance r away from the spur centre at time t is given by the equation

Temperature rise in spur

$$\Delta T(r, t) = \Delta T_{max}\left(1 + \frac{4\gamma t}{a^2}\right)^{-3/2} \exp\left(-\frac{r^2}{a^2 + 4\gamma t}\right),$$

in which a is the initial spur-size parameter, the Greek letter gamma, γ, is the thermal diffusivity of the medium (equal to heat conductivity divided by the product of density and specific heat at constant volume), and ΔT_{max} is the initial maximum temperature rise at the centre of the spur. Taking reasonable values for energy deposition (30 eV) and spur size a (20 angstroms, or 2×10^{-7} centimetre) and using (for water) density equal to one gram per cubic centimetre and specific heat 4×10^7 ergs per gram-degree, ΔT_{max} may be estimated to be 30° C (54° F). The time required for the central temperature to drop to half its initial value (*i.e.*, $t_{1/2}$) is given by $(1 + 4\gamma t_{1/2}/a^2)^{3/2} = 2$. With the thermal diffusivity, symbolized by the Greek letter gamma, γ, equal to 10^{-3} centimetre squared per second for water, $t_{1/2} = 6 \times 10^{-12}$ second. The conclusion is that for low-LET radiation the local temperature rise is too small and too brief to have any appreciable

chemical (or physical) effect. It is particularly notable that the actual temperature rise is smaller than that estimated here because part of the deposited energy is invariably utilized in ionization, dissociation, and similar processes. This part of the energy resides in the potential form and is not completely available for heating. With high-LET ionizing radiations (such as fission fragments, stripped nuclei, and α-particles), the situation is somewhat different. In such a case, a large amount of energy is deposited per unit path length, resulting in cylindrical tracks (rather than spherical spurs). The equation for temperature rise in this case is written in a form much like the equation for spur geometry; that is,

$$\Delta T(r, t) = \Delta T_{max}\left(1 + \frac{\gamma t}{a^2}\right)^{-1} \exp\left(\frac{-r^2}{a^2 + 4\gamma t}\right),$$

except that in this case a is the initial size parameter for the track cylinder and ΔT_{max} is the maximum initial temperature rise on its axis. For fission fragments with LET of 500 eV per angstrom (*i.e.*, 10^{-8} centimetre) and a equal to 20 angstroms, ΔT_{max} for water is 1.6×10^4 K. Admittedly, this figure is an overestimate for reasons similar to those that apply for low-LET radiations, but it is believed that the temperature rise is high and may approach 10,000 K.

Temperature rise to 10,000 K

The time for this temperature to drop to half the initial value is given by $t_{1/2} = a^2/4\gamma$, which in this case is estimated to be about 10^{-11} second. This time is not much larger than the corresponding time for survival of an isolated spur. Because of the high local temperature, however, the reaction time of the radiation-produced intermediates is also very small. In an intermediate + substrate (*i.e.*, solvent) reaction, for example, with an activation energy of approximately eight kilocalories per mole, the rate constant, k, for such a typical pseudo-first-order reaction may be written, according to the simple usage of chemical kinetics, in the form

$$k \simeq 10^{-11} \exp(-8{,}000/RT) \text{ cubic centimetre per molecule per second,}$$

in which R is the gas constant in units of two calories per degree and T is the absolute temperature. For a substrate concentration of 10^{22} molecules per cubic centimetre and a temperature of 10,000 K, $-d(\ln \nu)/dt = 6.6 \times 10^{10}$ reciprocal seconds, in which the Greek letter nu, ν, here denotes the concentration of intermediates still existent at time t. Therefore, the time required for the intermediate concentration to drop to an "e-folding" value (*i.e.*, to fraction $1/e$) is approximately 1.5×10^{-11} second, a time that compares favourably with the duration of the temperature pulse. The conclusion is that for very high LET radiation there is indeed a high degree of local heating, and, even though the heat pulse survives only a short time, that time still is long enough to bring about the acceleration of reactions between the short-lived intermediates and the ambient substrate.

CRYSTAL-LATTICE EFFECTS

In neutron irradiation of a solid, atoms are dislodged from normal lattice positions and set in motion (the Wigner effect). The fractional amount of energy transfer depends, as in any elastic collision, on the mass ratio of the neutron to that of the recoil atom. Thus, in graphite a carbon atom, on first collision with a neutron of 1,000,000-eV (produced, say, in a fission process), receives a kinetic energy of approximately 10^5 eV, which is large compared with its binding energy in the lattice (about 10 eV). It is estimated that the 1,000,000-electron-volt neutron strikes about 60 carbon atoms before it is thermalized or its speed is so much reduced that it cannot knock off other carbon atoms. Much of the structural damage caused by radiation is attributable to these (relatively heavy) carbon atoms rather than to the original neutron. In this sense, radiation damage by fast neutrons may be viewed as an indirect action. Slowing of the fast carbon (or other dislodged) atoms is basically governed by interaction time. This fact means that the stopping is light in the beginning of the journey of a dislodged atom and results only in occasional displacement of atoms. Toward the end of its career a large number of atoms are displaced in quick succession along a row, and finally a large amount of residual energy is dumped locally into a relatively small group of atoms. This process generates the displacement (or thermal) spike; the local temperature rise is estimated to be about 1,000 K. Even though the temperature rise lasts only about 10^{-11} second before the track is cooled down, this duration is enough for permanent structural damage. At least a part of the swelling of graphite under reactor-neutron irradiation is the result of this local heating; another part originates in change in lattice dimension under irradiation.

Cause of structural damage

The high temperature rise in a thermal spike probably results in local melting of the solid. Evidence in that direction has been obtained from a study of beta-brass (an alloy consisting of equal numbers of atoms of copper and zinc) under neutron bombardment at low temperature. Before irradiation, the alloy structure is ordered: each copper atom is surrounded by eight zinc atoms as nearest neighbours and vice versa. After irradiation, a general random rearrangement of the atoms can be detected, presumably the result of melting and refreezing.

Long-term effects of radiation on crystals are numerous, and the magnitudes of these effects depend on the crystal structure and previous history. Only some general features of these effects are recounted here.

1. Radiation damage may be thought to consist of pairs of interstitial atoms ejected from their normal lattice sites and the corresponding vacancies left behind. A vacancy-interstitial pair is called a Frenkel defect.

Frenkel defect

2. A solid has a tendency to recover spontaneously from radiation damage. If it were not for this property, it would indeed be extremely difficult to operate nuclear reactors that are permitted to heat up periodically to remove the effect in the graphite core. The healing (or so-called annealing) is presumably attributable to the recombination of interstitial atoms and vacancies, thereby removing Frenkel defects. It is not necessary that an interstitial atom always recombine with its corresponding vacancy. Often it may recombine with a vacancy that resembles the one that it left; the result is approximate restoration of the original properties of the crystal. Such annealing is facilitated by the increased mobility of the vacancies and interstitials at higher temperature. At a particular temperature called the annealing temperature, the healing becomes fast and essentially complete. The same substance may have somewhat different annealing temperatures depending on the particular property under study. Many experiments on radiation damage must be carried out at low temperatures to freeze in the defects produced. Pure metals are the most easily annealed substances. Annealing temperatures in such cases are relatively low. Accordingly, the annealing temperature for the increase of electrical resistance in pure copper is only around 40 K. On the other hand, changes in elastic modulus and hardness, such as are required to produce tuning-fork characteristics, persist up to room temperature—namely, 293 K. Quick annealing in pure metals is directly attributable to the high mobility of atoms in perfectly ordered structures. At the other extreme are organic solids, particularly polymers, that are composed of large molecules. In this case, the damage originates in the breaking of bonds that ordinarily do not rejoin in the original manner but instead produce chemically different material.

3. In simple metals irradiation decreases conductivity for both heat and electricity. Conduction of both in metallic crystals is attributable to their ordered structure. The more perfect the structure, the better is the conduction. Frenkel defects, generated by irradiation, therefore decrease both conductivities. In extreme cases conductivity decrease of orders of magnitude has been observed. With moderate irradiation, however, both thermal and electrical conductivities decrease usually by half. The thermal conductivity of graphite falls to roughly half the unirradiated value with an exposure of 3×10^{20} neutrons per square centimetre at room temperature. Like other property changes, this effect also can be annealed at elevated temperatures with concomitant release of stored energy. Energy storage in graphite amounts to about 200 calories per gram per 10^{20} neutrons per square centimetre total flux. Interstitial car-

Change in conductivities

bon atoms produced in the irradiation scatter electrons and thus decrease electrical conductivity. The pattern of conductivity decrease and increase depends on the nature of the graphite and the duration of exposure in a reactor. With ceramic materials, loss of thermal conductivity by a factor of about 3 to 5 may be observed under conditions in which the decrease is about one-half in graphite. In mica, on the other hand, the change is somewhat less than in graphite.

4. Hardness and ductility depend on perfection of the crystal structure. It is thus found that irradiation results in a loss of ductility and an increase in hardness. Such effects are attributed to glide-plane obstruction in the crystal. Most structured materials become harder, less ductile, and sometimes more brittle as the result of neutron irradiation. Similarly, most polymers also lose ductility on irradiation. In a certain sense radiation-induced damage to the crystal structure is qualitatively similar to that produced by cold-working (for example, by hammering). Neutron irradiation of pure copper, which is naturally soft at room temperature, makes it so hard that it can be made to sing like a tuning fork. Graphite experiences increase in strength and hardness upon irradiation. Annealing is faster at elevated temperatures; also, damage is less when the irradiation is at a higher temperature. A similar effect is seen for the compressive stress-strain curve. Studies of dynamic properties in ceramics indicate a saturation effect at large doses.

5. As was discussed above, irradiation causes expansion and lattice distortion in most cases. A perfect crystal of graphite consists of planes of carbon atoms layer upon layer. When irradiated by neutrons, graphite expands perpendicular to the base plane and contracts slightly parallel to it. After moderate exposure in a nuclear reactor, the expansion is approximately 1 percent for a flux of 10^{20} neutrons per square centimetre. The actual amount of expansion, of course, depends on the fabrication history and operating temperature of the graphite. Expansion of moderator materials such as graphite is of considerable importance in the design of nuclear reactors. Even a small percentage change in dimension can result in large total change in the reactor structure; if this change is not allowed for in the engineering design of the reactor, it may well create strained operating conditions eventually leading to failure. (M.Bu./A.Moz./M.Lu.)

Influence on nuclear-reactor design

SURFACE EFFECTS

A surface is distinct from bulk matter in that it constitutes the physical interface with the environment. Whether or not a metal will corrode in salt water, for example, or how much resistance to wear is inherent in the design of a bearing are concerns that relate primarily to the physical condition of surfaces. The latter, in turn, may be selectively modified by the application of coatings or by the action of radiation, or by both. Three of the most common examples of surface modification by radiation—ultraviolet curing, ion implantation, and sputtering—are considered here.

Ultraviolet curing

Ultraviolet curing is a process in which polymers, generally employed as coatings, are irradiated by ultraviolet light. Such action produces electronic excitation and ionization of the long chain molecules that make up the polymer, either directly or through the mediation of imbedded, light-sensitive "activators." This results in intermolecular bonding, a process called cross-linking. The entire polymeric coating, typically on the order of tenths of millimetres thick (depending on the application), becomes so highly cross-linked as to take on the character of a single giant molecule. The major effects of ultraviolet irradiation of polymers include reduction of friction, increased resistance to wear, increased hardness, and increased resistance to attack by acids and other corrosive agents. Ultraviolet curing is employed for diverse purposes ranging from the formation of "no-wax" coatings on floor tiles to application in the photolithographic process integral to the fabrication of solid-state electronic devices.

Ion implantation

Ion implantation involves the irradiation of solids by beams of energetic ions emanating from particle accelerators. Typical energies employed are on the order of 100 keV (100,000 electron volts). Typical depths of penetration are on the order of several thousand angstroms, depending on energy, ion type, and target material. In ion implantation, virtually any atomic species can be embedded to predetermined depths and with predetermined concentration profiles in any target material so as to modify the surface characteristics without affecting desirable bulk properties. A typical example is the implantation of titanium in iron alloys to reduce wear of bearings and gears. A particularly promising technique was developed by physicists Michael W. Ferralli and Luntz, in which vacuum deposition of polymeric coatings on metallic substrates and simultaneous ion-beam irradiation act to produce implanted hydrocarbon films. The latter can be made to vary in carbon-to-hydrogen ratio from very high values—with the implanted region having some characteristics of diamond—to values on the order of unity and corresponding polymeric characteristics. This is accomplished by a process called preferential sputtering (see below). The films so produced are highly resistant to corrosion and appear to possess important bio-compatibility properties, making them suitable for applications in, for example, the treatment of the surfaces of surgical implants such as artificial hip joints. Such effects of ion implantation result in part from structural changes induced by radiation damage (*e.g.,* implantation of boron or phosphorus in steel can render the surface amorphous so as to eliminate grain boundaries and other corrosion-sensitive sites), and in part from chemical changes arising from bonding of the implanted species with constituents of the substrate.

Sputtering

Sputtering is a process in which atoms, ions, and molecular species in the surface of a target material are ejected under the action of ion-beam irradiation. Energies typical of ion implantation are employed and, while any ion type may be used, noble (or rare) gases such as argon and neon are most common. The latter avoid unwanted chemical interactions between the ions of the beam and the substrate. Sputtering results from several interaction mechanisms. Conceptually, the simplest is rebound sputtering, in which an incident ion strikes an atom on the surface, causing it to recoil into the target. The recoiling atom promptly collides with a neighbouring atom in the target, rebounds elastically, and is ejected from the surface. A similar but somewhat more complex mechanism is recoil sputtering, in which a struck, recoiling surface atom undergoes a random sequence of elastic scatterings in the target material, ultimately migrating back to, and through, the surface. Yet another mechanism is prompt thermal sputtering, in which energized atoms in thermal spikes created close to the surface escape through the surface before annealing occurs. Certain materials (*e.g.,* crystalline alkali halides) are prone to electronic sputtering, in which energy associated with electronic excitations induced by the incident ion is transformed into atomic recoil kinetic energy, often sufficient to cause the ejection of ions through the surface. By means of any of these various mechanisms, several atoms may be sputtered for each ion incident on the target. The number of atoms sputtered per incident ion is called the sputtering yield.

Surface modifications caused by sputtering are characterized as structural (*e.g.,* phase conversion from crystalline to amorphous and vice versa), topographical (*e.g.,* alteration of the shape of surface protrusions such as grain boundaries, development of facets, and the removal of surface contaminants), electronic (*e.g.,* radiation-induced chemical changes), and compositional (*e.g.,* preferential sputtering of a particular atomic species resulting in changes in the composition of alloys). (M.Lu.)

Biologic effects of ionizing radiation

The biomedical effects of ionizing radiation have been investigated more thoroughly than those of any other environmental agent. Evidence that harmful effects may result from small amounts of such radiation has prompted growing concern about the hazards that may be associated with low-level irradiation from the fallout of nuclear weapons, medical radiography, nuclear power plants, and other sources.

Assessment of the health impact of ionizing radiation requires an understanding of the interactions of radiation with living cells and the subsequent reactions that lead to injury. These subjects are surveyed in the following sections, with particular reference to the principal sources and levels of radiation in the environment and the different types of biologic effects that may be associated with them.

HISTORICAL BACKGROUND

Within weeks after Röntgen revealed the first X-ray photographs in January 1896, news of the discovery spread throughout the world. Soon afterward, the penetrating properties of the rays began to be exploited for medical purposes, with no inkling that such radiation might have deleterious effects.

Early reports of burns and cancer induced by X rays

The first reports of X-ray injury to human tissue came later in 1896. Elihu Thomson, an American electrical engineer, deliberately exposed one of his fingers to X rays and provided accurate observations on the burns produced. That same year, Thomas Alva Edison was engaged in developing a fluorescent X-ray lamp when he noticed that his assistant, Clarence Dally, was so "poisonously affected" by the new rays that his hair fell out and his scalp became inflamed and ulcerated. By 1904 Dally had developed severe ulcers on both hands and arms, which soon became cancerous and caused his early death.

During the next few decades, many investigators and physicians developed radiation burns and cancer, and more than 100 of them died as a result of their exposure to X rays. These unfortunate early experiences eventually led to an awareness of radiation hazards for professional workers and stimulated the development of a new branch of science—namely, radiobiology.

Radiations from radioactive materials were not immediately recognized as being related to X rays. In 1906 Henri Becquerel, the French physicist who discovered radioactivity, accidentally burned himself by carrying radioactive materials in his pocket. Noting that, Pierre Curie, the co-discoverer of radium, deliberately produced a similar burn on himself. Beginning about 1925, a number of women employed in applying luminescent paint that contained radium to clock and instrument dials became ill with anemia and lesions of the jawbones and mouth; some of them subsequently developed bone cancer.

In 1933 Ernest O. Lawrence and his collaborators completed the first full-scale cyclotron at the University of California at Berkeley. This type of particle accelerator was a copious source of neutrons, which had recently been discovered by Sir James Chadwick in England. Lawrence and his associates exposed laboratory rats to fast neutrons produced with the cyclotron and found that such radiation was about two and a half times more effective in killing power for rats than were X rays.

Considerably more knowledge about the biologic effects of neutrons had been acquired by the time the first nuclear reactor was built in 1942 in Chicago. The nuclear reactor, which has become a prime source of energy for the world, produces an enormous amount of neutrons as well as other forms of radiation. The widespread use of nuclear reactors and the development of high-energy particle accelerators, another prolific source of ionizing radiation, have given rise to health physics. This field of study deals with the hazards of radiation and protection against such hazards. Moreover, since the advent of spaceflight in the late 1950s, certain kinds of radiation from space and their effects on human health have attracted much attention. The protons in the Van Allen radiation belts (two doughnut-shaped zones of high-energy particles trapped in the Earth's magnetic field), the protons and heavier ions ejected in solar flares, and similar particles near the top of the atmosphere are particularly important.

Health physics

UNITS FOR MEASURING IONIZING RADIATION

Ionizing radiation is measured in various units. The oldest unit, the roentgen (R), denotes the amount of radiation that is required to produce 1 electrostatic unit of charge in 1 cubic centimetre of air under standard conditions of pressure, temperature, and humidity. For expressing the dose of radiation absorbed in living tissue, the principal units are the gray (Gy; 1 Gy = 1 joule of radiation energy absorbed per kilogram of tissue) and the rad (1 rad = 100 ergs per gram of tissue = 0.01 Gy). The sievert (Sv) and the rem make it possible to normalize doses of different types of radiation in terms of relative biologic effectiveness (RBE), since particulate radiations tend to cause greater injury for a given absorbed dose than do X rays or gamma rays. The dose equivalent of a given type of radiation (in Sv) is the dose of the radiation in Gy multiplied by a quality factor that is based on the RBE of the radiation. Hence, one sievert, defined loosely, is that amount of radiation roughly equivalent in biologic effectiveness to one gray of gamma rays (1 Sv = 100 rem). Because the sievert and the rem are inconveniently large units for certain applications, the milligray (mGy; 1 mGy = 1/1000 Gy) and millisievert (mSv; 1 mSv = 1/1000 Sv) are often substituted.

For expressing the collective dose to a population, the person-Sv and person-rem are the units used. These units represent the product of the average dose per person times the number of people exposed (*e.g.*, 1 Sv to each of 100 persons = 100 person-Sv = 10,000 person-rem).

The units employed for measuring the amount of radioactivity contained in a given sample of matter are the becquerel (Bq) and the curie (Ci). One becquerel is that quantity of a radioactive element in which there is one atomic disintegration per second; one curie is that quantity in which there are 3.7×10^{10} atomic disintegrations per second (1 Bq = 2.7×10^{-11} Ci). The dose that will accumulate over a given period (say, 50 years) from exposure to a given source of radiation is called the committed dose, or dose commitment.

Table 3: Cosmic-Radiation Exposure

location	mean dose in millisievert (mSv)* per year
Sea level, temperate zone	0.20–0.40
1,500 metres	0.40–0.60
3,000 metres	0.80–1.20
12,000 metres	28
36–600 kilometres	70–150
Interplanetary space	180–250
Van Allen radiation belt (protons)	<15,000
Single solar flare (protons and helium)	<10,000

*Millisievert is a radiation dose-equivalent unit: it corresponds to a dose equivalent in biologic effectiveness to 10 ergs energy of gamma radiation transferred to one gram of tissue.

SOURCES AND LEVELS OF RADIATION IN THE ENVIRONMENT

Natural sources. From the beginning, life has evolved in the presence of natural background ionizing radiation. The principal types and sources of such radiation are: (1) cosmic rays, which impinge on the Earth from outer space (Table 3; Figure 4); (2) terrestrial radiations, which

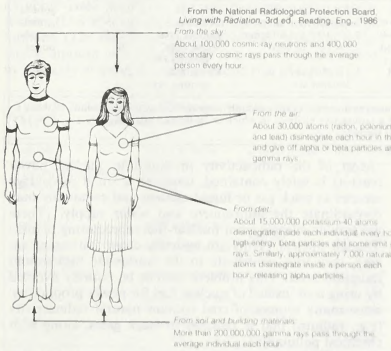


Figure 4: Major sources of natural background radiation and their respective contributions to the exposure of the average person.

Table 4: External Dose Due to Natural Radioactivity in Soil or Rock

source	dose in mSv per year
Ordinary regions	0.25–1.6
Active regions	
Granite in France	1.8–3.5
Houses in Switzerland (alum shale)	1.58–2.2
Monazite alluvial deposits in Brazil	mean 5; max 10
Monazite sands, Kerala, India	3.7–28

are released by the disintegration of radium, thorium, uranium, and other radioactive minerals in the Earth's crust (Table 4; Figure 4); and (3) internal radiations, which are emitted by the disintegration of potassium-40, carbon-14, and other radioactive isotopes that are normally present within living cells (Table 5; Figure 4). The average total dose received from all three sources by a person residing at sea level is approximately 0.91 mSv per year (Table 6); however, a dose twice this size may be received by a person residing at a higher elevation such as Denver, Colo., where cosmic rays are more intense (Table 3), or by a person residing in a geographic region where the radium content of the soil is relatively high (Table 4). In the latter type of region, the radioactive gas radon, which is formed in the decay of radium, may enter a dwelling through its floor or basement walls and accumulate in the indoor air unless the dwelling is well ventilated periodically; occupants of such a dwelling may therefore receive a dose as high as 100 mSv per year in their lungs from inhalation of the entrapped radon and its disintegration products (Table 5; Figure 4).

Average total dose from natural sources

Artificial sources. In addition to natural background radiation, people are exposed to radiation from various man-made sources, the largest of which is the application of X rays in medical diagnosis. Although the doses delivered in different types of X-ray examinations vary from a small fraction of a mGy to tens of mGy (Table 7), the average annual dose per capita from medical and dental irradiation in developed countries of the world now approaches in magnitude the dose received from natural background radiation (Table 6). Less significant artificial sources of radiation include radioactive minerals in crushed rock, building materials, and phosphate fertilizers; radiation-emitting components of television sets, smoke detectors, and various other consumer products; radioactive fallout from nuclear weapons (Table 8); and radiation released in nuclear power production (Table 6).

Use of X rays in medical diagnosis

Table 5: Average Dose Due to Natural Radioactivity Deposited Internally

isotope	radioactivity in millibecquerel (mBq)*	radiation	dose in mSv (per year)	critical organ
Carbon-14	2.2×10^{-7} per kilogram	beta rays	0.016	gonads
Potassium-40	3.9×10^{-7} per kilogram	beta rays	0.165	gonads
Potassium-40	5.6×10^{-8} per kilogram	gamma rays	0.023	gonads
Radium and daughters	3.7×10^{-9} in body	alpha, beta, gamma rays	7.6	bones
Radon and daughters	1.2×10^{-2} per 1 in inhaled air	alpha, beta, gamma rays	20	lungs

*Millibecquerel is a unit of radioactive disintegration rate; it corresponds to that quantity of a radioactive element in which there is one disintegration every 1,000 seconds.

Most of the radioactivity produced in nuclear power reactors is safely contained; however, a small percentage escapes as stack gas or liquid effluent and eventually may contaminate the atmosphere and water supply. (There are similar releases from nuclear-fuel reprocessing plants.) Though nuclear plants are basically clean sources of energy, they thus contribute to the worldwide background radiation level. This problem cannot be entirely avoided by using coal instead of nuclear fuel for power production, since many sources of coal contain natural radioactivity (*e.g.*, radium) that is released in stack gases, along with chemical pollutants.

From Table 6 it is evident that the human population is now exposed to about twice as much radiation from all sources combined as it receives from natural sources alone. Hence, it is important to understand the possible consequences, if any, that may result from the additional exposure to radiation.

In comparison with the relatively small amounts of radiation described above, the dose typically administered to a patient in the treatment of cancer is thousands of times larger; *i.e.*, a total dose of 50 Sv or more is usually delivered to a tumour in daily exposures over a period of four to six weeks. To protect the normal tissues of the patient against injury from such a large dose, as well as to protect medical personnel against excessive occupational exposure to stray radiation, precautions are taken to restrict exposure to the tumour itself insofar as possible. Comparable safeguards are utilized to minimize the exposure of workers employed in other activities involving radiation or radioactive material. Similarly, elaborate safety measures are required for disposal of radioactive wastes from nuclear reactors due in part to the slow rate at which certain fission products decay. A given amount of plutonium-239, for example, still retains about one-half of its radioactivity after 25,000 years, so that reactor wastes containing this long-lived radionuclide must be safely isolated for centuries.

Table 6: Estimates of Average Annual Dose Equivalent to the Whole Body from Various Sources of Irradiation Received by Members of the U.S. Population

source of radiation	average dose rates (mSv/year)
Natural	
Environmental	
Cosmic radiation	0.27 (0.27–1.30)*
Terrestrial radiation	0.28 (0.30–1.15)†
Internal radioactive isotopes	0.36
Subtotal	0.91
Man-made	
Environmental	
Technologically enhanced	0.04
Global fallout	0.04
Nuclear power	0.002
Medical	
Diagnostic	0.78
Radiopharmaceuticals	0.14
Occupational	0.01
Miscellaneous	0.05
Subtotal	1.06
Total	1.97

*Values in parentheses indicate range over which average levels for different states vary with elevation. †Range of variation (shown in parentheses) attributable largely to geographic differences in the content of potassium-40, radium, thorium, and uranium in the Earth's crust.

In the event of an atmospheric nuclear bomb explosion, large quantities of radioactivity are released, the dispersal of which depends on the prevailing weather conditions as well as on the height and nature of the blast. Although the level of contamination resulting from such an explosion or from a nuclear-power plant accident is generally highest in the immediate vicinity of the event itself, both radioactive gas and dust may be transported via air or water for many hundreds of kilometres and eventually contaminate the entire globe.

MECHANISM OF BIOLOGIC ACTION

As ionizing radiation penetrates living matter, it gives up its energy through random interactions with atoms and molecules in its path, leading to the formation of reactive

Table 7: Typical Doses to Exposed Tissue Received in Routine X-Ray Diagnosis

examination	dose per exposure in milligray (mGy)*
X-ray photograph	
Chest	0.4–10
Abdominal	10
Extremities	2.5–10
Fluoroscopy	100–200 per minute
X-ray movies	250 per examination
CAT scan	50–100 per examination

*Milligray is a unit of absorbed radiation dose; it corresponds to 1/1,000 joule of radiation energy absorbed per kilogram of tissue.

Table 8: Worldwide Dose Commitment from Radioactive Fallout from Nuclear Tests Prior to 1970*

source	isotope	half-life	due to bone surfaces (mGy)
External radiation	short lived (*e.g.*, iodine-131)	8 days	360
	longer lived (*e.g.*, cesium-137)	30 years	360
Internal radiation	strontium-89 and -90	50 days	1,310
	cesium-137	28 years	210
	carbon-14†	5,730 years	160
Total			2,400

*North temperate zones; doses calculated for bone surface. †Calculated to year 2000 only.

ions and free radicals. It is the molecular alterations resulting from these ionizations and, in turn, the resultant biochemical changes that give rise to various types of injury. X rays and gamma rays, for example, impart their energy to "planetary" atomic electrons, which are thereby ejected from their orbits. Such an ejection of a planetary electron results in an ion pair consisting of a free electron and the electrically charged atom from which it was ejected. The ejected electron may give rise to a highly reactive free radical, which in turn may diffuse far enough to attack a biologically important target molecule in its vicinity. This so-called indirect action process, through which radiation causes damage via radiation-induced free radicals, may be envisioned as follows:

Process of indirect action

incident X-ray photon
↓
fast electron (e^-)
↓
ion radical
↓
free radical
↓
chemical change due to breakage of bonds
↓
biologic effect.

While the initial steps in the above process occur almost instantaneously, expression of the biologic effect may take years or decades, depending on the type of injury involved. The indirect action of radiation is more important in the biologic effects of low-LET radiations than in those of high-LET radiations (see above), but the latter have a greater capacity to cause injury through direct interaction with biologic targets.

Target theory

Direct biologic actions, studied in detail between 1927 and 1947, gave rise to a target theory of radiobiology that has provided a quantitative treatment of many of the biologic effects of radiation, particularly in the field of genetics. According to this theory, a tissue or cell undergoing irradiation is likened to a field traversed by machine-gun fire, in which the production of a given effect requires one or more hits by an ionized track on a sensitive target. The probability of obtaining the effect is thus dependent on the probability of obtaining the requisite number of hits on the appropriate target or targets.

The distribution of ionizing atomic interactions along the path of an impinging radiation depends on the energy, mass, and charge of the radiation. The ionizations caused by neutrons, protons, and alpha particles are characteristically clustered more closely together than are those caused by X rays or gamma rays (Figure 5). Thus, because the probability of injury depends on the concentration of molecular damage produced at a critical site, or target, in the cell (*e.g.*, a gene or a chromosome), charged particles generally cause greater injury for a given total dose to the cell than do X rays or gamma rays; *i.e.*, they have a high RBE. At the same time, however, charged particles usually penetrate such a short distance in tissue (Figure 5) that they pose relatively little hazard to tissues within the body unless they are emitted by a radionuclide, or radioactive isotope, that has been deposited internally.

RADIONUCLIDES AND RADIOACTIVE FALLOUT

Radionuclides emit various ionizing radiations (*e.g.*, electrons, positrons, alpha particles, gamma rays, or even characteristic X rays), the precise types of which depend on the radionuclide in question. Exposure to a radionuclide and its emissions may be external, in which case the penetrating power of the radiation is an important factor in determining the probability of injury. Alpha particles, for example, do not penetrate deeply enough into the skin to cause damage, whereas energetic beta particles or X rays can be hazardous to the skin and deeper tissues (Figure 5).

From J. Shapiro, *Radiation Protection: A Guide for Scientists and Physicians*, 2nd ed., Harvard University Press, 1981

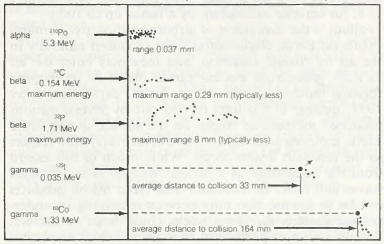


Figure 5: Density of distribution and range of ionization events along tracks of radiations of different types in soft tissue. Each ionization event is denoted by a dot.

Accumulation in critical organs. Radionuclides can enter the body by ingestion, inhalation, or injection. Once taken into the body, their radiation effects depend on their anatomic distribution, duration of retention in the body, and rate of radioactive decay (see above), as well as on the energies of their emitted radiations. An internally deposited radioactive element may concentrate in, and thus irradiate, certain organs more than others. Radioiodine, for example, collects in the thyroid gland, whereas radium and strontium accumulate chiefly in the bones. Different radioelements also vary in their rates of removal. Radioiodine, for instance, is normally eliminated from the thyroid rapidly enough so that its concentration is halved within days. Strontium-90, on the other hand, is retained in high concentrations in the skeleton for years.

The term critical organ refers to the part of the body most vulnerable to a given isotope. The critical organ for plutonium, radium, strontium, and many other fission products is bone and the adjacent bone marrow. For iodine, the critical organ is the thyroid gland. Insoluble airborne radioactive dust often settles in the alveoli of the lungs, while small colloidal particles may become deposited in the bone marrow, liver, or spleen. Table 9 gives an abbreviated list of the maximum permissible concentrations (U.S. recommendations) of some radionuclides for humans. (The maximum permissible concentration is the largest amount of a radionuclide that can be accumulated in the body without producing undue risk of injury.)

Maximum permissible concentration (MPC)

Table 9: Values for the Maximum Permissible Concentration (MPC) of Certain Radionuclides

isotope	chemical form	critical organ	mBq in body
Tritium (hydrogen-3)	water		7.4×10^{-3}
Carbon-14	carbon dioxide		1.5×10^{-5}
Strontium-90*	water-soluble salt		1.5×10^{-6}
		bone	1.5×10^{-7}
Iodine-131	water-soluble salt		1.8×10^{-6}
		thyroid	2.6×10^{-8}
Cesium-137	water-soluble salt		1.1×10^{-6}
Radon-222†	gas		
Radium-226‡	water-soluble salt		7.4×10^{-8}
		bone	3.7×10^{-8}
Uranium	water-soluble salt		7.4×10^{-8}
		kidney	1.8×10^{-10}
Plutonium-239	water-soluble salt		1.5×10^{-8}
		bone	1.5×10^{-9}

*MPC in drinking water: 3.7×10^{-9} micro Bq per litre. †MPC in air: 3.7×10^{-11} micro Bq per litre. ‡MPC in drinking water: 3.7×10^{-10} micro Bq per litre.

Since a radionuclide delivers radiation continuously to the surrounding tissue, the effect of such protracted continuous exposure must be distinguished from that of a single exposure or of periodically repeated exposures. From experiments with divided doses of gamma radiation or X radiation, it has been found that up to about 60 percent of the radiation effect from a single brief exposure is repaired within several hours. The body therefore is able to tolerate a larger total dose when the dose is accumulated slowly or when part of it is absorbed at a later time. There is less recovery with neutron and alpha radiation, however. (Neutrons are generally more effective agents of mutation than are X rays: for a single brief exposure, by a factor 1 to 8; for chronic irradiation, by a factor up to 100.)

Fallout is the deposition of airborne radioactive contaminants on Earth. Radioisotopes are produced naturally in the air by cosmic radiation, and they may enter the air in stack gases from nuclear power plants or be released through industrial accidents or nuclear explosions. After 1954, nuclear bomb tests carried out by several nations produced measurable fallout on the surface of the entire Earth, arousing great concern and controversy with respect to the resultant health effects. While much of the hazard from the detonation of a nuclear weapon is due to blast waves and heat, the radiation dose from fission products can be so intense that only persons remaining in underground shelters for some weeks could hope to survive. Usually the most prominent isotopes in fallout are fission products; however, all materials exposed to nuclear blasts may become radioactive.

Fallout from a nuclear bomb blast

The hazards of long-lived radioisotopes. Several of the radioisotopes contained in fallout are especially hazardous because they remain radioactive for relatively long periods. Cesium-137, strontium-90, and plutonium-239 may be the most significant among these. Fallout material can cover external surfaces and foliage and later be washed into the soil, from which plants may absorb strontium-90, along with the chemically similar calcium, and cesium-137 with potassium. Humans take in these radioactive materials chiefly from drinking water and from plant and animal foods, including milk. Many fallout isotopes that reach the sea and inland waterways eventually end up in concentrated form in the bodies of waterborne animals and plants, becoming a source of concern when they are part of the human food chain.

The most easily detectable fallout product in humans and other animals is iodine-131, an isotope that emits beta and gamma rays and is enriched about 100 times in the thyroid gland through selective accumulation. Because of its relatively short half-life (eight days), iodine-131 is probably not the most hazardous fallout isotope; yet, excessive amounts of radiation from this isotope can lead to metabolic disturbances and an increased incidence of thyroid cancer, especially in children.

A mixture of radioactive gases is discharged into the atmosphere in small amounts by nuclear power reactors. Reactors are thus generally placed at sites where atmospheric mixing and transport are such that the short-lived gases decay and are diluted before they can be inhaled in appreciable amounts by human populations.

Protection against radioactive fallout

Methods that have been developed for biologic protection against fallout range from measures designed to keep radioisotopes out of the body to biochemical means for rapidly eliminating such isotopes from tissues. At times of nuclear emergencies, airborne radioactive particles may be kept from the lungs by staying indoors or by wearing masks with suitable filtration. Absorption of ingested isotopes via the intestinal tract may be inhibited by certain mucoprotein substances that possess great surface affinity for adsorption of strontium and other substances; sodium alginate prepared from seaweed kelp is such a substance. It is possible with appropriate chemicals to remove virtually all radioactive strontium from cow's milk without affecting its essential nutritive components. Certain chelates—for example, EDTA (ethylenediaminetetraacetic acid)—will react with strontium and "cover" this atom. As a result, the presence of EDTA in the blood reduces the deposition of strontium in bones (elimination of already deposited isotopes also is somewhat accelerated). Unfortunately, however, EDTA and most other chelating agents are not specific for strontium; they also chelate the closely related and important element calcium. Consequently, their use requires expert medical supervision and is limited in effectiveness. On the other hand, the uptake of radioactive iodine by the thyroid gland may be reduced by the ingestion of large amounts of stable iodine, which is relatively nontoxic except to those with special sensitivity.

MAJOR TYPES OF RADIATION INJURY

Any living organism can be killed by radiation if exposed to a large enough dose, but the lethal dose varies greatly from species to species. Mammals can be killed by less than 10 Gy, but fruit flies may survive 1,000 Gy. Many bacteria and viruses may survive even higher doses. In general, humans are among the most radiosensitive of all living organisms, but the effects of a given dose in a person depend on the organ irradiated, the dose, and the conditions of exposure.

Somatic and genetic effects

The biologic effects of radiation in humans and other mammals are generally subdivided into (1) those that affect the body of the exposed individual—somatic effects—and (2) those that affect the offspring of the exposed individual—genetic, or heritable, effects. Among the somatic effects, there are those that occur within a short period of time (*e.g.,* inhibition of cell division) and those that may not occur until years or decades after irradiation (*e.g.,* radiation-induced cancer). In addition, there are those, called non-stochastic effects, that occur only in response to a considerable dose of radiation (*e.g.,* ulceration of the skin) and those, termed stochastic, for which no threshold dose is known to exist (*e.g.,* radiation-induced cancer).

Every type of biologic effect of radiation, irrespective of its precise nature, results from injury to the cell, the microscopic building block of which all living organisms are composed. It therefore seems useful to open a review of such effects with a discussion of the action of radiation on the cell.

Effects on the cell. The effects of radiation on the cell include interference with cell division, damage to chromosomes, damage to genes (mutations), neoplastic transformation (a change analogous to the induction of cancer), and cell death. The mechanisms through which these changes are produced are not yet fully understood, but each change is thought to be the end result of chemical alterations that are initiated by radiation as it randomly traverses the cell.

Any type of molecule in the cell can be altered by irradiation, but the DNA of the genetic material is thought to be the cell's most critical target, since damage to a single gene may be sufficient to kill or profoundly alter the cell. A dose that can kill the average dividing cell (say, 1–2 Sv) produces dozens of lesions in the cell's DNA molecules. Although most such lesions are normally reparable through the action of intracellular DNA repair processes, those that remain unrepaired or are misrepaired may give rise to permanent changes in the affected genes (*i.e.,* mutations) or in the chromosomes on which the genes are carried, as discussed below.

Variations in radiosensitivity among cells

In general, dividing cells (such as cancer cells) are more radiosensitive than nondividing cells. As noted above, a dose of 1–2 Sv is sufficient to kill the average dividing cell, whereas nondividing cells can usually withstand many times as much radiation without overt signs of injury. It is when cells attempt to divide for the first time after irradiation that they are most apt to die as a result of radiation injury to their genes or chromosomes.

The percentage of human cells retaining the ability to multiply generally decreases exponentially with increasing radiation dose, depending on the type of cell exposed and the conditions of irradiation. With X rays and gamma rays, traversal by two or more radiation tracks in swift succession are usually required to kill the cell. Hence, the survival curve is typically shallower at low doses and low dose rates than at high doses and high dose rates (Figure 6). The reduced killing effectiveness of a given dose when it is delivered in two or more widely spaced fractions is attributed to the repair of sublethal damage between successive exposures. With densely ionizing par-

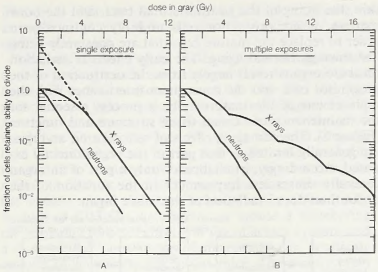


Figure 6: *Typical curves showing relation of proliferative ability to radiation dose for mammalian cells exposed to X rays and fast neutrons.*
(A) Dose delivered in a single exposure. In the case of X rays, the dose-effect curve has a large initial shoulder; for fast neutrons, the initial shoulder is smaller and the final slope is steeper. (B) Multiple successive exposures. The shoulder of the X-ray survival curve is reexpressed after each dose fraction.
Adapted from E.J. Hall, *Radiology for the Radiologist*, 2nd ed., Harper and Row, 1978

ticulate radiations, on the other hand, the survival curve is characteristically steeper than with X rays or gamma rays, and its slope is relatively unaffected by the dose or the dose rate (Figure 6), implying that the death of the cell usually results from a single densely ionizing particle track and that the injury produced by such a track is of a relatively irreparable type.

Damage to genes (mutations). Gene mutations resulting from radiation-induced damage to DNA have been produced experimentally in many types of organisms. In general, the frequency of a given mutation increases in proportion to the dose of radiation in the low-to-intermediate dose range. At higher doses, however, the frequency of mutations induced by a given dose may be dependent on the rate at which the dose is accumulated, tending to be lower if the dose is accumulated over a long period of time.

In human white blood cells (lymphocytes), as in mouse spermatogonia and oocytes, the frequency of radiation-induced mutations approximates 1 mutation per 100,000 cells per genetic locus per Sv. This rate of increase is not large enough to detect with existing methodology in the children of the atomic-bomb survivors of Hiroshima and Nagasaki, owing to their limited numbers and the comparatively small average dose of radiation received by their parents. Accordingly, it is not surprising that heritable effects of irradiation have not been observable thus far in this population or in any other irradiated human population, in spite of exhaustive efforts to detect them.

The observed proportionality between the frequency of induced mutations and the radiation dose has important health implications for the human population, since it implies that even a small dose of radiation given to a large number of individuals may introduce mutant genes into the population, provided that the individuals are below reproductive age at the time of irradiation. The effect on a population of a rise in its mutation rate depends, however, on the role played by mutation in determining the characteristics of the population. Although deleterious genes enter the population through mutations, they tend to be eliminated because they reduce the fitness of their carriers. Thus, a genetic equilibrium is reached at the point where the entry of deleterious genes into the population through mutation is counterbalanced by their loss through reduction in fitness. At the point of equilibrium, an increase of the mutation rate by a given percentage causes a proportionate increase in the gene-handicapped fraction in the population. The full increase is not manifested immediately, however, but only when genetic equilibrium is again established, which requires several generations.

The mutation-rate doubling dose

The capacity of radiation to increase the frequency of mutations is often expressed in terms of the mutation-rate doubling dose, which is the dose that induces as large an additional rate of mutations as that which occurs spontaneously in each generation. The more sensitive the genes are to radiation, the lower is the doubling dose. The doubling dose for high-intensity exposure in several different organisms has been found experimentally to lie between about 0.3 and 1.5 Gy. For seven specific genes in the mouse, the doubling dose of gamma radiation for spermatogonia is about 0.3 Gy for high-intensity exposure and about 1.0 Gy for low-intensity exposure. Little is known about the doubling dose for human genes, but most geneticists assume that it is about the same as the doubling dose for those of mice. Studies of the children of atomic-bomb survivors are consistent with this view, as noted above.

From the results of experiments with mice and other laboratory animals, the dose required to double the human mutation rate is estimated to lie in the range of 0.2–2.5 Sv, implying that less than 1 percent of all genetically related diseases in the human population is attributable to natural background irradiation (Table 10). Although natural background irradiation therefore appears to make only a relatively small contribution to the overall burden of genetic illness in the world's population, millions of individuals may be thus affected in each generation.

Table 10: Estimated Contribution of Natural Background Irradiation to the Occurrence of Genetic Abnormalities in the Human Population

	incidence per million offspring		
	number attributable to all causes	number attributable to natural background irradiation*	
type of genetic abnormality		first generation	equilibrium generations
Dominant traits + diseases	10,000	3–15	100–300
Chromosomal + recessive traits + diseases	10,000	<30	<120
Recognized abortions			
Aneuploidy + polyploidy	35,000	33	33
XO	9,000	9	9
Unbalanced rearrangements	11,000	216	276
Congenital anomalies	20,000		
Anomalies expressed after birth	10,000	150	30–300
Constitutional + degenerative diseases	15,000		
Total (rounded)	120,000	<300–500	<600–900

*Equivalent to 1 mSv per year, the average dose from background radiation, or 30 mSv per generation.

Notwithstanding the fact that the vast majority of mutations are decidedly harmful, those induced by irradiation in seeds are of interest to horticulturists as a means of producing new and improved varieties of plants. Mutations produced in this manner can affect such properties of the plant as early ripening and resistance to disease, with the result that economically important varieties of a number of species have been produced by irradiation. In their effects on plants, fast neutrons and heavy particles have been found to be up to about 100 times more mutagenic than X rays. Radioactive elements taken up by plants also can be strongly mutagenic. In the choice of a suitable dose for the production of mutations, a compromise has to be made between the mutagenic effects and damaging effects of the radiation. As the number of mutations increases, so also does the extent of damage to the plants. In the irradiation of dry seeds by X rays, a dose of 10 to 20 Gy is usually given.

Damage to chromosomes. By breaking both strands of the DNA molecule, radiation also can break the chromosome fibre and interfere with the normal segregation of duplicate sets of chromosomes to daughter cells at the time of cell division, thereby altering the structure and number of chromosomes in the cell. Chromosomal changes of this kind may cause the affected cell to die when it attempts to divide, or they may alter its properties in various other ways.

Chromosome breaks

Chromosome breaks often heal spontaneously, but a break that fails to heal may cause the loss of an essential part of the gene complement; this loss of genetic material is called gene deletion. A germ cell thus affected may be

capable of taking part in the fertilization process, but the resulting zygote may be incapable of full development and may therefore die in an embryonic state.

When adjoining chromosome fibres in the same nucleus are broken, the broken ends may join together in such a way that the sequence of genes on the chromosomes is changed. For example, one of the broken ends of chromosome A may join onto a broken end of chromosome B, and vice versa in a process termed translocation. A germ cell carrying such a chromosome structural change may be capable of producing a zygote that can develop into an adult individual, but the germ cells produced by the resulting individual may include many that lack the normal chromosome complement and so yield zygotes that are incapable of full development; an individual affected in this way is termed semisterile. Because the number of his descendants is correspondingly lower than normal, such chromosome structural changes tend to die out in successive generations.

As would be expected from target theory considerations, X rays and gamma rays given at high doses and high dose rates induce more two-break aberrations per unit dose than are produced at low doses and low dose rates (Figure 7). With densely ionizing radiation, by comparison, the yield of two-break aberrations for a given dose is higher than with sparsely ionizing radiation and is proportional to the dose irrespective of the dose rate (Figure 7). From these comparative dose-response relationships, it is inferred that a single X-ray track rarely deposits enough energy at any one point to break two adjoining chromosomes simultaneously, whereas the two-break aberrations that are induced by high-LET irradiation result preponderantly from single particle tracks.

From D.C. Lloyd and R.J. Purrott. "Chromosome Aberration Analysis in Radiological Protection Dosimetry." *Radiological Protection Dosimetry 1*, 1981

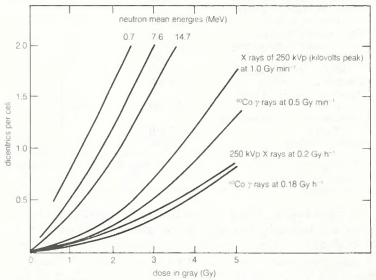


Figure 7: Frequency of dicentric chromosome aberrations in human lymphocytes irradiated in vitro, as a function of the dose and dose rate of neutrons, X rays, and gamma rays.

In irradiated human lymphocytes, the frequency of chromosome aberrations varies so predictably with the dose of radiation (Figure 7) that it is used as a crude biologic dosimeter of exposure in radiation workers and other exposed persons. What effect, if any, an increase in the frequency of chromosome aberrations may have on the health of an affected individual is uncertain. Only a small percentage of all chromosome aberrations is attributable to natural background radiation; the majority result from other causes, including certain viruses, chemicals, and drugs.

Effects on organs of the body (somatic effects). A wide variety of reactions occur in response to irradiation in the different organs and tissues of the body. Some of the reactions occur quickly, while others occur slowly. The killing of cells in affected tissues, for example, may be detectable within minutes after exposure, whereas degenerative changes such as scarring and tissue breakdown may not appear until months or years afterward.

In general, dividing cells are more radiosensitive than nondividing cells (see above), with the result that radiation injury tends to appear soonest in those organs and tissues in which cells proliferate rapidly. Such tissues include the skin, the lining of the gastrointestinal tract, and the bone marrow, where progenitor cells multiply continually in order to replace the mature cells that are constantly being lost through normal aging. The early effects of radiation on these organs result largely from the destruction of the progenitor cells and the consequent interference with the replacement of the mature cells, a process essential for the maintenance of normal tissue structure and function (Figure 8). The damaging effects of radiation on an organ are generally limited to that part of the organ directly exposed. Accordingly, irradiation of only a part of an organ generally causes less impairment in the function of the organ than does irradiation of the whole organ.

From A.C. Upton. "Cancer Induction and Non-Stochastic Effects." *British Journal of Radiology*, vol. 60, 1987

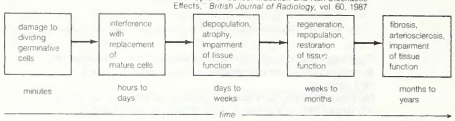


Figure 8: Characteristic sequence of events in the pathogenesis of early and late non-stochastic effect of radiation.

Skin. Radiation can cause various types of injury to the skin, depending on the dose and conditions of exposure. The earliest outward reaction of the skin is transitory reddening (erythema) of the exposed area, which may appear within hours after a dose of 6 Gy or more. This reaction typically lasts only a few hours and is followed two to four weeks later by one or more waves of deeper and more prolonged reddening in the same area. A larger dose may cause subsequent blistering and ulceration of the skin and loss of hair, followed by abnormal pigmentation months or years later.

Bone marrow. The blood-forming cells of the bone marrow are among the most radiosensitive cells in the body. If a large percentage of such cells are killed, as can happen when intensive irradiation of the whole body occurs, the normal replacement of circulating blood cells is impaired. As a result, the blood cell count may become depressed and, ultimately, infection, hemorrhage, or both may ensue. A dose below 0.5–1 Sv generally causes only a mild, transitory depletion of blood-forming cells; however, a dose above 8 Sv delivered rapidly to the whole body usually causes a fatal depression of blood-cell formation.

High radiosensitivity of blood-forming cells of the bone marrow

Gastrointestinal tract. The response of the gastrointestinal tract is comparable in many respects to that of the skin. Proliferating cells in the mucous membrane that lines the tract are easily killed by irradiation, resulting in the denudation and ulceration of the mucous membrane. If a substantial portion of the small intestine is exposed rapidly to a dose in excess of 10 Gy, as may occur in a radiation accident, a fatal dysentery-like reaction results within a very short period of time.

Reproductive organs. Although mature spermatozoa are relatively resistant to radiation, immature sperm-forming cells (spermatogonia) are among the most radiosensitive cells in the body. Hence, rapid exposure of both testes to a dose as low as 0.15 Sv may interrupt sperm-production temporarily, and a dose in excess of 4 Sv may be sufficient to cause permanent sterility in a certain percentage of men.

In the human ovary, oocytes of intermediate maturity are more radiosensitive than those of greater or lesser maturity. A dose of 1.5–2.0 Sv delivered rapidly to both ovaries may thus cause only temporary sterility, whereas a dose exceeding 2–3 Sv is likely to cause permanent sterility in an appreciable percentage of women.

Lens of the eye. Irradiation can cause opacification of the lens, the severity of which increases with the dose. The effect may not become evident, however, until many months after exposure. During the 1940s, some physicists who worked with the early cyclotrons developed cataracts as a result of occupational neutron irradiation, indicating for the first time the high relative biologic effectiveness of neutrons for causing lens damage. The threshold for a progressive, vision-impairing opacity, or cataract, varies

Cataract development

from 5 Sv delivered to the lens in a single exposure to as much as 14 Sv delivered in multiple exposures over a period of months.

Brain and sensory organs. Generally speaking, humans do not sense a moderate radiation field; however, small doses of radiation (less than 0.01 Gy) can produce phosphene, a light sensation on the dark-adapted retina. American astronauts on the first spacecraft that landed on the Moon (Apollo 11, July 20, 1969) observed irregular light flashes and streaks during their flight, which probably resulted from single heavy cosmic-ray particles striking the retina. In various food-preference tests, rats, when given the choice, avoid radiation fields of even a few mGy. A dose of 0.03 Gy is sufficient to arouse a slumbering rat, probably through effects on the olfactory system, and a dose of the same order of magnitude can accelerate seizures in genetically susceptible mice. The mature brain and nervous system are relatively resistant to radiation injury, but the developing brain is radiosensitive to damage (see below).

Symptoms of the acute radiation syndrome

Radiation sickness. The signs and symptoms resulting from intensive irradiation of a large portion of the bone marrow or gastrointestinal tract constitute a clinical picture known as radiation sickness, or the acute radiation syndrome. Early manifestations of this condition typically include loss of appetite, nausea, and vomiting within the first few hours after irradiation, followed by a symptom-free interval that lasts until the main phase of the illness (Table 11).

Table 11: Symptoms of Acute Radiation Sickness (Hematopoietic Form)

time after exposure	supralethal dose range (6–10 Gy)	midlethal dose range (2.5–5 Gy)	sublethal dose range (1–2 Gy)
Several hours	no definite symptoms	nausea and vomiting	
First week	diarrhea, vomiting, inflammation of throat	no definite symptoms	
Second week	fever, rapid emaciation leading to death for 100 percent of the population		
Third week		loss of hair begins loss of appetite general malaise fever, hemorrhages, pallor leading to rapid emaciation and death for 50 percent of the population	loss of appetite sore throat pallor and diarrhea recovery begins (no deaths in absence of complications)

The main phase of the intestinal form of the illness typically begins two to three days after irradiation, with abdominal pain, fever, and diarrhea, which progress rapidly in severity and lead within several days to dehydration, prostration, and a fatal, shocklike state. The main phase of the hematopoietic form of the illness characteristically begins in the second or third week after irradiation, with fever, weakness, infection, and hemorrhage. If damage to the bone marrow is severe, death from overwhelming infection or hemorrhage may ensue four to six weeks after exposure unless corrected by transplantation of compatible unirradiated bone marrow cells.

The higher the dose received, the sooner and more profound are the radiation effects. Following a single dose of more than 5 Gy to the whole body, survival is improbable (Table 11). A dose of 50 Gy or more to the head may cause immediate and discernible effects on the central nervous system, followed by intermittent stupor and incoherence alternating with hyperexcitability, epileptiform seizures, and death within several days (the cerebral form of the acute radiation syndrome).

When the dose to the whole body is between 6 and 10 Gy, the earliest symptoms are loss of appetite, nausea, and vomiting, followed by prostration, watery and bloody diarrhea, abhorrence of food, and fever (Table 11). The blood-forming tissues are profoundly injured, and the white blood cell count may decrease within 15–30 days from about 8,000 per cubic millimetre to as low as 200. As a result of these effects, the body loses its defenses against microbial infection, and the mucous membranes lining the gastrointestinal tract may become inflamed. Furthermore, internal or external bleeding may occur because of a reduction in blood platelets. Return of the early symptoms, frequently accompanied by delirium or coma, presage death; however, symptoms may vary significantly from individual to individual. Complete loss of hair within 10 days has been taken as an indication of a lethally severe exposure.

In the dose range of 1.5–5.0 Gy, survival is possible (though in the upper range improbable), and the symptoms appear as described above but in milder form and generally following some delay. Nausea, vomiting, and malaise may begin on the first day and then disappear, and a latent period of relative well-being follows. Anemia and leukopenia set in gradually. After three weeks, internal hemorrhages may occur in almost any part of the body, but particularly in mucous membranes. Susceptibility to infection remains high, and some loss of hair occurs. Lassitude, emaciation, and fever may persist for many weeks before recovery or death occurs.

Moderate doses of radiation can severely depress the immunologic defense mechanisms, resulting in enhanced sensitivity to bacterial toxins, greatly decreased fixation of antigens, and reduced efficiency of antibody formation. Antibiotics, unfortunately, are of limited effectiveness in combating postirradiation infections. Hence, of considerable value are plastic isolators that allow antiseptic isolation of a person from his environment; they provide protection against infection from external sources during the period critical for recovery.

Below a dose of 1.5 Gy, an irradiated person is generally able to survive intensive whole-body irradiation. The symptoms following exposure in this dose range are similar to those already described but milder and delayed. With a dose under 1 Gy, the symptoms may be so mild that the exposed person is able to continue his normal occupation in spite of measurable depression of his bone marrow. Some persons, however, suffer subjective discomfort from doses as low as 0.3 Gy. Although such doses may cause no immediate reactions, they may produce delayed effects that appear years later (see below).

Effects on the growth and development of the embryo. The tissues of the embryo, like others composed of rapidly proliferating cells, are highly radiosensitive. The types and frequencies of radiation effects, however, depend heavily on the stage of development of the embryo or fetus at the time it is exposed. For example, when exposure occurs while an organ is forming, malformation of the organ may result. Exposure earlier in embryonic life is more likely to kill the embryo than cause a congenital malformation, whereas exposure at a later stage is more likely to produce a functional abnormality in the offspring than a lethal effect or a malformation.

Radiation-induced malformations

A wide variety of radiation-induced malformations have been observed in experimentally irradiated rodents. Many of these are malformations of the nervous system, including microcephaly (reduced size of brain), exencephaly (part of the brain formed outside the skull), hydrocephalus (enlargement of the head due to excessive fluid), and anophthalmia (failure of the eyes to develop). Such effects may follow a dose of 1–2 Gy given at an appropriate stage of development. Functional abnormalities produced in laboratory animals by prenatal irradiation include abnormal reflexes, restlessness, and hyperactivity, impaired learning ability, and susceptibility to externally induced seizures. The abnormalities induced by radiation are similar to those that can be caused by certain virus infections, neurotropic drugs, pesticides, and mutagens.

Abnormalities of the nervous system, which occur in 1–2 percent of human infants, were found with greater frequency among children born to women who were pregnant and residing in Hiroshima or Nagasaki at the time of the atomic explosions. The incidence of reduced head size and mental retardation in such children was increased by about 40 percent per Gy when exposure occurred between the eighth and 15th week of gestation, the age of greatest susceptibility to radiation.

The period of maximal sensitivity for each developing

organ is sharply circumscribed in time, with the result that the risk of malformation in a particular organ depends heavily on the precise stage of development at which the embryo is irradiated. The risk that a given dose will produce a particular malformation is thus much smaller if the dose is spread out over many days or weeks than if it is received during the few hours of the critical period itself. It also would appear that the induction of a malformation generally requires injury to many cells in a developing organ, so that there is little likelihood of such an effect resulting from the low doses and dose rates characteristic of natural background radiation.

Effects on the incidence of cancer. Atomic-bomb survivors, certain groups of patients exposed to radiation for medical purposes, and some groups of radiation workers have shown dose-dependent increases in the incidence of certain types of cancer. The induced cancers have not appeared until years after exposure, however, and they have shown no distinguishing features by which they can be identified individually as having resulted from radiation, as opposed to some other cause. With few exceptions, moreover, the incidence of cancer has not been increased detectably by doses of less than 0.01 Sv.

Because the carcinogenic effects of radiation have not been documented over a wide enough range of doses and dose rates to define the shape of the dose-incidence curve precisely, the risk of radiation-induced cancer at low levels of exposure can be estimated only by extrapolation from observations at higher dose levels, based on assumptions about the relation between cancer incidence and dose. For most types of cancer, information about the dose-incidence relationship is rather meagre. The most extensive data available are for leukemia and cancer of the female breast.

Radiation as a factor in leukemia and breast cancer

The overall incidence of all forms of leukemia other than the chronic lymphatic type has been observed to increase roughly in proportion to dose during the first 25 years after irradiation. Different types of leukemia, however, vary in the magnitude of the radiation-induced increase for a given dose, the age at which irradiation occurs, and the time after exposure. The total excess of all types besides chronic lymphatic leukemia, averaged over all ages, amounts to approximately one to three additional cases of leukemia per year per 10,000 persons at risk per sievert to the bone marrow.

Cancer of the female breast also appears to increase in incidence in proportion to the radiation dose. Furthermore, the magnitude of the increase for a given dose appears to be essentially the same in women whose breasts were irradiated in a single, brief exposure (*e.g.*, atomic-bomb survivors), as in those who were irradiated over a period of years (*e.g.*, patients subjected to multiple fluoroscopic examinations of the chest or workers assigned to coating watch and clock dials with paint containing radium), implying that even small exposures widely separated in time exert carcinogenic effects on the breast that are fully additive and cumulative. Although susceptibility decreases sharply with age at the time of irradiation, the excess of breast cancer averaged over all ages amounts to three to six cases per 10,000 women per sievert each year.

Additional evidence that carcinogenic effects can be produced by a relatively small dose of radiation is provided by the increase in the incidence of thyroid tumours that has been observed to result from a dose of 0.06–2.0 Gy of X rays delivered to the thyroid gland during infancy or childhood, and by the association between prenatal diagnostic X irradiation and childhood leukemia. The latter association implies that exposure to as little as 10–50 mGy of X radiation during intrauterine development may increase the subsequent risk of leukemia in the exposed child by as much as 40–50 percent.

Although some, but not all, other types of cancer have been observed to occur with greater frequency in irradiated populations (Table 12), the data do not suffice to indicate whether the risks extend to low doses. It is apparent, however, that the dose-incidence relationship varies from one type of cancer to another. From the existing evidence, the overall excess of all types of cancer combined may be inferred to approximate 0.6–1.8 cases per 1,000 persons per sievert per year when the whole body is exposed to radiation, beginning two to 10 years after irradiation. This increase corresponds to a cumulative lifetime excess of roughly 20–100 additional cases of cancer per 1,000 persons per sievert, or to an 8–40 percent per sievert increase in the natural lifetime risk of cancer.

Table 12: Estimated Lifetime Cancer Risks Attributed to Low-Level Irradiation

site irradiated	cancers per 10,000 person-Sv*
Bone marrow (leukemia)	15–20
Thyroid	25–120
Breast (women only)	40–200
Lung	25–140
Stomach	
Liver	5–60 (each)
Colon	
Bone	
Esophagus	
Small intestine	
Urinary bladder	5–30 (each)
Pancreas	
Lymphatic tissue	
Skin	10–20
Total (both sexes)	125–1,000

*The unit person-Sv represents the product of the average dose per person times the number of people exposed (1 sievert to each of 10,000 persons = 10,000 person-Sv); all values provided here are rounded.
Source: National Academy of Sciences Advisory Committee on the Biological Effects of Ionizing Radiation, *The Effects on Populations of Exposure to Low Levels of Ionizing Radiation* (1972, 1980); United Nations Scientific Committee on the Effects of Atomic Radiation, *Sources and Effects of Ionizing Radiation* (1977 report to the General Assembly, with annexes).

The above-cited risk estimates imply that no more than 1–3 percent of all cancers in the general population result from natural background ionizing radiation. At the same time, however, the data suggest that up to 20 percent of lung cancers in nonsmokers may be attributable to inhalation of radon and other naturally occurring radionuclides present in air.

Shortening of the life span. Laboratory animals whose entire bodies are exposed to radiation in the first half of life suffer a reduction in longevity that increases in magnitude with increasing dose. This effect was mistakenly interpreted by early investigators as a manifestation of accelerated or premature aging. The shortening of life in irradiated animals, however, has since been observed to be attributable largely, if not entirely, to the induction of benign and malignant growths. In keeping with this observation is the finding that mortality from diseases other than cancer has not been increased detectably by irradiation among atomic-bomb survivors.

PROTECTION AGAINST EXTERNAL RADIATION

A growing number of substances have been found to provide some protection against radiation injury when administered prior to irradiation (Table 13). Many of

Table 13: Some Chemicals That Exert Radioprotective Effects in Laboratory Animals

class	specific chemical	effective dose (in milligrams per kilogram of tissue)
Sulfur compounds	glutathione	1,000
	cysteine	1,000
	cysteamine	150
	AET*	350
Hormones	estradiolbenzoate	12
	ACTH	25 for 7 days
Enzyme inhibitors	sodium cyanide	5
	carbon monoxide	by inhalation
	mercaptoethylamine (MEA)	235
	para-aminopropiophenone (PAPP)	30
Metabolites	formic acid	90
Vasoconstrictors	serotonin	50
Nervous system drugs	amphetamine	1
	chlorpromazine	20

*Aminoethylisothiuronium bromide hydrobromide.

them apparently act by producing anoxia or by competing for oxygen with normal cell constituents and radiation-produced radicals. All of the protective compounds tried thus far, however, are toxic, and anoxia itself is hazardous. As a consequence, their administration to humans is not yet practical.

Diurnal changes in the radiosensitivity of rodents indicate that the factors responsible for daily biologic rhythms may also alter the responses of tissues to radiation. Such factors include the hormone thyroxine, a normal secretion of the thyroid gland. Other sensitizers at the cellular level include nucleic-acid analogues (*e.g.*, 5-fluorouracil) as well as certain compounds that selectively radiosensitize hypoxic cells such as metronidazole.

Radiosensitivity is also under genetic control to some degree, susceptibility varying among different inbred mouse strains and increasing in the presence of inherited deficiencies in capacity for repairing radiation-induced damage to DNA. Germ-free mice, which spend their entire lives in a sterile environment, also exhibit greater resistance to radiation than do animals in a normal microbial environment owing to elimination of the risk of infection.

For many years it was thought that radiation disease was irreversible once a lethal dose had been received. It has since been found that bone-marrow cells administered soon after irradiation may enable an individual to survive an otherwise lethal dose of X rays, because these cells migrate to the marrow of the irradiated recipient, where they proliferate and repopulate the blood-forming tissues. Under these conditions bone-marrow transplantation is feasible even between histo-incompatible individuals, because the irradiated recipient has lost the ability to develop antibodies against the injected "foreign" cells. After a period of some months, however, the transplanted tissue may eventually be rejected, or it may develop an immune reaction against the irradiated host, which also can be fatal. The transplantation of bone-marrow cells has been helpful in preventing radiation deaths among the victims of reactor accidents, as, for example, those injured in 1986 at the Chernobyl nuclear power plant in Ukraine, then in the Soviet Union. It should be noted, however, that cultured or stored marrow cells cannot yet be used for this purpose.

Bone-marrow transplantation as a means of preventing radiation deaths

CONTROL OF RADIATION RISKS

In view of the fact that radiation is now assumed to play a role in mutagenic or carcinogenic activity, any procedure involving radiation exposure is considered to entail some degree of risk. At the same time, however, the radiation-induced risks associated with many activities are negligibly small in comparison with other risks commonly encountered in daily life. Nevertheless, such risks are not necessarily acceptable if they can be easily avoided or if no measurable benefit is to be gained from the activities with which they are associated. Consequently, systematic efforts are made to avoid unnecessary exposure to ionizing radiation in medicine, science, and industry. Toward this end, limits have been placed on the amounts of radioactivity (Tables 9 and 12) and on the radiation doses (Table 14) that the different tissues of the body are permitted to accumulate in radiation workers or members of the public at large.

Although most activities involving exposure to radiation for medical purposes are highly beneficial, the benefits cannot be assumed to outweigh the risks in situations where radiation is used to screen large segments of the population for the purpose of detecting an occasional person with an asymptomatic disease. Examples of such applications include the "annual" chest X-ray examination and routine mammography. Each use of radiation in medicine (and dentistry) is now evaluated for its merits on a case-by-case basis.

Other activities involving radiation also are assessed with care in order to assure that unnecessary exposure is avoided and that their presumed benefits outweigh their calculated risks. In operating nuclear power plants, for example, much care is taken to minimize the risk to surrounding populations. Because of such precautions, the total impact on health of generating a given amount of electricity from nuclear power is usually estimated to be smaller than that resulting from the use of coal for the same purpose, even after allowances for severe reactor accidents such as the one at Chernobyl. (C.A.T./A.C.U.)

Table 14: Recommended Exposure Limits*

type of exposure	maximum permissible dose
Occupational exposures (annual)†	
Effective dose-equivalent limit (stochastic effects)	50 mSv
Dose-equivalent limits for tissues and organs (non-stochastic effects)	
Lens of eye	150 mSv
All others (*e.g.*, red bone marrow, breast, lung, gonads, skin, and extremities)	500 mSv
Guidance: cumulative exposure	10 mSv × age
Planned special occupational exposure, effective dose-equivalent limit‡	100 mSv
Guidance for emergency occupational exposure	§
Public exposures (annual)	
Effective dose-equivalent limit, continuous or frequent exposure†	1 mSv
Effective dose-equivalent limit, infrequent exposure†	5 mSv
Remedial action recommended when:	
Effective-dose equivalent	>5 mSv
Exposure to radon and its decay products	>0.007 Jhm^{-3} ‖
Dose-equivalent limit for lens of eye, skin, and extremities	
Education and training exposures (annual)†	
Effective dose-equivalent limit	1 mSv
Dose-equivalent limit for lens of eye, skin, and extremities	50 mSv
Embryo-fetus exposures†	
Total dose-equivalent limit	5 mSv
Dose-equivalent limit in a month	0.01 mSv
Negligible individual risk level (annual)†	
Effective dose equivalent per source of practice	0.5 mSv

*Including background but excluding internal exposures. †Sum of external and internal exposures, excluding medical exposures. ‡Effective dose equivalent in any one planned event or cumulative effective dose equivalent in planned special exposures over a working lifetime should not exceed 100 mSv (10 rem). §Acute exposures in excess of 100 mSv (10 rem) are justified only by life-saving actions in emergency situations. ‖ Joule hours per cubic metre.
Source: National Council on Radiation Protection and Measurements, *Ionizing Radiation Exposure of the Population of the United States* (1987).

Biologic effects of non-ionizing radiation

EFFECTS OF HERTZIAN WAVES AND INFRARED RAYS

Hertzian waves. The effects of Hertzian waves (electromagnetic waves in the radar and radio range) and of infrared rays usually are regarded as equivalent to the effect produced by heating. The longer radio waves induce chiefly thermal agitation of molecules and excitation of molecular rotations, while infrared rays excite vibrational modes of large molecules and release fluorescent emission as well as heat. Both of these types of radiation are preferentially absorbed by fats containing unsaturated carbon chains.

Effect of long radio waves

The fact that heat production resulted from bombardment of tissue with high-frequency alternating current (wavelengths somewhat longer than the longest radio waves) was discovered in 1891, and the possibility of its utilization for medical purposes was realized in 1909, under the term diathermy. This method of internal heating is beneficial for relieving muscle soreness and sprain (see also below). Diathermy can be harmful, however, if so much internal heat is given that the normal cells of the body suffer irreversible damage. Since humans have heat receptors primarily in their skin, they cannot be forewarned by pain when they receive a deep burn from diathermy. Sensitive regions easily damaged by diathermy are those having reduced blood circulation. Cataracts of the eye lens have been produced in animals by microwave radiation applied in sufficient intensity to cause thermal denaturation of the lens protein.

Microwaves and their application

Microwave ovens have found widespread use in commercial kitchens and private homes. These can heat and cook very rapidly and, if used properly, constitute no hazard to operators. In the radio-television industry and in the radar

division of the military, persons are sometimes exposed to high densities of microwave radiation. The hazard is particularly pronounced with exposure to masers, capable of generating very high intensities of microwaves (*e.g.*, carbon dioxide masers). The biologic effects depend on the absorbency of tissues. At frequencies higher than 150 megahertz, significant absorption takes place. The lens of the human eye is most susceptible to frequencies around 3,000 megahertz, which can produce cataracts. At still higher frequencies, microwaves interact with superficial tissues and skin, in much the same manner as infrared rays.

Acute effects of microwaves become significant if a considerable temperature rise occurs. Cells and tissues eventually die at temperatures of about 43° C. Microwave heating is minimized if the heat that results from energy absorption is dissipated by radiation, evaporation, and heat conduction. Normally one-hundredth of a watt (10 milliwatts) can be so dissipated, and this power limit generally has been set as the permissible dose. Studies with animals have indicated that, below the permissible levels, there are negligible effects to various organ systems. Microwaves or heat applied to testes tend strongly to decrease the viability of sperm. This effect, however, is not significant at the "safe" levels.

In the late 1980s, some investigators in the Soviet Union documented a variety of nonthermal effects of microwaves and recommended about 1,000 times lower safe occupational dose levels than are still in force in the United States today. Most prominent among the nonthermal effects appear to be those on the nervous system. Such effects have resulted in untimely tiring, excitability, and insomnia registered by persons handling high-frequency radio equipment. Nonthermal effects have been observed on the electroencephalogram of rabbits. These effects may be due to changes in the properties of neural membranes or to denaturation of macromolecules.

Infrared rays. A significant part of solar energy reaches the Earth in the form of infrared rays. Absorption and emission by the human body of these rays play an important part in temperature exchange and regulation of the body. The principles of infrared emission and absorption must be considered in the design of air conditioning and clothing.

Hazards of infrared radiation

Overdosage of infrared radiation, usually resulting from direct exposure to a hot object (including heating lamps) or flame, can cause severe burns. While infrared exposure is a hazard near any fire, it is particularly dangerous in the course of nuclear chain reactions. In the course of a nuclear detonation, a brief but very intense emission of infrared occurs, together with visible and ultraviolet light emitted from the fireball (flash burns). Of the total energy of nuclear explosion, as much as one-third may be in the form of thermal radiation, moving with the velocity of light. The rays will arrive almost instantaneously at regions removed from the source by only a few kilometres. Smoke or fog can effectively scatter or absorb the infrared components, and even thin clothing can greatly reduce the severity of burn effects.

EFFECTS OF VISIBLE AND ULTRAVIOLET LIGHT

Life could not exist on Earth without light from the Sun. Plants utilize the energy of the Sun's rays in the process of photosynthesis to produce carbohydrates and proteins, which serve as basic organic sources of food and energy for animals. Light has a powerful regulating influence on many biologic systems. Most of the strong ultraviolet rays of the Sun, which are hazardous, are effectively absorbed by the upper atmosphere. At high altitudes and near the Equator, the ultraviolet intensity is greater than at sea level or at northern latitudes.

Ultraviolet light of very short wavelength, below 2200 angstroms, is highly toxic for cells; in the intermediate range, the greatest killing effectiveness on cells is at about 2600 angstroms. The nucleic acids of the cell, of which genetic material is composed, strongly absorb rays in this region. This wavelength, readily available in mercury vapour, xenon, or hydrogen arc lamps, has great effectiveness for germicidal purification of the air.

Since penetration of visible and ultraviolet light in body tissues is small, only the effects of light on skin and on the visual apparatus are of consequence. When incident light exerts its action on the skin without additional external predisposing factors, scientists speak of intrinsic action. In contrast, a number of chemical or biologic agents may condition the skin for action of light; these latter phenomena are grouped under photodynamic action. Visible light, when administered following lethal doses of ultraviolet, is capable of causing recovery of the cells exposed. This phenomenon, referred to as photorecovery, has led to the discovery of various enzyme systems that are capable of restoring damaged nucleic acids in genes to their normal form. It is probable that photorecovery mechanisms are continually operative in some plants exposed to the direct action of sunlight.

The surface of the Earth is protected from the lethal ultraviolet rays of the Sun by the top layers of the atmosphere, which absorb far ultraviolet, and by ozone molecules in the stratosphere, which absorb most of the near ultraviolet. Even so, it is believed that an enzymatic mechanism operating in the skin cells of individuals continually repairs the damage caused by ultraviolet rays to the nucleic acids of the genes. Many scientists believe that chlorofluorocarbons used in aerosol spray products and in various technical applications are depleting the stratospheric ozone layer, thus exposing persons to more intense ultraviolet radiation at ground level.

Quality of light and its effects

There is some evidence to indicate that not only overall light intensity but also special compositions have differential effects on organisms. For example, in pumpkins, red light favours the production of pistillate flowers, and blue light leads to development of staminate flowers. The ratio of females to males in guppies is increased by red light. Red light also appears to accelerate the rate of proliferation of some tumours in special strains of mice. The intensity of incident light has an influence on the development of light-sensing organs; the eyes of primates reared in complete darkness, for instance, are much retarded in development.

Intrinsic action. Light is essential to the human body because of its biosynthetic action. Ultraviolet light induces the conversion of ergosterol and other vitamin precursors present in normal skin to vitamin D, an essential factor for normal calcium deposition in growing bones. While some ultraviolet light appears desirable for the formation of vitamin D, an excess amount is deleterious. Humans have a delicate adaptive mechanism that regulates light exposure of the more sensitive deeper layers of the skin. The transmission of light depends on the thickness of the upper layers of the skin and on the degree of skin pigmentation. All persons, with the exception of albinos, are born with varying amounts of melanin pigment in their skin. Exposure to light further enhances the pigmentation already present and can induce production of new pigment granules. The therapeutic possibilities of sunlight and ultraviolet light became apparent around 1900, with popularization of the idea that exposure of the whole body to sunlight promotes health.

By that time, it was already known that large doses of ultraviolet radiation cause sunburn, the wavelength of about 2800 angstroms being most effective. It induces reddening and swelling of the skin (owing to dilation of the blood vessels), usually accompanied by pain. In the course of recovery, epidermal cells are proliferated, melanin is secreted, and the outer corneal layer of dead cells is thickened. In 1928 it was first shown clearly that prolonged or repeated exposure to ultraviolet light leads to the delayed development of skin cancer. The fact that ultraviolet light, like X radiation, is mutagenic may explain its ability to cause skin cancer, but the detailed mechanism of cancer induction is not yet completely understood. There seems very little doubt, however, that skin cancer in humans is in some cases correlated with prolonged exposure to large doses of sunlight. Among blacks who are protected by rich melanin formation and thickened corneal structure of the skin, incidence of cancer of the skin is several times less frequent than it is among whites living at the same latitude.

Light sensitization

Photodynamic action. There are a number of diseases in humans and other animals in which light sensitivity is involved; for example, hydroa, which manifests itself in blisters on parts of the body exposed to sunlight. It has been suggested that this disease results from a light-sensitive porphyrin compound found in the blood.

Actually there are many organic substances and various materials of biologic origin that make cells sensitive to light. When eosin is added to a suspension of human red blood corpuscles exposed to light, the red corpuscles will break up in a process called hemolysis. Other typical photodynamic substances are rose bengal, hematoporphyrin, and phylloerythrin—all are dyes capable of fluorescence. Their toxicity manifests itself only in the presence of light and oxygen.

Some diseases in domestic animals result from ingestion of plants having photodynamic pigments. For example, St. Johnswort's disease is caused by the plant *Hypericum.* Fagopyrism results from eating buckwheat. In geeldikopp ("yellow thick head"), the photodynamic agent is produced in the animal's own intestinal tract from chlorophyll derived from plants. In humans the heritable condition of porphyria frequently is associated with light sensitivity, as are a number of somewhat ill-defined dermatologic conditions that result from exposure to sunlight. The recessively inherited rare disease xeroderma pigmentosum also is associated with light exposure; it usually results in death at an early age from tumours of the skin that develop on exposed areas. The cells of such individuals possess a serious genetic defect: they lack the ability to repair nucleic-acid lesions caused by ultraviolet light.

Certain drugs (*e.g.,* sulfanilamide) sensitize some persons to sunlight. Many cases are known in which ingestion of or skin contact with a photodynamic substance was followed by increased light sensitivity.

Effects on development and biologic rhythms. In addition to its photosynthetic effect, light exerts an influence on growth and spatial orientation of plants. This phototropism is associated with yellow pigments and is particularly marked in blue light. The presence of illumination is a profound modifier of the cellular activities in plants as well. For example, while some species of blue-green algae carry out photosynthesis in the presence of light, they do not undergo cell division.

Diffuse sensitivity to light also exists in several phyla of animals. Many protozoans react to light. Chameleons, frogs, and octopuses change colour under the influence of light. Such changes are ascribed to special organs known as chromatophores, which are under the influence of the nervous system or endocrine system. The breeding habits and migration of some birds are set in motion by small consecutive changes in the daily cycle of light.

Effects of light on biological rhythms

Light is an important controlling agent of recurrent daily physiological alterations (circadian rhythms) in many animals, including humans in all likelihood (see BEHAVIOUR, ANIMAL: *Periodic biological phenomena*). Lighting cycles have been shown to be important in regulating several types of endocrine function: the daily variation in light intensity keeps the secretion of adrenal steroids in synchrony; the annual breeding cycles in many mammals and birds appear to be regulated by light. Ambient light somehow influences the secretions of a tiny gland, the pineal body, located near the cerebellum. The pineal body, under the action of enzymes, produces melanotonin, which in higher concentrations slows down the estrous cycle; low levels of melanotonin, caused by exposure of animals to light, accelerates estrus. It is believed that light stimulates the retina, and information is then transmitted by sympathetic nerves to the pineal body.

Effects on the eyes. The wavelength of light that produces sunburn also can cause inflammation of the cornea of the eye. This is what occurs in snow blindness or after exposure to strong ultraviolet light sources. Unusual sensitivities have been reported. Ultraviolet light, like infrared or penetrating radiations, can also cause cataract of the eye lens, a condition characterized by denatured protein in the fibrous cells forming the lens (see above). The retina usually is not reached by ultraviolet light, but large doses of visible and infrared light can irreversibly bleach the visual pigments, as in sun blindness. Numerous pathological conditions of the eye are accompanied by abnormal light sensitivity and pain, a condition that is known as photophobia. The pain appears to be associated with reflex movements of the iris and reflex dilation of the blood vessels of the conjunctiva. Workers exposed to ultraviolet-light sources or to atomic flashes need to wear protective glasses. (C.A.T.)

Applications of radiation

MEDICAL APPLICATIONS

The uses of radiation in diagnosis and treatment have multiplied so rapidly in recent years that one or another form of radiation is now indispensable in virtually every branch of medicine. The many forms of radiation that are used include electromagnetic waves of widely differing wavelengths (*e.g.,* radio waves, visible light, ultraviolet radiation, X rays, and gamma rays), as well as particulate radiations of various types (*e.g.,* electrons, fast neutrons, protons, alpha particles, and pi-mesons).

Imaging techniques. Advances in techniques for obtaining images of the body's interior have greatly improved medical diagnosis. New imaging methods include various X-ray systems, positron emission tomography, and nuclear magnetic resonance imaging.

X-ray systems. In all such systems, a beam of X radiation is shot through the patient's body, and the rays that pass through are recorded by a detection device. An image is produced by the differential absorption of the X-ray photons by the various structures of the body. For example, the bones absorb more photons than soft tissues; they thus cast the sharpest shadows, with the other body components (organs, muscles, etc.) producing shadows of varying intensity.

The conventional X-ray system produces an image of all structures in the path of the X-ray beam, so that a radiograph of, say, the lungs shows the ribs located in front and as well as in back. Such extraneous details often make it difficult for the physician examining the X-ray image to identify tumours or other abnormalities on the lungs. This problem has been largely eliminated by computerized tomographic (CT) scanning, which provides a cross-sectional image of the body part being scrutinized. Since its introduction in the 1970s, CT scanning, also called computerized axial tomography (CAT), has come to play a key role in the diagnosis and monitoring of many kinds of diseases and abnormalities.

Computerized tomographic scanning

In CT scanning a narrow beam of X rays is rotated around the patient, who is surrounded by several hundred X-ray photon detectors that measure the strength of the penetrating photons from many different angles. The X-ray data are analyzed, integrated, and reconstructed by a computer to produce images of plane sections through the body onto the screen of a television-like monitor. Computerized tomography enables more precise and rapid visualization and location of anatomic structures than has been possible with ordinary X-ray techniques. In many cases, lesions can be detected without resorting to exploratory surgery.

Positron emission tomography (PET). This imaging technique permits physicians to determine patterns of blood flow, blood volume, oxygen perfusion, and various other physiological, metabolic, and immunologic parameters. It is used increasingly in diagnosis and research, especially of brain and heart functions.

PET involves the use of chemical compounds "labeled" with short-lived positron-emitting isotopes such as carbon-11 and nitrogen-13, positron cameras consisting of photomultiplier-scintillator detectors, and computerized tomographic reconstruction techniques. After an appropriately labeled compound has been injected into the body, quantitative measurements of its activity are made throughout the sections of the body being scanned by the detectors. As the radioisotope disintegrates, positrons are annihilated by electrons, giving rise to gamma rays that are detected simultaneously by the photomultiplier-scintillator combinations positioned on opposite sides of the patient.

Nuclear magnetic resonance (NMR) imaging. This

Application of magnetic resonance imaging

method, also referred to as magnetic resonance imaging (MRI), involves the beaming of high-frequency radio waves into the patient's body while it is subjected to a strong magnetic field. The nuclei of different atoms in the body absorb radio waves at different frequencies under the influence of the magnetic field. The NMR technique makes use of the fact that hydrogen nuclei (protons) respond to an applied radio frequency by reemitting radio waves of the same frequency. A computer analyzes the emissions from the hydrogen nuclei of water molecules in body tissues and constructs images of anatomic structures based on the concentrations of such nuclei. This use of proton density makes it possible to produce images of tissues that are comparable, and in some cases superior, in resolution and contrast to those obtained with CT scanning. Moreover, since macroscopic movement affects NMR signals, the method can be adapted to measure blood flow. The ability to image atoms of fluorine-19, phosphorus-31, and other elements besides hydrogen permit physicians and researchers to use the technique for various tracer studies as well. (For information on tracer studies, see ATOMS: *Radioactivity.*)

Other radiation-based medical procedures. *Radionuclides in diagnosis.* Radionuclides have come to play a key role in certain diagnostic procedures. These procedures may be divided into two general types: (1) radiographic imaging techniques for visualizing the distribution of an injected radionuclide within a given organ as a means of studying the anatomic structure of the organ; and (2) quantitative assay techniques for measuring the absorption and retention of a radionuclide within an organ as a means of studying the metabolism of the organ.

Notable among the radionuclides used for imaging purposes is technetium-99*m*, a gamma-ray emitter with a six-hour half-life, which diffuses throughout the tissues of the body after its administration. Among the radionuclides suitable for metabolic studies, iodine-131 is one of the most widely used. This gamma-ray emitter has a half-life of eight days and concentrates in the thyroid gland, and so provides a measure of thyroid function.

Important gamma-ray emitters

Treating cancer and other diseases with highly energetic forms of ionizing radiation. In addition to X rays and gamma rays, densely ionizing particles—neutrons, protons, mesons, alpha particles, and heavy ions, for example—have been used increasingly to treat cancer and other lesions. Such high-LET radiations (see above) offer potential advantages over conventional X rays and gamma rays in that they have per given dose greater capacity to damage tumours, particularly deep-seated ones, and can be applied more precisely to the lesion under treatment, causing less injury to surrounding tissue. The results of these radiations in cancer treatment, though preliminary, are promising.

Ultraviolet radiation therapy. Ultraviolet radiation ("Wood's" light) is used diagnostically to detect fluorescent materials that are present in certain disorders—*e.g.*, some fungal diseases of the skin. It is also widely employed in combination with a radiosensitizing agent such as 8-methoxypsoralen to treat psoriasis. In this approach, known as PUVA therapy, the entire surface of the skin is bathed repeatedly with ultraviolet radiation.

Phototherapy. Intense visible light is used in treating newborns' jaundice, a disease characterized by the accumulation of the pigment bilirubin in the bloodstream during the first few days of life. Since wavelengths of 420–480 nanometres absorbed in the skin expedite detoxification and elimination of the pigment, the affected infant is bathed in visible light for 12–24 hours in treating the disorder.

Laser surgery

Treatment with lasers. The laser is used increasingly for surgery, as it has proved to be a finely controlled and relatively bloodless means of dissecting and destroying tissue. By "tuning" the laser to different wavelengths, one can vary the extent to which its light is absorbed in particular cells or cellular inclusions. Certain types of lesions, such as birthmarks of the "port-wine stain" variety, can thus be destroyed more or less selectively, with minimal damage to surrounding tissues.

The laser also is well-suited for treating lesions of the inner eye, since a beam of laser light can pass through the intact cornea and lens without harming them. In addition, lasers are used together with optical fibres to treat lesions inside blood vessels and in other locations that are not readily accessible to standard surgical intervention. In this procedure, a fibre-optic probe is inserted into a vessel or body cavity by means of cannulas.

Diathermy. Microwave radiation has long been used for warming internal parts of the body in treating deep-seated inflammations and various other disorders. This approach, termed diathermy, is also being explored as a means of inducing hyperthermia in tumour tissue as an adjunct to radiation therapy (or chemotherapy) in the treatment of certain types of cancer. (A.C.U.)

APPLICATIONS IN SCIENCE AND INDUSTRY

Photochemistry. The principal applications of photochemistry (including photography) are in the initiation of reactions by light that can pass through glass or quartz windows. Such light has a wavelength of not less than about 185 nanometres. Light of shorter wavelength is also effective, but the windows required (sapphire, lithium fluoride, or extraordinarily thin aluminum) and the associated mechanical difficulties seriously limit application of photochemical methods in the range from 185 nanometres down to a conceivable lower limit of about 85 nanometres. Photochemical techniques are particularly applicable when a specific initial process (the breakage of a particular bond in a molecule of a particular substance, for example) is required. For such purposes, high-intensity ultraviolet lamps are generally employed, the window is either glass or quartz, and the initiation reaction is limited to the relatively thin layers in which the light is absorbed. The processes include the photochlorination of aromatic compounds (such as benzene, toluene, and xylene), sulfhydration of olefins, production of cyclohexanone oxime, photopolymerization (principally in surface-curing processes), sulfoxidation, and vitamin-D synthesis. Tunable lasers provide a potential means of initiating photochemical processes of practical interest, one such example being the separation of isotopes.

High-energy radiation. The large-scale use of such ionizing radiation for modifying and synthesizing materials, known as radiation processing, represents a minor yet significant technology. It involves irradiating materials either with a beam of electrons produced by a high-voltage particle accelerator or with gamma rays emitted by the radioisotope cobalt-16 or, in a few cases, cesium-137. The electrons are generally accelerated to an energy range of 0.15–10 MeV. (By comparison, the electron energy in a typical television set is only 0.025 MeV.) The gamma rays given off by cobalt-16 have an energy of 1.25 MeV, while those emitted by cesium-137 have approximately half that amount.

Electron and gamma-ray irradiation

Exposure to such electrons and gamma rays does not induce radioactivity in the materials irradiated, and so the technique can be used in the manufacture or processing of many kinds of consumer and industrial products. Moreover, radiation processing has several major advantages over conventional technologies. It consumes far less energy than thermally and chemically initiated processes and at the same time causes less environmental pollution. Paints and certain other coatings, for example, can be cured at room temperature with one-tenth of the energy required in heat curing. Radiation preservation of food involves substantially less energy expenditure than that associated with either refrigeration or canning. A radiation source that releases 1 kilowatt of gamma energy (roughly equivalent to the electrical requirements of a toaster) can irradiate 10 tons of potatoes per hour; the exposure to a small dose of ionizing radiation inhibits sprouting and thereby delays spoilage.

Radiation processing in the modification of plastics

Because of such advantages, radiation processing has found increasingly wider application. It has proved particularly valuable in the processing of plastics. Chemical reactions induced by electron-beam irradiation permit the cross-linking of polymers that make up the foamed plastic used for sound and thermal insulation. A large fraction of the wire and cable employed in high-temperature

applications and much of the wiring in telecommunications equipment are covered by insulation cross-linked by electron irradiation. The heat-shrinkable polyethylene packaging for hams and turkeys and various other poultry products is manufactured by the same process. The coating of certain audio and video recording tape is cured by exposure to electron beams, as is the rubber in a large percentage of automobile tires. Sterilization of disposable medical supplies, such as syringes, blood transfusion kits, and hospital gowns, is usually done with gamma rays. Other potential applications of radiation processing include the treatment of a wide assortment of food products so as to reduce the amount of chemical preservatives employed, the treatment of sewage for pathogen reduction, and the precipitation of sulfur dioxide and nitrogen oxide (the primary source materials for acid rain) from the stack gases of electric power plants and smelting facilities that burn fossil fuels.

Radiation source technology has developed to a point where reliable, safe, and inexpensive sources are readily available. When electron accelerators are used, radioactivity is not involved in any aspect of the process and there is no conceivable hazard to the surrounding community. In processing facilities that use gamma radiation, the source is encapsulated in a double layer of stainless steel to prevent the escape of radioactivity to the environment. Other safeguards minimize the possibility of accidental exposure of either the plant personnel or the population at large.

Lasers. As noted above, lasers have become a valuable tool in medicine. They also have important uses in a number of other areas, as, for example, communications. Laser light can carry voice messages and digitally encoded information and can do so in large amounts because of its high frequency. Except in satellite-to-satellite communications, laser beams are transmitted via optical fibres. The speed with which the focal spot of a narrow laser beam can be controlled makes it suitable for a variety of applications in information processing—*e.g.,* use in optical scanners, optical disc storage systems, and certain types of computer printers.

The destructive action of lasers

A highly intense laser beam can instantly vaporize the surface of a target. When laser pulses are concentrated on frozen deuterium-tritium pellets, they can initiate nuclear fusion (see ATOMS: *Energy from atoms: Nuclear fusion*). High-powered lasers can be used as space weapons to destroy reconnaissance and communications satellites and perhaps even ballistic missiles. These same capabilities have led to the use of lasers in research as well as in surgery. The laser microprobe is used for microanalysis of surface composition. Laser beams have been found to have a selective effect on cellular components, or organelles: those components that absorb light of the wavelength of the beam are destroyed, whereas transparent parts of the cells remain unaffected. Organelles such as mitochondria, which are responsible for cell respiration, or chloroplasts, which are involved in plant-cell photosynthesis, can be separately studied in this manner.

An intense beam of laser light can be used for small-scale cutting, scribing, and welding in certain industrial processes. Laser "pens" capable of producing such high-intensity light beams have proved useful in the assembly of various electronic components, such as computer memory and logic units consisting of integrated arrays of microcircuit elements (Figure 9).

The use of special dyes can alter laser action. The availability of high-pulse-intensity laser beams is also revolutionizing microscopy. It is possible to photograph microaction in a small fraction of a second and to use holography for image synthesis. (C.A.T./Jo.Si./Ed.)

By courtesy of General Electric Research and Development Center

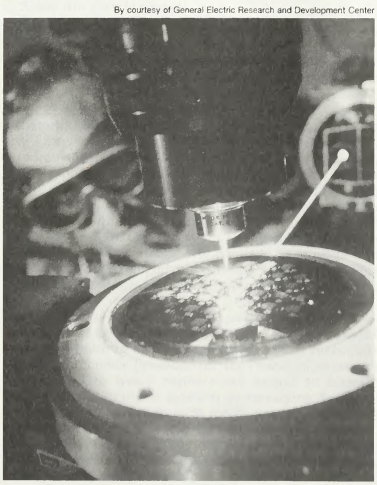


Figure 9: A laser scribing microscopic lines that connect integrated circuits.

BIBLIOGRAPHY

General: Historical works include MAX PLANCK, *Introduction to Theoretical Physics,* vol. 4, *Theory of Light* (1932, reprinted 1957; originally published in German, 1927), the classic work on the subject of light and quanta. Other forms of electromagnetic radiation are covered in OTTO GLASSER, *Wilhelm Conrad Röntgen and the Early History of the Roentgen Rays* (1933; originally published in German, 1931). See also R.W. DITCHBURN, *Light,* 3rd ed., 2 vol. (1976), a well-presented text on physical optics that, though not too mathematical, does require understanding of the use of differential equations.

Interaction of radiation with matter: GERHARD K. ROLLEFSON and MILTON BURTON, *Photochemistry and the Mechanism of Chemical Reactions* (1939, reprinted 1946); WILLIAM ALBERT NOYES and PHILIP ALBERT LEIGHTON, *The Photochemistry of Gases* (1941, reprinted 1966), are both classic works that include material on internal conversion and predissociation. The language of intersystem crossing is discussed in detail in a comprehensive text, JACK G. CALVERT and JAMES N. PITTS, *Photochemistry* (1966). The actual effects of radiation on solids are thoroughly summarized in HANS A. BETHE and JULIUS ASHKIN, "Passage of Radiations Through Matter," in EMILIO SEGRÈ (ed.), *Experimental Nuclear Physics,* vol. 1 (1953), pp. 166–357; G.J. DIENES and G.H. VINEYARD, *Radiation Effects in Solids* (1957); and DOUGLAS S. BILLINGTON and JAMES H. CRAWFORD, JR., *Radiation Damage in Solids* (1961). For a survey of radiation effects on aqueous solutions and organic compounds, see J.W.T. SPINKS and R.J. WOODS, *An Introduction to Radiation Chemistry,* 2nd ed. (1976); *Actions chimiques et biologiques des radiations* (annual 1955–71), the first survey on a large variety of subjects in radiation chemistry written by scientists largely about their own work—some volumes have been translated into English with the title, *The Chemical and Biological Action of Radiations;* and MAX S. MATHESON and LEON M. DORFMAN, *Pulse Radiolysis* (1969), an excellent book on techniques in radiation chemistry. *Advances in Photochemistry* (irregular) is concerned mainly with surveys of advances in the field. See also J.F. ZIEGLER (ed.), *Ion Implantation: Science and Technology* (1984), a treatment of ion implantation mechanisms, techniques, effects, and practical applications; and ORLANDO AUCIELLO and ROGER KELLY (eds.), *Ion Bombardment Modification of Surfaces: Fundamentals and Applications* (1984), covering surface alteration mechanisms with major emphasis on topographical effects. (M.Bu./A.Moz./M.Lu.)

Radiological units and measurements: For descriptions, see RALPH E. LAPP and HOWARD L. ANDREWS, *Nuclear Radiation Physics,* 4th ed. (1972); and INTERNATIONAL COMMISSION ON RADIATION UNITS AND MEASUREMENTS, *Radiation Quantities and Units* (1980).

Biologic effects of radiation: (Ionizing): General information is given in CHARLES WESLEY SHILLING (ed.), *Atomic Energy Encyclopedia in the Life Sciences* (1964). Introductory information on radiation biology is given in J.E. COGGLE, *Biological Effects of Radiation,* 2nd ed. (1983); JOHN W. GOFMAN, *Radiation and Human Health* (1981); DANIEL S. GROSCH and LARRY E. HOPWOOD, *Biological Effects of Radiations,* 2nd ed. (1979); and ERIC J. HALL, *Radiation and Life,* 2nd ed. (1984). More specialized topics are covered in ASSEMBLY OF LIFE SCIENCES (U.S.) COMMITTEE ON THE BIOLOGICAL EFFECTS OF IONIZING RADIATIONS, *The Effects on Populations of Exposure to Low Levels of Ionizing Radiation, 1980* (1980); MERRILL EISENBUD, *Environ-*

mental Radioactivity: From Natural, Industrial, and Military Sources, 3rd ed. (1987); DONALD J. PIZZARELLO and RICHARD L. WITCOFSKI, *Medical Radiation Biology*, 2nd ed. (1982); UNITED NATIONS SCIENTIFIC COMMITTEE ON THE EFFECTS OF ATOMIC RADIATION, *Ionizing Radiation: Sources and Biological Effects* (1982), and *Genetic and Somatic Effects of Ionizing Radiation* (1986); ARTHUR C. UPTON, *Radiation Injury: Effects, Principles, and Perspectives* (1969); and ARTHUR C. UPTON *et al.* (eds.), *Radiation Carcinogenesis* (1986).

(*Non-ionizing*): Microwave radiation is treated in NATIONAL COUNCIL OF RADIATION PROTECTION AND MEASUREMENTS, *Biological Effects of Ultrasound: Mechanisms and Clinical Implications* (1983), and *Biological Effects and Exposure Criteria for Radiofrequency and Electromagnetic Fields* (1986); and R.C. PETERSEN, "Bioeffects of Microwaves: A Review of Current Knowledge," *Journal of Occupational Medicine*, 25(2):103–110 (February 1983). Visible and ultraviolet radiations are the subject of WALTER HARM, *Biological Effects of Ultraviolet Radiation* (1980); A. JARRET (ed.), *The Photobiology of the Skin: Lasers and the Skin* (1984); KENDRIC C. SMITH (ed.), *Topics in Photomedicine* (1984); and RICHARD J. WURTMAN, MICHAEL J. BAUM, and JOHN T. POTTS, JR. (eds.), *The Medical and Biological Effects of Light* (1985).

(*Nuclear war*): Radiation effects of a nuclear war are discussed in SAMUEL GLASSTONE and PHILIP J. DOLAN (eds.), *The Effects of Nuclear Weapons*, 3rd ed. (1977); JEAN PETERSEN and DON HINRICHSEN (eds.), *Nuclear War: The Aftermath* (1982); JULIUS LONDON and GILBERT F. WHITE (eds.), *The Environmental Effects of Nuclear War* (1984); and FREDRIC SOLOMON and ROBERT Q. MARSTON (eds.), *The Medical Implications of Nuclear War* (1986).

(*Radiation protection and safety*): Procedures and recommendations for protection are analyzed in INTERNATIONAL COMMISSION ON RADIOLOGICAL PROTECTION, *Recommendations of the International Commission on Radiological Protection* (1977, reprinted with supplements, 1987), and *Nonstochastic Effects of Ionizing Radiation* (1984); NATIONAL COUNCIL ON RADIATION PROTECTION AND MEASUREMENTS, *Ionizing Radiation Exposure of the Population of the United States* (1987); and MARILYN F. NOZ and GERALD Q. MAGUIRE, JR., *Radiation Protection in the Radiologic and Health Sciences*, 2nd ed. (1985).

Applications of radiation: (*Medical*): Radiological imaging techniques are explored in R.P. CLARK and M.P. GOFF (eds.), *Recent Developments in Medical and Physiological Imaging* (1986); W.-D. HEISS and M.F. PHELPS (eds.), *Positron Emission Tomography of the Brain* (1983); ALEXANDER R. MAGULIS and CHARLES A. GOODING (eds.), *Diagnostic Radiology, 1987* (1987); and ALBERT A. MOSS, ERNEST J. RING, and CHARLES B. HIGGINS (eds.), *NMR, CT, and Interventional Radiology* (1984). Radiation therapy is addressed by GILBERT H. FLETCHER, *Textbook of Radiotherapy*, 3rd ed. (1980); and ERNEST J. RING and GORDON K. MCLEAN, *Interventional Radiology: Principles and Techniques* (1981). Specific uses of phototherapy are outlined in AUDREY K. BROWN and JANE SHOWACRE (eds.), *Phototherapy for Neonatal Hyperbilirubinemia: Long-Term Implications* (1977); WAYNE F. MARCH (ed.), *Ophthalmic Lasers: Current Clinical Uses* (1984); and WARWICK L. MORISON, *Phototherapy and Photochemotherapy of Skin Disease* (1983). (A.C.U.)

(*Scientific and industrial*): A review of ionizing radiation processing in medicine and industrial manufacturing is found in VITOMIR MARKOVIC, "Modern Tools of the Trade," *International Atomic Energy Agency Bulletin*, 27(1):33–39 (Spring 1985). Industrial uses are explored in JOSEPH SILVERMAN, "Radiation Processing: The Industrial Applications of Radiation Chemistry," *Journal of Chemical Education*, 58(2):168–173 (Feb. 1981). INTERNATIONAL ATOMIC ENERGY AGENCY, *Industrial Application of Radioisotopes and Radiation Technology* (1982), is a collection of conference papers. (Jo.Si.)

Relativity

Relativity is concerned with measurements made by different observers moving relative to one another. In classical physics it was assumed that all observers anywhere in the universe, whether moving or not, obtained identical measurements of space and time intervals. According to relativity theory, this is not so, but their results depend on their relative motions.

There are actually two distinct theories of relativity known in physics, one called the special theory of relativity, the other the general theory of relativity. Albert Einstein proposed the first in 1905, the second in 1916. Whereas the special theory of relativity is concerned primarily with electric and magnetic phenomena and with their propagation in space and time, the general theory of relativity was developed primarily in order to deal with gravitation. Both theories centre on new approaches to space and time, approaches that differ profoundly from those useful in everyday life; but relativistic notions of space and time are inextricably woven into any contemporary interpretation of physical phenomena ranging from the atom to the universe as a whole.

This article will set forth the principal ideas comprising both special and general relativity. It will also deal with some implications and applications of these theories.

The article is divided into the following sections:

The special theory of relativity 501
  Historical background
  Relativity of space and time
  Consequences
The general theory of relativity 504
  Physical origins
  The principle of equivalence
  Curved space-time
  Confirmation of the theory
  Conceptual implications of general relativity
  Schwarzschild's solution of the field equations
Applications of relativistic principles 506
  Particle accelerators
  Relativistic particle physics
  Relativistic cosmology
Modifications of general relativity 507
Bibliography 508

THE SPECIAL THEORY OF RELATIVITY

Historical background. Classical physics owes its definitive formulation to the British scientist Sir Isaac Newton. According to Newton, when one physical body influences another body, this influence results in a change of that body's state of motion, its velocity; that is to say, the force exerted by one particle on another results in the latter's changing the direction of its motion, the magnitude of its speed, or both. Conversely, in the absence of such external influences, a particle will continue to move in one unchanging direction and at a constant rate of speed. This statement, Newton's first law of motion, is known as the law of inertia.

As motion of a particle can be described only in relation to some agreed frame of reference, Newton's law of inertia may also be stated as the assertion that there exist frames of reference (so-called inertial frames of reference) with respect to which particles not subject to external forces move at constant speed in an unvarying direction. Ordinarily, all laws of classical mechanics are understood to hold with respect to such inertial frames of reference. Each frame of reference may be thought of as realized by a grid of surveyor's rods permitting the spatial fixation of any event, along with a clock describing the time of its occurrence.

According to Newton, any two inertial frames of reference are related to each other in that the two respective grids of rods move relative to each other only linearly and uniformly (with constant direction and speed) and without rotation, whereas the respective clocks differ from each other at most by a constant amount (as do the clocks adjusted to two different time zones on Earth) but go at the same rate. Except for the arbitrary choice of such a constant time difference, the time appropriate to various inertial frames of reference then is the same: If a certain physical process takes, say, one hour as determined in one inertial frame of reference, it will take precisely one hour with respect to any other inertial frame; and if two events are observed to take place simultaneously by an observer attached to one inertial frame, they will appear simultaneous to all other inertial observers. This universality of time and time determinations is usually referred to as the absolute character of time. The idea that a universal time can be used indiscriminately by all, irrespective of their varying states of motion—that is, by a person at rest at his home, by the driver of an automobile, and by the passenger aboard an airplane—is so deeply ingrained in most people that they do not even conceive of alternatives. It was only at the turn of the 20th century that the absolute character of time was called into question as the result of a number of ingenious experiments described below.

The idea of absolute time

As long as the building blocks of the physical universe were thought to be particles and systems of particles that interacted with each other across empty space in accordance with the principles enunciated by Newton, there was no reason to doubt the validity of the space-time notions just sketched. This view of nature was first placed in doubt in the 19th century by the discoveries of a Danish physicist, Hans Christian Ørsted, the English scientist Michael Faraday, and the theoretical work of the Scottish-born physicist James Clerk Maxwell, all concerned with electric and magnetic phenomena. Electrically charged bodies and magnets do not affect each other directly over large distances, but they do affect one another by way of the so-called electromagnetic field, a state of tension spreading throughout space at a high but finite rate, which amounts to a speed of propagation of approximately 186,000 miles (300,000 kilometres) per second. As this value is the same as the known speed of light in empty space, Maxwell hypothesized that light itself is a species of electromagnetic disturbance; his guess has been confirmed experimentally, first by the production of lightlike waves by entirely electric and magnetic means in the laboratory by a German physicist, Heinrich Hertz, in the late 19th century.

Both Maxwell and Hertz were puzzled and profoundly disturbed by the question of what might be the carrier of the electric and magnetic fields in regions free of any known matter. Up to their time, the only fields and waves known to spread at a finite rate had been elastic waves, which appear to the senses as sound and which occur at low frequencies as the shocks of earthquakes, and surface waves, such as water waves on lakes and seas. Maxwell called the mysterious carrier of electromagnetic waves the ether, thereby reviving notions going back to antiquity. He attempted to endow his ether with properties that would account for the known properties of electromagnetic waves, but he was never entirely successful. The ether hypothesis, however, led two U.S. scientists, Albert Abraham Michelson and Edward Williams Morley, to conceive of an experiment (1887) intended to measure the motion of the ether on the surface of the Earth in their laboratory. On the reasonable hypothesis that the Earth is not the pivot of the whole universe, they argued that the motion of the Earth relative to the ether should result in slight variations in the observed speed of light (relative to the Earth and to the instruments of a laboratory) travelling in different directions. The measurement of the speed

Maxwell's ether hypothesis

of light requires but one clock, if, by use of a mirror, a pencil of light is made to travel back and forth so that its speed is measured by clocking the total time elapsed in a round trip at one site; such an arrangement obviates the need for synchronizing two clocks at the ends of a one-way trip. Finally, if one is concerned with variations in the speed of light, rather than with an absolute determination of that speed itself, then it suffices to compare with each other round-trip-travel times along two tracks at right angles to each other, and that is essentially what Michelson and Morley did. To avoid the use of a clock altogether, they compared travel times in terms of the numbers of wavelengths travelled, by making the beams travelling on the two distinct tracks interfere optically with each other. (If the waves meet at a point when both are in the same phase—*e.g.*, both at their peak—the result is visible as the sum of the two in amplitude; if the peak of one coincides with the trough of the other, they cancel each other and no light is visible. Since the wavelengths are known, the relative positions of the peaks give an exact measure of how far one wave has advanced with respect to the other.) This highly precise experiment, repeated many times with ever-improved instrumental techniques, has consistently led to the result that the speed of light relative to the laboratory is the same in all directions, regardless of the time of the day, the time of the year, and the elevation of the laboratory above sea level.

The special theory of relativity resulted from the acceptance of this experimental finding. If an Earth-bound observer could not detect the motion of the Earth through the ether, then, it was felt, probably any observer, regardless of his state of motion, would find the speed of light the same in all directions.

Relativity of space and time. An Irish and a Dutch physicist, George Francis FitzGerald and Hendrik Antoon Lorentz, independently showed that the negative outcome of Michelson's and Morley's experiment could be reconciled with the notion that the Earth is travelling through the ether, if one hypothesizes that any body travelling through the ether is foreshortened in the direction of travel (though its dimensions at right angles to the motion remain undisturbed) by a ratio that increases with increasing speed. If v denotes the speed of the body relative to the ether, and c is the speed of light, that ratio equals the quantity $(1 - v^2/c^2)^{1/2}$. At ordinary speeds, c is so much greater than v that the fraction, practically speaking, is zero, and the ratio becomes $\sqrt{1}$, which is 1; *i.e.*, the foreshortening is nil; as v approaches c, however, the fraction becomes significant. The travelling body would be flattened completely if its speed through the ether should ever reach that of light.

Suppose, now, that the variations in the speed of light were to be determined not by interference but by means of an exceedingly accurate clock and assume further that in such a modified experiment (whose actual performance is precluded at present, because even the best atomic clocks available do not possess the requisite accuracy) the motion through the ether were still imperceptible, then. Lorentz showed, one would have to conclude that all clocks moving through the ether are slowed down compared to clocks at rest in the ether, again by the factor $(1 - v^2/c^2)^{1/2}$. Thus, all rods and all clocks would be modified systematically, regardless of materials and construction design, whenever they were moving relative to the ether. Accordingly, for theoretical analysis, one would have to distinguish between "apparent" and "true" space and time measurements, with the further proviso that "true" dimensions and "true" times could never be determined by any experimental procedure.

"Apparent" and "true" time and space

Conceptually, this was an unsatisfactory situation, which was resolved by Albert Einstein in 1905. Einstein realized that the key concept, on which all comparisons between differently moving observers and frames of reference depended, is the notion of universal, or absolute, simultaneity; that is to say, the proposition that two events that appear simultaneous to any one observer will also be judged to take place at the same time by all other observers. This appears to be a straightforward proposition, provided that knowledge of distant events can be obtained practically instantaneously. Actually, however, there is no known method of signalling faster than by means of light or radio waves or any other electromagnetic radiation, all of which travel at the same rate, c.

Suppose, now, that someone on Earth observes two events, say two supernovae (suddenly erupting very bright stars) appearing in different parts of the sky. Nothing can be said about whether these two supernovae emerged simultaneously or not from merely noting their appearance in the sky; it is necessary to know also their respective distances from the observer, which typically may amount to several hundred or several thousand light-years (one light-year, the distance light moves in one year, equals approximately 5.88×10^{12} miles, or 9.46×10^{12} kilometres). By the time one sees the eruption of a supernova, it has in actuality faded back into invisibility hundreds of years ago. Applying this simple idea to the observations and measurements made by different observers of the same events, Einstein demonstrated that if each observer applied the same method of analysis to his own data, then events that appeared simultaneous to one would appear to have taken place at different times to observers in different states of motion. Thus, it is necessary to speak of relativity of simultaneity.

Once this theoretical deduction is accepted, the findings of FitzGerald and Lorentz lend themselves to a new interpretation. Whenever two observers are associated with two distinct inertial frames of inference in relative motion to each other, their determinations of time intervals and of distances between events will disagree systematically, without one being "right" and the other "wrong." Nor can it be established that one of them is at rest relative to the ether, the other in motion. In fact, if they compare their respective clocks, each will find that his own clock will be faster than the other; if they compare their respective measuring rods (in the direction of mutual motion), each will find the other's rod foreshortened. The speed of light will be found to equal the same value, $c = 186{,}000$ miles per second, relative to every inertial frame of reference and in all directions. The status of Maxwell's ether is thereby cast in doubt, as its state of motion cannot be ascertained by any conceivable experiment. Consequently, the whole notion of an ether as the carrier of electromagnetic phenomena has been eliminated in contemporary physics.

The Lorentz transformations

The mathematical equations that relate space and time measurements of one observer to those of another, moving observer are known as Lorentz transformations. If the relative motion is measured along the x-axis and if its magnitude is v, these expressions are (see Figure 1):

$$x' = (1 - v^2/c^2)^{-1/2}\,(x - vt),\ y' = y,\ z' = z,$$

$$t' = (1 - v^2/c^2)^{-1/2}\,(t - vx/c^2).$$

Consequences. *The limiting character of the speed of light.* As the speed of one inertial frame of reference relative to another is increased, its rods appear increasingly foreshortened and its clocks more and more slowed down. As this relative speed approaches c, both of these effects increase indefinitely. The relative speed of the two frames cannot exceed c if light and other electromagnetic phenomena are to travel at the speed c in all directions when viewed from either frame of reference. Hence the special theory of relativity forecloses relative speeds of frames of reference greater than c. As an inertial frame of reference can be associated with any material object in uniform nonrotational motion, it follows that no material object can travel at a rate of speed exceeding c.

This conclusion is self-consistent only because under the

From P. Bergmann, *The Riddle of Gravitation*; Charles Scribner's Sons

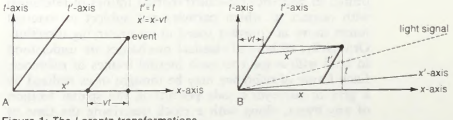


Figure 1: *The Lorentz transformations.*
(A) Nonrelativistic transformation; (B) relativistic transformation (see text).

Lorentz transformations the velocity of a body with respect to one inertial frame of reference is related to its velocity with respect to another frame not by the Newtonian rule that the difference in velocities equals the relative velocity between the two frames but by a more involved formula, which takes into account the changes in scale length, in clock time, and in simultaneity. If all velocities involved have the same direction, then the velocity (see Figure 2) in one frame, u, is related to the velocity in the other frame, u', by the expression stating that u' equals the sum of u and v divided by 1 plus the product of u and v divided by the square of c:

$$u' = (u + v)/(1 + uv/c^2).$$

As long as neither u nor v exceeds the speed of light, c, u' also will be less than c.

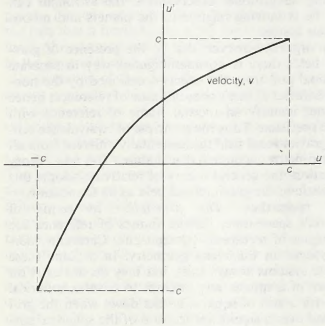


Figure 2: Velocities of the same body in two frames of reference (see text).

Variable mass. The mass of a material body is a measure of its resistance to a change in its state of motion caused by a given force. The larger the mass, the smaller the acceleration. If a material body is already moving at a speed approaching the speed of light, it must offer increasing resistance to any further acceleration so as not to cross the threshold of c. Hence the special theory of relativity leads to the conclusion that the mass of a moving body m is related to the mass that it would have if at rest, m_0, by a formula in which m equals m_0 divided by the square root of one minus the fraction v^2/c^2:

$$m = (1 - v^2/c^2)^{-1/2} m_0.$$

Relativistic mass

This changing value of the mass of the moving body, m, is called the relativistic mass. As v approaches c, the figure within the parentheses approaches zero and the mass m becomes infinitely large.

The relativistic mass formula may be interpreted as indicating that the relativistic mass of a body exceeds its rest mass m_0 by an amount that equals its kinetic energy E, divided by c^2: $m - m_0 = E/c^2$. Hence the hypothesis that generally the energy is c^2 times the mass, or $E = mc^2$, and that energy and mass are, in fact, equivalent physical concepts, differing only by the choice of their units. This hypothesis has been verified experimentally, in that all massive particles have been converted into forms of energy (for instance, gamma radiation) and conversely have been created out of pure energy. It was in part the recognition of this relationship that led to research out of which grew the technology of nuclear fission and fusion.

Invariant intervals. Data on pure time intervals obtained with respect to two relatively moving inertial frames of reference will differ and so will data on spatial distances. It is possible, however, to form from time intervals plus distances a single expression that will have the same value with respect to all inertial frames of reference. If the time interval between two distant events be denoted by T and their distance from each other by L, an expression involving a quantity symbolized by τ can be derived in which τ squared equals the square of the time interval minus the fraction of distance squared over speed of light squared: $\tau^2 = T^2 - L^2/c^2$. This will have the same value as $\overline{T}^2 - \overline{L}^2/c^2$, with $\overline{T}$ and $\overline{L}$ having been obtained in another inertial frame of reference. If τ^2 is positive, then τ is called the invariant (timelike) interval between the two events. If τ^2 is negative, then the expression λ, derived from the above as $\lambda^2 = L^2 - c^2T^2$, will be called the invariant (spacelike) interval.

The invariant interval between two instants in the history of one physical system equals the ordinary time lapse T measured by means of a clock at rest relative to that physical system, because, in such a comoving frame of reference, L vanishes. That is why such an invariant (timelike) interval is also referred to as the "proper time" elapsed between the two instants. Any clock will read its own proper time.

The "twin paradox." Given an inertial frame of reference and two similar material systems ("twins")—for instance, two atomic clocks of identical design—suppose that one of these clocks remains permanently at rest in the given frame, whereas the other clock is moved at a high speed first in one direction away from the first clock and subsequently in the opposite direction until the two clocks are again close to each other. According to the Lorentz transformation, the second clock has been slower than the first throughout its journey, and hence it shows a smaller lapse of time than the clock that has remained at rest. By reading the clocks, one can then tell which clock has remained at rest, which one has moved. This difference in behaviour of the two clocks has been called the clock paradox or the twin paradox.

The clock "paradox"

The "paradox" supposedly consists of a violation of the principle of relativity, according to which no asymmetric distinctions exist between different inertial frames of reference. The fallacy of this argument lies in the fact that no inertial frame of reference is associated with the second clock, as it cannot have moved free of acceleration throughout its journey: at least once its velocity (*i.e.*, the direction of its motion) must have been changed drastically, so as to enable it ever to return to its mate. Hence no violation of the principle of relativity; no paradox is involved. Various experiments on moving particles and atoms have indeed confirmed the predictions of the theory.

Four-dimensional space-time. The German mathematical physicist Hermann Minkowski pointed out that the invariant interval between two events has some of the properties of the distance in Euclidean geometry. Based on Euclidean geometry, the Cartesian coordinate system is designed to identify any point (event) in space by its reference to three mutually perpendicular lines or axes meeting at an arbitrary point of origin. The distance s between two events, in accordance with Pythagoras' theorem, in any Cartesian (rectilinear) coordinate system is obtained by taking the square root of the sum of the squares of coordinate distances, $s^2 = x^2 + y^2 + z^2$, and its value is independent of the choice of coordinate system, though the values of x, y, and z are not. The invariant interval, similarly, is the square root of a sum and difference of squares of intervals of both space and time. Accordingly, Minkowski suggested that space and time should be thought of as comprising a single four-dimensional continuum, space-time, often also referred to as the Minkowski universe. Events, localized as regards both space and time, are the natural analogues of points in ordinary three-dimensional geometry; in the history of one particle, its proper time resembles the arc length of a curve in three-space.

In Minkowski's space-time the invariant interval may be either timelike or spacelike. If $L^2 - c^2T^2$ for two events happens to be zero, the invariant interval is neither, but null, or lightlike, as a light signal emanating from the earlier of the two events may just pass the second as the latter occurs. By contrast, in ordinary geometry the distance between two points, s, vanishes only if the two points coincide. To this extent the analogy between space-time and ordinary space is imperfect.

Minkowski's four-dimensional, geometric approach to relativity appears to add to the original physical concepts of relativity mostly a new terminology but not much

else. Nevertheless, for the further conceptual development of relativity Minkowski's contribution has been of inestimable value.

THE GENERAL THEORY OF RELATIVITY

Physical origins. The general theory of relativity derives its origin from the need to extend the new space and time concepts of the special theory of relativity from the domain of electric and magnetic phenomena to all of physics and, particularly, to the theory of gravitation. As space and time relations underlie all physical phenomena, it is conceptually intolerable to have to use mutually contradictory notions of space and time in dealing with different kinds of interactions, particularly in view of the fact that the same particles may interact with each other in several different ways—electromagnetically, gravitationally, and by way of so-called nuclear forces.

Newton's theory of gravity

Newton's explanation of gravitational interactions must be considered one of the most successful physical theories of all time. It accounts for the motions of all the constituents of the solar system with uncanny accuracy, permitting, for instance, the prediction of eclipses hundreds of years ahead. But Newton's theory visualizes the gravitational pull that the Sun exerts on the planets and the pull that the planets in turn exert on their moons and on each other as taking place instantaneously over the vast distances of interplanetary space, whereas according to relativistic notions of space and time any and all interactions cannot spread faster than the speed of light. The difference may be unimportant, for practical reasons, as all of the members of the solar system move at relative speeds far less than $^1/_{1,000}$ of the speed of light; nevertheless, relativistic space-time and Newton's instantaneous action at a distance are fundamentally incompatible. Hence Einstein set out to develop a theory of gravitation that would be consistent with relativity.

Proceeding on the basis of the experience gained from Maxwell's theory of the electric field, Einstein postulated the existence of a gravitational field that propagates at the speed of light, c, and that will mediate an attraction as closely as possible equal to the attraction obtained from Newton's theory. From the outset it was clear that mathematically a field theory of gravitation would be more involved than that of electricity and magnetism. Whereas the sources of the electric field, the electric charges of particles, have values independent of the state of motion of the instruments by which these charges are measured, the source of the gravitational field, the mass of a particle, varies with the speed of the particle relative to the frame of reference in which it is determined and hence will have different values in different frames of reference. This complicating factor introduces into the task of constructing a relativistic theory of the gravitational field a measure of ambiguity, which Einstein resolved eventually by invoking the principle of equivalence.

The principle of equivalence. Everyday experience indicates that in a given field of gravity, such as the field caused by the Earth, the greater the mass of a body the greater the force acting on it. That is to say, the more massive a body the more effectively will it tend to fall toward the Earth; in fact, in order to determine the mass of a body one weighs it—that is to say, one really measures the force by which it is attracted to the Earth, whereas the mass is properly defined as the body's resistance to acceleration. Newton noted that the ratio of the attractive force to a body's mass in a given field is the same for all bodies, irrespective of their chemical constitution and other characteristics, and that they all undergo the same acceleration in free fall; this common rate of acceleration on the surface of the Earth amounts to an increase in speed by approximately 32 feet (about 9.8 metres) per second every second.

This common rate of gravitationally caused acceleration is illustrated dramatically in space travel during periods of coasting. The vehicle, the astronauts, and all other objects within the space capsule undergo the same acceleration, hence no acceleration relative to each other. The result is apparent weightlessness: no force holds the astronaut to the floor of his cabin or a liquid in an open container.

Weightlessness

To this extent, the behaviour of objects within the freely coasting space capsule is indistinguishable from the condition that would be encountered if the space capsule were outside all gravitational fields in interstellar space and moved in accordance with the law of inertia. Conversely, if a space capsule were to be accelerated upward by its rocket engines in the absence of gravitation, all objects inside would behave exactly as if the capsule were at rest but in a gravitational field. The principle of equivalence states formally the equivalence, in terms of local experiments, of gravitational forces and reactions to an accelerated noninertial frame of reference (*e.g.*, the capsule while the rockets are being fired) and the equivalence between inertial frames of reference and local freely falling frames of reference. Of course, the principle of equivalence refers strictly to local effects: looking out of his window and performing navigational observations, the astronaut can tell how he is moving relative to the planets and moons of the solar system.

Einstein argued, however, that in the presence of gravitational fields there is no unambiguous way to separate gravitational pull from the effects occasioned by the noninertial character of one's chosen frame of reference; hence one cannot identify an inertial frame of reference with complete precision. Thus the principle of equivalence renders the gravitational field fundamentally different from all other force fields encountered in nature. The new theory of gravitation, the general theory of relativity, adopts this characteristic of the gravitational field as its foundation.

Curved space-time. *The principles.* In terms of Minkowski's space-time, inertial frames of reference are the analogues of rectilinear (straight-line) Cartesian coordinate systems in Euclidean geometry. In a plane these coordinate systems always exist, but they do not exist on the surface of a sphere: any attempt to cover a spherical surface with a grid of squares breaks down when the grid is extended over a significant fraction of the spherical surface. Thus a plane is a flat surface, whereas the surface of a sphere is curved. This distinction, based entirely on internal properties of the surface itself, classifies the surface of a cylinder as flat, as it can be rolled off on a plane and thus is capable of being covered by a grid of squares.

Einstein conjectured that the presence of a gravitational field causes space-time to be curved (whereas in the absence of gravitation it is flat), and that this is the reason that inertial frames cannot be constructed. The curved trajectory of a particle in space and time resulting from the effects of gravitation would then represent not a straight line (which exists only in flat spaces and space-times) but the straightest curve possible in a curved space-time, a geodesic. Geodesics on a sphere (such as the surface of the Earth) are the great circles. (The plane of any great circle goes through the centre of the Earth.) They are the least curved lines one can construct on the surface of a sphere, and they are the shortest curves connecting any two points. The geodesics of space-time connect two events (or two instants in the history of one particle) with the greatest lapse of proper time, as was indicated in the earlier discussion of the twin paradox.

If the presence of a gravitational field amounts to a curvature of space-time, then the description of the gravitational field in turn hinges on a mathematical elucidation of the curvature of four-dimensional space-time. Before Einstein, the German mathematician Bernhard Riemann (1826–66) had developed methods related directly to the failure of any attempt to construct square grids. If one were to construct within any small piece of (two-dimensional) surface a quadrilateral whose sides are geodesics, if the surface were flat, the sum of the angles at the four corners would be 360°. If the surface is not flat, the sum of the angles will not be 360°. The deviation of the actual sum of the angles from 360° will be proportional to the area of the quadrilateral; the amount of deviation per unit of surface will be a measure of the curvature of that surface. If the surface is imbedded in a higher dimensional continuum, then one can consider similarly unavoidable angles between vectors constructed as parallel as possible to each other at the four corners of the quadrilateral, and thus associate several distinct components of curvature

Riemannian space

with one surface. And, of course, there are several independent possible orientations of two-dimensional surfaces—for instance, six in a four-dimensional continuum, such as space-time. Altogether there are 20 distinct and independent components of curvature defined at each point of space-time; in mathematics these are referred to as the 20 components of Riemann's curvature tensor.

The mathematical expression. Einstein discovered that he could relate 10 of these components in a natural way to the sources of the gravitational field, mass (or energy), density, momentum density, and stress, if he were to duplicate approximately Newton's equations of the gravitational field and, at the same time, formulate laws that would take the same form regardless of the choice of frame of reference. The remaining 10 components may be chosen arbitrarily at any one point but are related to each other by partial differential equations at neighbouring points. Einstein derived a field equation that, along with the rule that a freely falling body moves along a geodesic, forms the comprehensive treatment of gravitation known as the general theory of relativity.

Confirmation of the theory. The general theory of relativity is constructed so that its results are approximately the same as those of Newton's theories as long as the velocities of all bodies interacting with each other gravitationally are small compared with the speed of light—*i.e.*, as long as the gravitational fields involved are weak. The latter requirement may be stated roughly in terms of the escape velocity. The escape velocity is defined as the minimal speed with which a projectile must be endowed at any given location to enable it to fly off to infinitely removed regions of the universe without the application of further force. On the surface of the Earth the escape velocity is approximately 11.2 kilometres (6.95 miles) per second. A gravitational field is considered strong if the escape velocity approaches the speed of light, weak if it is much smaller. All gravitational fields encountered in the solar system are weak in this sense.

The success of Newton's theory, incidentally, must be considered a confirmation of the general theory of relativity to the extent that that application of the theory remains confined to situations involving small relative speeds and weak fields. Obviously, any superiority of the new theory over the old one may be inferred only if their predictions disagree and if those of the general theory of relativity are confirmed by experiment and observation.

As the principle of equivalence forms the cornerstone of general relativity, its verification is crucial. Highly precise experiments with this objective were performed between 1888 and 1922 by a Hungarian physicist, Roland, Baron von Eötvös, and his collaborators, who confirmed the principle to an accuracy of one part in 10^8, and in the 1960s by an American physicist, Robert Dicke, who achieved an accuracy of one part in 10^{11}. Subsequently the Soviet physicist V.L. Braginsky further improved the accuracy to one part in 10^{12}. Through this work the principle of equivalence has become one of the most precisely confirmed general principles of contemporary physics.

Tests of the principle of equivalence

Some other new predictions of general relativity are explained below.

Advance of Mercury's perihelion. The major axes of the elliptical trajectories of the planets about the Sun turn slowly within their planes because of the interactions of the planets with each other, but it was discovered in the 19th century that interplanetary perturbations could not account fully for the turning rate of Mercury's orbit, leaving unexplained about 43″ of arc per century. The general theory of relativity, however, accounts exactly for this discrepancy. In 1967 Dicke—and more recently Henry Allen Hill, also of the United States—suggested that a small part of Mercury's perihelion advance may be caused by the slight flattening of the Sun at its poles, thus opening the way for possible modification of general relativity. On the other hand, support for Einstein's original version of the theory has come from a comprehensive evaluation of solar system data by the American investigator Ronald W. Hellings and from investigations of the binary pulsar system PSR 1913+16 by the American astronomer Joseph H. Taylor.

Gravitational redshift. General relativity predicts that the wavelength of light emanating from sources within a gravitational field will increase (shift toward the red end of the spectrum) by an amount proportional to the gravitational potential at the site of the source. This effect was found first in astronomical objects, particularly in stars called white dwarfs, on whose surfaces the gravitational potential is large. The best quantitative confirmation of gravitational redshift was obtained in laboratory experiments in Great Britain and the United States in the 1960s; an accuracy of one part in 100 was achieved in measuring the minute difference in gravitational potential between two sites differing in altitude by a few metres.

Optical effects of gravitation. General relativity predicts that the curvature of space-time results in the apparent bending of light rays passing through gravitational fields and in an apparent reduction of their speeds of propagation. The bending was first observed, within a couple of years of Einstein's publication of the new theory, during a total eclipse, when stellar images near the occulted disk of the Sun appeared displaced by fractions of 1″ of arc from their usual locations in the sky. The associated delay in travel time was observed in the late 1960s, when ultraintense radar pulses were reflected off Mercury and Venus just as these planets were passing behind the Sun. These experiments are difficult to perform and their accuracy is difficult to evaluate, but it seems conservative to conclude that they confirm the relativistic effect within a few parts in 100. Finally, extended massive objects such as galaxies may act as "gravitational lenses," providing more than one optical path for light emanating from a source far behind the lens and thus producing multiple images. Such multiple images, typically of quasars, had been discovered by the early 1980s.

Gravitational waves. General relativity predicts the occurrence of gravitational waves, whose properties should resemble in some respects those of electromagnetic waves: they should travel at the same speed, c, and they should be polarized. Joseph Weber, an American physicist, announced in 1969 that he had detected events that might be caused by incoming gravitational waves—namely, vibrations occurring simultaneously in pairs of large aluminum cylinders, about 1,000 kilometres apart and each weighing several tons. Although these detectors had been insulated with great care from all other potential sources of such vibrations, the separation of gravitational signals from ordinary thermal noise (Brownian motion) presents delicate problems of instrumentation and interpretation, which proved difficult to resolve to the satisfaction of other experimenters attempting to repeat Weber's observations.

Weber's approach has been refined by the choice of different materials for the vibrating masses, by cryogenic techniques reducing the level of thermal noise, and by other improvements. A fundamentally different technique, replacing Weber's stationary cylinders by independently moving masses whose distances from each other would be measured by interferometric means, also has been investigated. While these efforts at direct detection of gravitational waves were under way, observations of the binary pulsar PSR 1913+16 indicated that this double star system is losing energy at precisely the rate that corresponds to the emission of gravitational radiation according to the theory of general relativity.

The discovery of gravitational waves would represent an important confirmation of the validity of the theory. Also, such waves might become the basis of an entirely new technology of astronomical observation, as they are believed to be the most penetrating kind of radiation imaginable.

Future astrophysical tests. The properties of certain astronomical objects, such as quasars (see below *Relativistic cosmology*), pulsars (extremely dense stars that emit electromagnetic pulses with great regularity), very bright galaxies at the cores of which extraordinary amounts of energy are being emitted, and jets of matter moving at relativistic speeds, imply that there are processes involving gravitational fields so strong that general relativity is needed to interpret the observations, which in turn will provide new tests of that theory.

Conceptual implications of general relativity. The gen-

eral theory of relativity represents a further modification of classical concepts of space and time that goes far beyond those implicit in the special theory. The special theory does away with the absolute character of time and with the absolute distance between two objects that are at rest relative to each other. The geometric concepts appropriate to the special theory are the four-dimensional space-time continuum, in which events that are fixed in space and in time are represented by points, often referred to as world points (to distinguish them from the points of ordinary three-dimensional space), and the histories of particles moving through space in the course of time by curves (world curves); the representations of particles that are not accelerated by forces are straight lines.

Minkowskian space-time

Minkowski's space-time is a rigidly flat continuum, as is the three-dimensional space of Euclid's geometry. Distances between world points are measured by the invariant intervals, whose magnitudes do not depend on the particular coordinate system, or frame of reference, used. The Minkowski universe is homogeneous; that is to say, geometric figures constructed at any site may be transferred to another site without distortion. Finally, among all the possible frames of reference there is a special set, the inertial frames of reference, just as in ordinary space the rectilinear coordinate systems are distinguished by their simplicity among all conceivable coordinate systems. Space-time serves as the immutable backdrop of all physical processes, without being affected by them.

In general relativity, space-time also is a four-dimensional continuum, with invariant intervals being defined at least locally between events taking place close to each other. But only small regions of space-time resemble the continuum envisaged by Minkowski, just as small bits of a spherical surface appear nearly planar. In the broad sense, according to general relativity, space-time is curved, and this curvature is equivalent to the presence of a gravitational field. Far from being rigid and homogeneous, the general-relativistic space-time continuum has geometric properties that vary from point to point and that are affected by local physical processes. Space-time ceases to be a stage, or scaffolding, for the dynamics of nature; it becomes an integral part of the dynamic process. General relativity, it has been said, makes physics part of geometry. It may also be claimed that general relativity makes geometry part of physics, that is to say, of a natural science. Not only are the properties of space and time subject to scientific investigation, to a study by means of experiments, but specific properties, such as the amount of curvature in a particular location at a specified time, are to be measured with the help of physical instruments.

Though the general theory of relativity is universally accepted as the most satisfactory basis of the gravitational force now known, it has not been completely fused with quantum mechanics, of which the central concept is that energy and angular momentum exist only in finite and discrete lumps, called quanta. Since the 1920s quantum mechanics has been the sole rationale of the forces that act between subatomic particles; gravitation doubtless is one of these forces, but its effects are unobservably small in comparison to electromagnetic and nuclear forces. Relativistic phenomena in the subatomic realm have been adequately dealt with by merging quantum mechanics with the special, not the general, theory.

Many physicists, foremost among them Einstein himself, tried during the first half of the 20th century to enrich the geometric structure of space-time so as to encompass all known physical interactions. Their goal, a unified field theory, remained elusive but was brought nearer during the late 1960s by the successful unification of the electromagnetic and the so-called weak nuclear force.

Unified field theory

Schwarzschild's solution of the field equations. Immediately on publication of Einstein's paper on general relativity, the German astronomer Karl Schwarzschild found a mathematical solution to the new field equations, which corresponds to the gravitational field of a compact massive body, such as a star or planet, and which is now referred to as Schwarzschild's field. If the mass that serves as the source of the field is fairly diffuse, so that the gravitational field on the surface of the astronomical body is fairly weak, Schwarzschild's field will exhibit physical properties similar to those described by Newton. Gross deviations will be found if the mass is so highly concentrated that the field on the surface is strong. At the time of Schwarzschild's work, in 1916, this appeared to be a theoretical speculation; but with the discovery of pulsars and their interpretation as probable neutron stars composed of matter that has the same density as atomic nuclei (so-called nuclear matter), the possibility exists that strong fields may soon be accessible to astronomical observation.

The most conspicuous feature of the Schwarzschild field is that if the total mass is thought of as concentrated at the very centre, a point called a singularity, then at a finite distance from that centre, the Schwarzschild radius, the geometry of space-time changes drastically from that to which we are accustomed. Particles and even light rays cannot penetrate from inside the Schwarzschild radius to the outside and be detected. Conversely, to an outside observer any objects approaching the Schwarzschild radius appear to take an infinite time to penetrate toward the inside. There cannot be any effective communication between the inside and the outside, and the boundary between them is called an event horizon.

The exterior and the interior of the Schwarzschild radius are not cut off from each other entirely, however. Suppose an observer were to attach himself to a particle that is falling freely straight toward the centre and that this observer is equipped with a clock that reads its own proper time. This observer would penetrate the Schwarzschild radius within a finite proper time (see Figure 3); moreover, he would find no abnormalities in his environment as he did so. The reason is that his clock would deviate from one permanently kept outside and at a constant distance from the centre, so grossly that the same event that seen from the outside takes forever occurs within a finite time to the free-falling observer.

From P. Bergmann, *The Riddle of Gravitation*; Charles Scribner's Sons

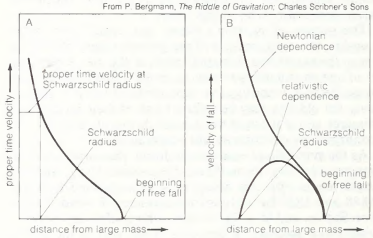


Figure 3: Free fall in a Schwarzschild field (A) seen by an outside observer and (B) in terms of proper time of the falling object.

These peculiarities of the Schwarzschild field may well have practical applications in astronomy. In 1931 the Indian-born U.S. astrophysicist Subrahmanyan Chandrasekhar, and in 1939 the U.S. physicist J. Robert Oppenheimer, established that a star whose mass exceeds the mass of the Sun by an appreciable factor is bound to contract and, eventually, to collapse under the influence of its own gravity, no matter how resistant its constituent matter. As many stars are believed to have such large masses, it is likely that there already exist some collapsed stars, so-called black holes. Though continuing to make its presence known by the gravitational attraction it exerts on other stars, a black hole would not emit light, and thus be invisible, hence its name.

APPLICATIONS OF RELATIVISTIC PRINCIPLES

Particle accelerators. Modern particle accelerators raise particles to speeds very near that of light. At these energies and speeds the differences in behaviour predicted by classical physics and by the special theory of relativity are huge; the machines must be designed in accordance with relativistic principles, or they will not operate.

Electron synchrotrons operate at energies of several thousand million electron volts, which means that the rela-

tivistic mass of an electron orbiting at maximum energy is roughly 10,000 times its rest mass. Accordingly, the magnetic field required to maintain the electrons in orbit is 10,000 times as powerful as it would have to be if nonrelativistic physics held, at the same speed. On the other hand, at that given energy the speed of the electrons is in fact very nearly equal to the speed of light, the difference amounting to no more than one part in 100,000,000 (10^8). At the same energy, but by nonrelativistic mechanics, the speed of the electrons would be about 100 times the speed of light. This difference has a very practical consequence: in those particle accelerators designed for highly relativistic energies, the synchrotrons, particles are injected into a circular orbit already near the speed of light, and their velocities change only slightly as their energies are brought up to the highest design value. If the orbit diameter is kept nearly constant, particles at all energies will circulate at the same frequency, and only the magnetic field that keeps them in orbit needs to be increased to keep pace with the increasing mass. The accelerating voltage is applied at the constant frequency required so that the particles will always be accelerated forward.

Operations of the synchrotron

Relativistic particle physics. The physics of subatomic particles depends on the principles of the special theory of relativity. These principles have their most direct application when particles are created, annihilated, or converted into different particles. In most particle transformations, large amounts of energy are involved; the total (rest) masses of the particles involved in the transformations will change, and this change will be related to the amounts of energy expended or gained by the rule that the change in mass (Δm_0) is balanced by a corresponding change in energy (ΔE), divided by the square of the speed of light (c^2): $\Delta m_0 = -c^{-2}\Delta E$. This rule has been confirmed universally and, by now, is being taken for granted.

The units, or quanta, of electromagnetic energy, called photons, have long been regarded as a species of particle in which are combined the properties of zero rest mass with nonvanishing relativistic mass, because they travel at the speed of light. The relativistic mass equals its total energy E divided by c^2. The energy of a photon also is equal to the product of its frequency ν and Planck's constant h. The relativistic mass of a photon can be checked experimentally if the photon is absorbed or deflected in its interactions with particles, when the change in its linear momentum (product of velocity and relativistic mass) results in a recoil by the other particles. If a high-frequency photon, a gamma photon, collides with a free electron, the result is called the Compton effect, which involves both an observable recoil on the part of the electron and an altered frequency of the deflected photon. Again, relativity is confirmed by experiment.

It has been conjectured that gravitational waves, also, are composed of zero-rest-mass quanta travelling at the speed of light (gravitons). As the quantum theory of the gravitational field has not been definitely established and as the detection of individual gravitons may remain beyond experimental capabilities for years to come, the existence of gravitons cannot be considered assured.

There is another species of zero-rest-mass particles, produced in radioactive decay involving the ejection of electrons or positrons from atomic nuclei (so-called beta decay). These particles, known as neutrinos, have no electric charge and travel at the speed of light. Several distinct species of neutrinos are now known, each produced in a different kind of beta decay. Neutrinos interact with other particles extremely weakly. As a result, they can traverse large amounts of matter with little chance of being deflected or absorbed. Though their existence has been confirmed beyond a doubt, their detection and detailed examination remain challenging problems.

Neutrinos

Relativistic cosmology. Theories concerning the structure and history of the whole universe have assumed an increasingly empirical aspect in the 20th century. Beginning in the 1960s, particularly, a combination of new observation techniques, new discoveries, and applications of special and general relativity has resulted in considerable progress. The most important techniques added to those of observations by means of visible light were radio astronomy; infrared, ultraviolet, X-ray, and gamma-ray astronomy from extraterrestrial platforms; cosmic-ray investigations; neutrino astronomy; and examination of the Moon and other astronomical bodies by unmanned and manned space exploration.

Edwin Powell Hubble, a U.S. astronomer, had discovered that the more distant astronomical objects exhibited a shift of spectral lines toward the red (long wavelength) end of the spectrum, the extent of the shift increasing the greater their distance from Earth. This cosmological red shift has been generally interpreted as evidence of rapid recession of these distant objects in an expanding universe. The present rate of expansion is expressed as the amount of recession per unit distance and is known as the Hubble constant. It amounts to about a mile per second recessional velocity for a distance of 10^5 light-years. Equivalently, if the expansion has been occurring at a constant rate, it must have started about 2×10^{10} years ago.

Quasars, also called quasi-stellar objects (QSO's), appear to be structures that combine extreme luminosity (100 times that of a bright galaxy) with great compactness, taking up less space than the distance between the Sun and its nearest neighbour star. Wherever a spectral analysis of a quasar's emitted light has been possible, the spectral lines have been found considerably red shifted. If these red shifts are cosmological (an interpretation now accepted by most astronomers), some quasars are more distant from the Galaxy than any other known objects. As such they may provide indications of the large-scale structure of the universe, which could not be obtained from investigations confined to "close" surroundings. The term close is to be understood in relation to distance in which Hubble's red shift becomes large ("cosmological distances"), distances amounting to thousands of millions of light-years.

Finally, the term primeval fireball refers to the discovery of an all-pervasive background of radiation whose frequencies lie in the border region between microwave radio frequencies and infrared, corresponding to wavelengths of the order of millimetres and centimetres. In the early 1970s this radiation was interpreted as a remnant of the original intensive radiation that must have been associated with the early history of the universe, when matter was both extremely dense and extremely hot; hence its name. Its spectral composition, which has been the object of intensive investigation, might provide some clues to the early history of the universe.

The notion of the primeval fireball

General relativity contributes to a theoretical discussion of cosmology the idea that the universe as a whole need not be flat even on the average and that it probably is not. Even if one were to assume that on a very large (cosmological) scale the universe is homogeneous and isotropic (*i.e.*, having the same properties in all directions), which appears a reasonable working hypothesis in the absence of any evidence to the contrary, there are a number of different possibilities. The universe might be spatially open (as a flat one surely is), or it might be closed, somewhat as the surface of a sphere is closed, without boundaries. Likewise, in the time direction the universe might be either open or closed; it is a little difficult to visualize a time-wise closed universe, which appears to be inconsistent with ordinary notions of cause and effect. But because these notions are distilled out of normal experience, they might be inapplicable on the scale of billions of years. In brief, many different cosmological models have been proposed and investigated theoretically, but observational information does not seem to favour one particular type. The information appears to favour types that expand from an early stage involving fireball conditions.

MODIFICATIONS OF GENERAL RELATIVITY

An outgrowth of a unified field theory of the early 1920s has been the development of a class of theories based on the hypothesis that underlying the four-dimensional space-time of our experience is a manifold having a higher dimensionality, whose geometric structure can accommodate all known force fields, including those associated with stable and unstable subatomic particles. Though these concepts remained highly speculative, they offered much promise and occupied many investigators.

Apart from the attempts to devise unified field theories, several modifications of general relativity have been proposed during the late 20th century. One of these was presented by the British scientist Fred Hoyle, whose results, together with the proposals of the astronomers Hermann Bondi and Thomas Gold, became the basis of the so-called steady-state cosmological theory. Bondi, Gold, and Hoyle opposed the "big-bang" theory of the origin of the universe, arguing instead that matter is being created continuously at a very low rate, just sufficient to maintain the constant average density of the universe in spite of the observed expansion. Though the steady-state hypothesis evoked much interest for some years, the existence of the cosmic background radiation (established in the 1960s) has been generally accepted as proof that the universe has in fact passed through a highly dense stage.

BIBLIOGRAPHY. Among expositions for general readers are ALBERT EINSTEIN, *Relativity: The Special and General Theory: A Popular Exposition,* 17th ed. (1961; originally published in German, 1917), a popularization for the lay reader of a classic work written by one of the greatest scientists of all time; BERTRAND RUSSELL, *The ABC of Relativity,* 4th rev. ed. edited by FELIX PIRANI (1985); ALBERT EINSTEIN and LEOPOLD INFELD, *The Evolution of Physics* (1938, reissued 1961); LEOPOLD INFELD, *Albert Einstein: His Work and Its Influence on Our World* (1950), two books that cover the whole of physics, with special emphasis on relativity (Infeld was one of Einstein's chief collaborators in the 1930s); HERMANN BONDI, *Relativity and Common Sense: A New Approach to Einstein* (1964, reissued 1980); ROBERT GEROCH, *General Relativity from A to B* (1978), a beautiful book explaining general relativity in an exciting and insightful manner to an audience of humanists; PETER G. BERGMANN, *The Riddle of Gravitation,* rev. and updated ed. (1987, reissued 1992), a work that emphasizes the general theory of relativity and includes a discussion of research; SAM LILLEY, *Discovering Relativity for Yourself* (1981), a work that covers both theories; GEORGE F.R. ELLIS and RUTH M. WILLIAMS, *Flat and Curved Space-times* (1988); ERIC CHAISSON, *Relatively Speaking: Relativity, Black Holes, and the Fate of the Universe* (1988); and CLIFFORD M. WILL, *Was Einstein Right?: Putting General Relativity to the Test,* 2nd ed. (1993), the last two works stressing the astronomical aspect of relativity.

Presentations for readers with technical training include H.A. LORENTZ *et al., The Principle of Relativity* (1923, reissued 1952), a collection of fundamental research papers, all in English; ALBERT EINSTEIN, *The Meaning of Relativity,* 5th ed., trans. from German (1955, reprinted 1988), based on lectures by Einstein delivered in 1921, with two appendixes containing Einstein's views on cosmology through 1945, and his work on the "non-symmetric" unified field theory to the time of his death in 1955; ABRAHAM PAIS, *"Subtle Is the Lord—": The Science and the Life of Albert Einstein* (1982), containing a wealth of material on relativity, its history, and its relationship to the whole of physics; DAVID BOHM, *The Special Theory of Relativity* (1965, reprinted 1989), a thoroughgoing treatment of the special theory combined with a discussion of the philosophical foundations of physics; A.P. FRENCH, *Special Relativity* (1968, reissued 1984), an introduction at the undergraduate level; HERMANN BONDI, *Cosmology,* 2nd ed. (1961), a survey of cosmology at a technical level, including observational data through the late 1950s; PETER G. BERGMANN, *Introduction to the Theory of Relativity* (1942, reissued 1976); C. MØLLER, *The Theory of Relativity,* 2nd ed. (1972); J.L. SYNGE, *Relativity: The Special Theory,* 2nd ed. (1964, reissued 1972), and *Relativity: The General Theory* (1960, reissued 1971); CHARLES W. MISNER, KIP S. THORNE, and JOHN ARCHIBALD WHEELER, *Gravitation* (1973), technical texts, on the graduate level, that represent distinct approaches to the subject by active research workers; STEVEN WEINBERG, *Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity* (1972), by a Nobel laureate; J.L. MARTIN, *General Relativity: A Guide to Its Consequences for Gravity and Cosmology* (1988), a text on the general theory; S.W. HAWKING and G.F.R. ELLIS, *The Large Scale Structure of Space-Time* (1973), a work principally concerned with the geometric aspects of general relativity on a global scale; ROBERT M. WALD, *General Relativity* (1984), and *Space, Time, and Gravity,* 2nd ed. (1992), written by one of the experts and an active contributor in the field; ROBERTO TORRETTI, *Relativity and Geometry* (1983), an exposition of the general and special theories from a geometric perspective, for the advanced reader; and WOLFGANG RINDLER, *Introduction to Special Relativity,* 2nd ed. (1991). Two historical works are DON HOWARD and JOHN STACHEL (eds.), *Einstein and the History of General Relativity* (1989); and JEAN EISENSTAEDT and A.J. KOX (eds.), *Studies in the History of General Relativity* (1992).

(P.G.Be./Ed.)

The Study and Classification of Religions

The history of mankind has shown the pervasive influences of religion, and thus the study of religion, involving the attempt to understand its significance, its origins, and its myriad forms, has become increasingly important in modern times. Broadly speaking, the study of religion comprehends two aspects: assembling information and interpreting systematically the material gathered in order to elicit its meaning. The first aspect involves the psychological and historical study of religious life and must be supplemented by such auxiliary disciplines as archaeology, ethnology, philology, literary history, and other similar disciplines. The facts of religious history and insight into the development of the historical religious communities are the foundation of all else in the study of religion. Beyond the historical basis lies the task of seeing the entirety of human religious experience from a unified or systematic point of view. The student of religions attempts not only to know the variety of beliefs and practices of *homo religiosus* ("religious man"), but also to understand the structure, nature, and dynamics of religious experience. The student of religion attempts to discover principles that operate throughout religious life—on the analogy of a sociologist seeking the laws of human social behaviour—to find out whether there are also laws that operate in the religious sphere. Only with the attempt to discern the system and structure binding together the differentiated historical richness of religion does a true science of religion, or *Religionswissenschaft,* begin.

The 19th century saw the rise of the study of religion in the modern sense, in which the techniques of historical enquiry, the philological sciences, literary criticism, psychology, anthropology, sociology, and other disciplines were brought to bear on the task of estimating the history, origins, and functions of religion. Rarely, however, has there been unanimity among scholars about the nature of the subject, partly because assumptions about the revealed nature of the Christian (or other) religion or assumptions about the falsity of religion become entangled with questions concerning the historical and other facts of religion. Thus, the subject has, throughout its history, contained elements of controversy.

This article is divided into the following sections:

The descriptive study of essence and content 509
- Nature and significance 509
  - The essence of religion and the context of religious beliefs, practices, and institutions
  - Neutrality and subjectivity in the study of religion
- History of the study of religion 510
  - The Greco-Roman period
  - The Middle Ages to the Reformation
  - The beginnings of the modern period
- Basic aims and methods 513
  - Historical, archaeological, and literary studies
  - Anthropological approaches to the study of religion
  - Sociological studies of religion
  - The psychology of religion
  - Philosophy of religion
  - Theological studies
  - History and phenomenology of religion
- Problems and directions 522

The classification of forms and phenomena 523
- Function and significance 523
- Principles of classification 524
  - Normative
  - Geographical
  - Ethnographic-linguistic
  - Philosophical
  - Morphological
  - Phenomenological
  - Other principles
- Conclusion 529

Bibliography 529

The descriptive study of essence and content

NATURE AND SIGNIFICANCE

Attempts to arrive at a definition of religion

The essence of religion and the context of religious beliefs, practices, and institutions. An acceptable definition of religion itself is difficult to attain. Attempts have been made to find an essential ingredient in all religions (*e.g.,* the numinous, or spiritual, experience; the contrast between the sacred and the profane; belief in gods or in God), so that an "essence" of religion can be described. But objections have been brought against such attempts, either because the rich variety of men's religions makes it possible to find counterexamples or because the element cited as essential is in some religions peripheral. The gods play a very subsidiary role, for example, in most phases of Theravāda ("Way of the Elders") Buddhism. A more promising method would seem to be that of exhibiting aspects of religion that are *typical* of religions, though they may not by universal. The occurrence of the rituals of worship is typical, but there are cases, however, in which such rituals are not central. Thus, one of the tasks of a student of religion is to gather together an inventory of types of religious phenomena.

The fact that there is dispute over the possibility of finding an essence of religion means that there is likewise a problem about speaking of the study of religion or of religions, for it is misleading to think of religion as something that "runs through" religions. This brings to light one of the major questions of method in the study of the subject. In practice, a religion is a particular system, or a set of systems, in which doctrines, myths, rituals, sentiments, institutions, and other similar elements are interconnected. Thus, in order to understand a given belief that occurs in such a system, it is necessary to look at its particular context—that is, other beliefs held in the system, rituals, and other aspects. Belief in the lordship of Christ in the early Christian Church, for example, has to be seen in the context of a belief in the Creator and of the sacramental life of the community. This systematic character of a religion has been referred to by the 20th-century Dutch theologian Hendrik Kraemer as "totalitarian"; but a better term would be "organic." Thus, there arises the problem of whether or not one belief or practice embedded in an organic system can properly be compared to a similar item in another organic system. To put the matter in another way, every religion has its unique properties, and attempts to make interreligious comparisons may hide these unique aspects. Most students of religion agree, however, that valid comparisons are possible, though they are difficult to make. Indeed, one can see the very uniqueness of a religion through comparison, which includes a contrast. The importance of setting religions side by side is why the study of religions is sometimes referred to as the "comparative study of religion"—though a number of scholars prefer not to use this phrase, partly because some comparative work in the past has incorporated value judgments about other religions.

Uniqueness and similarities in religions

But even if an inventory of types of belief and practices can be gathered—so as to provide a typical profile of what counts as religion—the absence of a tight definition means that there will always be a number of cases about which it is difficult to decide. Thus, some ideologies, such as Soviet Marxism, Maoism, and Fascism, may have analogies to religion. Certain attempts at an essentialist definition of

religion, such as that of the German-American theologian Paul Tillich (1886–1965), who defined religion in terms of man's ultimate concern, would leave the way open to count these ideologies as proper objects of the study of religion. Tillich, incidentally, calls them "quasi-religions." Though there is no consensus on this point among scholars, it is not unreasonable to hold that the frontier between traditional religions and modern ideologies represents one part of the field to be studied.

Neutrality and subjectivity in the study of religion. Discussion about religion has been complicated further by the attempt of some Christian theologians, notably Karl Barth (1886–1968), to draw a distinction between the Gospel (the proclamation peculiar to Christianity) and religion. This distinction depends, to some extent, upon taking a projectionist view of religion as a human product. This tradition goes back in modern times to the seminal work of the German philosopher Ludwig Feuerbach (1804–72), who proposed that God was the extension of human aspirations, and is found in the work of Karl Marx, Sigmund Freud, and others. The distinction attempts to draw a line between the transcendent, as it reveals itself to men, and religion, as a human product involved in the response to revelation. The difficulty of the distinction consists chiefly in a denial that God (*e.g.,* Yahweh or Christ) as the object of man's response is a "religious" being (*i.e.,* God is transcendent, not "religious" in the sense of being a part of the human product), and thus the question about revelation as a religious fact needs to be answered. This account of religion, however, incorporates a theory about it, which is characteristic of a number of definitions of religion and creates a difficulty in that the field—namely, the study of religion—is being defined in terms of a theory within it.

Subjectivity in the study of religion. There are, however, doubts about how far there can be neutrality and objectivity in the study of religion. Is it possible indeed to understand a faith without holding it? If it is not possible, then cross-religious comparisons would mostly break down, for normally it is not possible to be inside more than one religion. But it is necessary to be clear about what objectivity and subjectivity in religion means. Religion can be said to be subjective in at least two senses. First, the practice of religion involves inner experiences and sentiments, such as feelings of God guiding the life of the devotee. Here religion involves subjectivity in the sense of individual experience. Religion may also be thought to be subjective because the criteria by which its truth is decided are obscure and hard to come by, so that there is no obvious "objective" test in the way in which there is for a large range of empirical claims in the physical world. As to the first sense, one of the challenges to the student of religion is the problem of evoking its inner, individual side, which is not observable in any straightforward way. In considering a religion, however, the scholar is not only concerned with individual responses but also with communal ones. In any case, very often he is confronted only with texts describing beliefs and stories, so that he needs to infer the inner sentiments that these both evoke and express. The adherent of a faith is no doubt authoritative as to his own experience, but he is not necessarily so in regard to the communal significance of the rites and institutions in which he participates. Thus, the matter of coming to understand the inner side of a religion involves a dialectic between participant observation and dialogical (interpersonal) relationship with the adherents of the other faith. Consequently, the study of religion has strong similarities to, and indeed overlaps with, anthropology. General agreement upon scholarly methods, however, does not exist, partly because different scholars have come to the study of religion from different disciplines and points of view—such as history, theology, philosophy of religion, and sociology.

Study of individual and communal responses

The other sense of the subjectivity of religion is properly a matter for philosophy of religion and theology (Christian and otherwise). The study of religion can roughly be divided between descriptive and historical inquiries, on the one hand, and normative inquiries, on the other. The latter primarily concern the truth of religious claims, the acceptability of religious values, and other such normative aspects; the former, only indirectly involved with the normative elements of religion, are primarily concerned with its history, structure, and similar descriptive elements. The distinction, however, is not an absolute one, for, as has been noted, descriptions of religion may sometimes incorporate theories about religion that imply something about the truth or other normative aspects of some or all religions. Conversely, theological claims may imply something about the history of a religion. The dominant sense in which one speaks nowadays of the study of religion is the descriptive sense.

Neutrality in the study of religion. The attempt to be descriptive about religious beliefs and practices, without judging them to be valuable or otherwise, is often considered to involve *epochē*—that is, the suspension of belief and the "bracketing" of the phenomena under investigation. The idea of *epochē* is borrowed from the philosophy of the German thinker Edmund Husserl (1859–1938), the father of Phenomenology, and the procedure is regarded as central to the phenomenology of religion.

In this context the term phenomenology refers first to the attempt to describe religious phenomena in a way that brings out the beliefs and attitudes of the adherents of the religion under investigation, but without either endorsing or rejecting these beliefs and attitudes. Thus, the bracketing means forgetting about one's own beliefs that might endorse or conflict with what is being investigated. Second, phenomenology of religion refers to the attempt to devise a typology of religious phenomena—to classify religious activities, beliefs, and institutions.

Phenomenology as a method attempting objectivity

To some extent the emphasis on neutral description arises in modern times as a reaction against "committed" accounts of religion, which were for long the norm and still exist where religion is treated from a theological point of view. The Christian theologian, for example, may see a particular historical process as providential or as providing significance for Christian living. This is a legitimate perspective from the standpoint of faith. But the historical process itself has to be investigated in the first instance "scientifically"—that is, by considering the evidence, using the techniques of historical enquiry and other scientific methods. Conflict sometimes arises because the committed point of view is likely to begin from a more conservative stance—*i.e.,* to accept at face value the scriptural accounts of events—whereas the "secular" historian may be more skeptical, especially of records of miraculous events. The study of religion may thus come to have a reflexive effect on religion itself, such as the manner in which modern Christian theology has been profoundly affected by the whole question of the historicity of the New Testament.

The reflexive effect of the study of religion on religion itself may in practice make it more difficult for the student of religion to adopt the detachment implied by bracketing. Scholars generally agree, however, that the pursuit of objectivity is desirable, provided this does not involve sacrificing a sense of the inner aspect of religion. Thus, the stress on the distinction between the descriptive and normative approaches is becoming more frequent among scholars of religion.

The study of religion may thus be characterized as concerned with man's religious behaviour in relation to the transcendent, to God or the gods, and whatever else is regarded as sacred or holy, and as a study that attempts to be faithful both to the outer and inner facts. Its present-day concern is predominantly descriptive and explanatory and hence embraces such various disciplines as history, sociology, anthropology, psychology, and archaeology. Traditionally, however, the study has been more oriented toward truth claims in religion—these being properly the concern of theology and philosophy of religion. Needless to say, there are different sorts of theology, related to the different religious traditions, such as Christian, Muslim, and Buddhist. But insofar as the theologian expresses and articulates a tradition, he belongs to it and thus is part of the subject matter studied by the student of religion.

HISTORY OF THE STUDY OF RELIGION

No single history of the study of religion exists since the major cultural traditions (Europe, the Middle East, India,

China) have been mutually independent over long periods. The primary impulse that prompts many to study religion, however, happens to be the Western one, especially because other cultural traditions utilized categories other than that denoted in the Western concept of "religion." On the whole, in the ancient world and in the Middle Ages the various approaches to religion grew out of attempts to criticize or defend particular systems and to interpret religion in harmony with changes in knowledge. The same is true of part of the modern period, but increasingly the idea of the nonnormative (descriptive-explanatory) study of religions, and at the same time the attempt to understand the genesis and function of religion, has become established. Viewed thus, the 19th century is the formative period for the modern study of religion. The ensuing accounts here of the history of the subject take it up to the modern period and then consider the various disciplines connected with religion in detail—*i.e.,* in relation to their development since the 19th century.

Western orientation to the study of religion

The Greco-Roman period. *Early attempts to study religion.* One of the earliest attempts to systematize the seemingly conflicting Greek myths and thereby bring order into this rather chaotic Greek tradition was the *Theogony* of the Greek poet Hesiod (flourished *c.* 800 BC), who rather laboriously put together the genealogies of the gods. His work remains an important source book of ancient myth. The rise of speculative philosophy among the Ionian philosophers (*e.g.,* Thales, Heracleitus, and Anaximander) led to a more critical and, to some extent, rationalistic treatment of the gods. Thus, Thales (6th century BC) and Heracleitus (flourished *c.* 500 BC) considered water and fire, respectively, to be the first substance, out of which everything else is made, though Aristotle reported mysteriously in the 4th century BC that Thales believed that everything was filled with the gods. Anaximander (6th century BC) called the primary substance the infinite (*apeiron*). In these various schemes of religious belief, there is a unitary something that transcends the many clashing forces in the world, transcending even the gods. Heracleitus refers to the controlling principle as *logos,* or reason, though the philosopher, poet, and religious reformer Xenophanes (6th–5th centuries BC) directly assailed the traditional mythology as immoral, out of his concern to express a monotheistic religion. This theme of criticism of the myths was taken over and elaborated in the 4th century BC by the philosopher Plato. More conservatively, the poet Theagenes (6th century BC) allegorized the gods, treating them as standing for natural and psychological forces. To some extent, this line was pursued in the works of the Greek tragedians and by the philosophers Parmenides and Empedocles (5th century BC). Criticism of the ancient Greek tradition was reinforced by the reports of travellers as Greek culture penetrated widely into various other cultures. The Greek historian Herodotus (5th century BC) attempted to solve the problem of the plurality of cults by identifying foreign deities with Greek deities (*e.g.,* those of the Egyptian Amon with Zeus). This kind of syncretism was widely employed in the merging of Greek and Roman culture in the Roman Empire (*e.g.,* Zeus as the Roman god Jupiter).

Skepticism in the study of religion

The plurality of cults and gods also induced skepticism, as with the Sophist Protagoras (*c.* 481–411 BC), who was driven from Athens because of his impiety in questioning the existence of the gods. Prodicus of Ceos (5th century BC) gave a rationalistic explanation of the origin of deities that foreshadowed Euhemerism (see below *Later attempts to study religion*), and another Sophist, Critias (5th century BC), considered that religion was invented to frighten men into adhering to morality and justice. Plato was not averse to providing new myths to perform this alleged function—as is seen in his conception of the "noble lie" (*i.e.,* the invention of myths to promote morality and order) in *The Republic.* He was strongly critical, however, of the older poets' (*e.g.,* Homer's) accounts of the gods and substituted a form of belief in a single creator, the Demiurge or supreme craftsman. This line of thought was developed in a stronger way by Aristotle, in his conception of a supreme intelligence that is the unmoved mover. Aristotle combined elements of earlier thinking in his account of the genesis of the gods (coming from the observation of cosmic order and stellar beauty and from dreams).

Later attempts to study religion. Later Greek thinkers tended to vary between the positions adumbrated in the earlier period. The Stoics (philosophers of nature and morality) opted for a form of naturalistic monotheism, while the philosopher Epicurus (341–270 BC) was skeptical of religion as ordinarily understood and practiced, though he did not deny that there were gods who, however, had no transactions with men. Of considerable influence was Euhemerus (*c.* 330–*c.* 260 BC), who gave his name to the doctrine called Euhemerism, namely, that the gods are divinized men. Though Euhemerus' own argument was based largely upon fantasy, there are certainly some examples, both in Greek religion (*e.g.,* the god Heracles) and elsewhere, of the tendency to make men into gods, but it is obviously not universal.

Most of the Greek concepts about religion proved to be influential in the Roman world also. The atheistic Atomism of the Roman natural historian Lucretius (*c.* 95–55 BC) owed much to Epicurus. The eclectic thinker and politician Cicero (106–43 BC), in his *De natura deorum* ("Concerning the Nature of the Gods"), criticized Stoic, Epicurean, and later Platonic ideas about religion, but the book remains, however, incomplete. Much of the skepticism about the gods in the ancient world was concerned with the older traditional religions, whether of Greece or Rome. But in the early empire, the mystery cults, ranging from the Eleusinian mysteries of Greece to those of the Anatolian Cybele and the Persian Mithra, together with philosophically based religions such as Neoplatonism and Stoicism, had the greatest vitality. The patterns of religious belief were complex and of different levels, with various types of religion existing side by side. Into this situation Christianity was injected, and in its encounter with classical civilization it absorbed a number of the critiques of the gods of the older thinkers. In particular, Euhemerism was fashionable among the Church Fathers (the religious teachers of the early church) as an account of paganism. On the "pagan" side, there were persistent attempts to justify the popular cults and myths by the extensive use of allegory—a technique well adapted to the attempt to synthesize philosophical and popular religion. Christianity's own contribution to theories of the genesis of polytheism was through the doctrine of the fall of man, in which pure monotheism was believed to have become overlaid by demonic cults of the gods. Such an account could help to explain some underlying similarities between the Judeo-Christian tradition and certain aspects of Greco-Roman paganism. In this view there is the germ of an evolutionary account of religion. On the whole, however, the theories of religion in the ancient world were naturalistic and rationalist in emphasis.

The Middle Ages to the Reformation. *Theories of the Middle Ages.* The spread of Christianity into northern Europe and elsewhere beyond the confines of the Roman Empire presented similar problems to those encountered in the pagan world. Similar solutions were offered—*e.g.,* the identification of northern and Roman and Greek gods, sometimes using etymologies owing much to superficial resemblances of names. Thus, the Icelandic historian Snorri Sturluson (1179–1241) made use of this method in his handbook of Icelandic mythology—a work necessitated by the need to pass on the myth-laden Norse poetic lore that had survived the Christianization of the north—by adding to it Euhemeristic elements.

European contacts with other religions

Meanwhile, Islāmic theology had had an impact on Western Christianity, notably upon medieval Scholastic philosophy, in which the values of both reason and revelation were maintained. Muslim knowledge of other religions was in advance of European knowledge, notably in the work of the theologian Ibn Ḥazm (994–1064). Nevertheless, the reports of some European travellers, such as the Italian Marco Polo (1254?–?1324) and also Odoric of Pordenone (14th century), gave Westerners some knowledge of Asian religions. This opened the way toward a more inductive treatment of the phenomena of other religions, based on factual knowledge. Though most Christian, as well as Islāmic and Jewish, theologians tended to consider

the question of whether or not natural religion gives insight in God's nature—treating religion as a relation to the first cause of the universe—the English philosopher Roger Bacon (*c.* 1220–*c.* 1292) preferred to categorize the various manifest types of religion as a preliminary effort to establishing a true theology. Theorists of the medieval period continued to accept the thesis that polytheism had its origin in the Fall of man, but two new theories modified attitudes of Christians to other faiths. First, the theory arose that God adapts customs and rites having a pagan style in order to combat paganism itself—as a concession to the human condition. This theory could be used to explain the divergencies of practice within Christendom and to show points of contact between Christianity and paganism. Second, the doctrine of man's innate capacity to know God by reason enabled thinkers to discern some measure of truth in other religions. The questions raised by such theories were intensified during the Renaissance.

Theories of the Renaissance and Reformation. The Renaissance consisted in the invigoration of European culture through the rediscovery of the Greek and Roman classics, art, and architecture and thus was bound to set up tensions among Christians about paganism. The Italian Humanist Giovanni Boccaccio (1313–75) attempted to resolve these tensions in a medieval way by extensively allegorizing the ancient myths. The Dutch Humanist Erasmus (*c.* 1466–1536) and others, however, went further in stating that the ancient thinkers had a direct knowledge of the highest truth and sometimes in comparing them favourably with Scholastic theologians. One of the interlocutors in his *Convivium Religiosum* suggests that it would be better to lose the Scholastic theologian Duns Scotus than the ancient Roman thinkers Cicero or Plutarch, while another speaker restrains himself with difficulty from praying to the Greek philosopher Socrates (*c.* 470–399 BC) as though he were a Catholic saint. But a new turn to the arguments about idolatry, which were essentially apologetic, was given by the Protestant Reformers' attack on idolatry within the Roman Catholic Church and by their comparison between what they took to be the Christianity of the New Testament and the religion of Rome.

The need for a comparative study of religion

Thus, the need for a comparative treatment of religion became clear, and this prepared the way for more modern developments. Also preparatory for the modern study of religion was the new trend toward more or less systematic compilations of mythological and other material, stimulated partly by the Renaissance itself and partly by the discovery of America and other lands—conveying to the inhabitants of Europe a new perspective on the richness and variety of man's customs and histories. The most important figures in the exploration of the religions of the non-European world were the Spanish monk Bernardino de Sahagún (*c.* 1499–1590), who conscientiously gathered information in New Spain, J. Lafitau (1685–1740), a French missionary in Canada, and the Italian Jesuits Roberto De Nobili (1577–1656) and Matteo Ricci (1552–1610). The last two, who brought to bear a deep understanding of Indian and Chinese cultures, were unparalleled in that area of study until modern times. Thus, some of De Nobili's discussions with Brahmins were probably the first profound dialogues between Hindus and Christians. The inquiries of the 16th to 18th centuries thus initiated an accumulation of data about other cultures that stimulated studies of other men's religions and went beyond apologetic concerns, which hitherto had been dominant.

The beginnings of the modern period. *The late 17th and 18th centuries.* Attempts at a developmental account of religion were begun in the late 17th and 18th centuries. Notable was the scheme worked out, though not in great detail, by the Italian philosopher Giambattista Vico (1668–1774), who suggested that Greek religion passed through various stages: the divinization of nature, then of those powers that man had come to control (such as fire and crops), then of institutions (such as marriage), and finally the process of humanizing the gods, as in the works of Homer. The English philosopher David Hume (1711–76) gave another account in his *Natural History of Religion,* which reflected the growing Rationalism of the epoch. For Hume, original polytheism was the result of a naive anthropomorphism (conceiving the divine in human form) in the assignment of causes to natural events. The intensification of propitiatory and other forms of worship, he believed, led to the exaltation of one infinite divine Being. His "Essay upon Miracles" was also important in posing vital questions about the historical treatment of sacred texts, a set of problems that was to preoccupy 19th- and 20th-century Christian theologians.

The Rationalism of the period often involved a rejection both of paganism and dogmatic Christianity in the name of "natural religion." This natural religion, also called Deism, was the intellectual counterpart of the more emotional antidogmatic faith of the Pietists, who advocated "heart religion" over "head religion." Among the French Philosophes and Encyclopaedists, Voltaire (1694–1778) espoused an anticlerical Deism, which viewed the genesis of polytheism in the work of priests—a point also developed by another Encyclopaedist, Denis Diderot (1713–84). Voltaire was, incidentally, somewhat influenced and impressed by reports of the ethics of the Chinese social and religious sage Confucius (6th century BC).

The culmination of 18th-century Rationalism was found in the works of the German philosopher Immanuel Kant (1724–1804), but it was a rationalism modified to leave room for religion, which he based essentially on ethics. He held that all men in their awareness of the categorical imperative (*i.e.,* the notion that one must act as though what one does can become the universal law for mankind) and reverence for it share in the one religion and that the preeminence of Christianity lay in the conspicuous way in which Jesus enshrined the moral ideal. A series of reactions against the highly influential Kantian account paved the way for the various approaches to religion in the 19th century. In the meantime, the first beginnings of the development of Oriental studies and of ethnology and anthropology were making available more data about religion, though discussion in the 18th century continued, as in earlier centuries, the concern for the problems of religions other than those of the Judeo-Christian tradition largely in terms of the paganism of the ancient world. In this connection, the French scholar and politician Charles de Brosses (1709–77) attempted to explain Greek polytheism partly through the fetishism (belief in the magical powers of certain objects) found in West Africa. This foreshadowed later attempts to use comparative material in the elucidation of Greek myths. The French abbé Bergier (1718–90) explained primitive religions by means of a belief in spirits arising from a variety of psychological causes, which thus was a precursor of animism (a belief in souls in persons or certain natural objects).

Beginnings in the modern study of myth and evolutionary theories

One of the critics of Kant's view of religion was the German philosopher Johann Gottfried von Herder (1744–1803), who adopted an evolutionary account of the human race and who saw in mythology something much deeper and more significant for the understanding of human language and thought than a record of follies. His concern with symbolic thinking makes him the first modern student of myth. The German philosopher Friedrich Schelling (1775–1854) continued this positive approach, in the tradition of Romanticism. Furthermore, the advances in the knowledge of non-European, especially Indian, religion gave a wider perspective to discussions of the nature of religion, as was clear in the work of the German philosopher G.W.F. Hegel (1770–1831). The latter's self-confidence, in supposing that his philosophy represented the culmination of the history of philosophy, may amuse present-day scholars in view of the fact that many changes have occurred in philosophical enquiry since his day. Hegel was, nevertheless, immensely influential over a wide range of scholarship, including the study of religion. His followers were in large measure the founders of modern scientific history. Admittedly his theory of the historical dialectic—in which one movement (the thesis) is countered by another (the antithesis), both in interplay giving rise to a third (the synthesis), which now becomes the thesis of a new dialectical interplay, and so on—has been viewed as too artificial. But in providing a theoretical skeleton, it inspired attempts to make sense of the multitude of historical data, so that scholars were driven

to the investigation and discovery of particular facts that might exhibit the universal patterns postulated. Hegel also had a modified relativism, which implied that each phase of religion has a limited truth. This, together with his dialectic scheme, led to a general theory of religions, which though dated, much too neat, and based on imperfect information, nevertheless represents an important attempt at a comparative treatment, and one that was evolutionary or developmental.

The early 19th century. Hegel, as an Idealist, stressed the formative power of the spiritual on human history. By contrast, the French social philosopher Auguste Comte (1798–1857), from a Positivistic and Materialist point of view, devised a different evolutionary scheme in which there are three stages of human history: the theological, in which the supernatural is important; the metaphysical, in which the explanatory concepts become more abstract; and the positivistic—*i.e.,* the empirical. A rather different Positivism was expressed by the English philosopher Herbert Spencer (1820–1903), in which religion has a place beside science in attempting to refer to the unknown, and unknowable, Absolute. Evolutionary accounts were much boosted in the latter part of the 19th century by the new theory of biological evolution and had a marked effect both on history of religions and anthropology.

Meanwhile, the German philosopher Ludwig Feuerbach (1804–72) propounded, in his *Lectures on the Essence of Religion,* a view of religion as a projection of the aspirations of men, a thesis that was to be taken up in various ways by, among others, Marx, Freud, and Barth.

These various movements were supplemented by the growth of scientific history, archaeology, anthropology, and other sciences, which increased comparative knowledge of civilizations and cultures. The major figures and trends in the relevant disciplines are dealt with below.

Though the 19th-century theories that form the starting point of the modern study of religion were often based directly on metaphysical schemes in competition with Christian and other theologies, there was a notably different atmosphere in comparison with preceding periods, and the stage was set for a more complex and mutual attempt to understand the history and nature of religion.

BASIC AIMS AND METHODS

The growth of various disciplines in the 19th century, notably psychology and sociology, stimulated a more analytic approach to religions, while at the same time theology became more sophisticated and, in a sense, scientific as it began to be affected by and thus to make use of historical and other methods. The interrelations of the various disciplines in relation to religion as an area of study can be described as follows.

The various dimensions of religion

Religions, being complex, have different aspects or dimensions. Thus, the major world religions typically possess doctrines, myths, ethical and social teachings, rituals, social institutions, and inner experiences and sentiments. These dimensions lie behind the creation of buildings, art, music, and other such extensions of basic beliefs and attitudes. But not all religions are like Christianity and Buddhism, for example, in possessing institutions such as the church and the *saṅgha* (Buddhist monastic order), which exist across national and cultural boundaries. In opposition to such institutionalized religions, tribal religion, for example, is not usually separately institutionalized but in effect is the religious side of communal life and is not treated as distinct from other things that go on in the community.

The various dimensions of religion noted above represent a cross section of a tradition; but to see the latter in a well-balanced perspective it is necessary to view it as historical—as a religion having a past and the capacity for development in the future ("dead" religions, obviously enough, being the exception). Thus, there are various disciplines that may examine a religion cross-sectionally to find its basic patterns or structures. Psychology views religious experience and feelings and to some extent the myths and symbols that express experience; sociology and social anthropology view the institutions of religious tradition and their relationship to its beliefs and values; and literary and other studies seek to elicit the meanings of myths and other items. These structural enquiries sometimes benefit from being comparative—as when recurrent motifs in the doctrines of different religions are noticed. On the other hand, the aforementioned disciplines need to be supplemented by history, archaeology, philology, and other such disciplines, which have their own various methods of elucidating the past. Philosophy generally has attempted wide-ranging accounts of the nature of religion and of religious concepts, but it is not always easy to disentangle these enquiries from issues raised by normative theology.

Historical, archaeological, and literary studies. *Historical and literary studies.* The expansion of European empires in the early 19th century and the growth of scientific methods in history and philology combined to place Oriental and other non-European studies on a new basis. Another stimulus to the new approach to history and philology was Napoleon's expedition to Egypt, which was accompanied by scholars and scientists; it was a notable attempt to gather knowledge of a culture systematically. The discovery and editing of sacred and other texts from other cultures also had profound effects upon European thinking. A notable publishing venture was the series *Sacred Books of the East,* edited under the leadership of the German Orientalist and philologist Max Müller (1823–1900), which placed at the disposal of Westerners translations of the major literary sources of the non-Christian world. Earlier, Müller had published translations of the more important Vedic texts (Hindu sacred works), of which the Ṛgveda was given a complete scholarly edition in 1861–77. Interest in these ancient Indian texts was intense among Europeans and Americans in that earlier reports had suggested that these represented a world outlook from the "dawn of humanity" and that the origin of polytheism lay in nature worship. The Vedas, however, turned out to be of a very different character. The length of human history and prehistory, as implied by evoutionary theory and the growing archaeological discoveries, precluded looking upon the Vedic hymns as anything but late; though the contents showed them to be highly artificial and complex compilations for use in a priest-dominated ritual context, they were not at that time seen as spontaneous outpourings of the human spirit. Müller himself reacted rather sharply by adopting a different theory, which expressed his philological slant—namely, that polytheism was the result of a disease of language, in which the terms for natural phenomena came to be treated as having independent and personal reality; *nomina* ("names") became *numina* ("spirits"). The theory was in vogue for a time but was later replaced by more realistic insights drawn from anthropology. Furthermore, study of the greater part of the corpus of Indian sacred writings, including those in vernacular languages (especially Tamil), gradually modified the preoccupation with the earliest texts—the Vedic hymns and the *Upaniṣads* (philosophical treatises).

Discovery and publication of non-Western sacred writings

Throughout the development of the study of non-European languages there was a supposition that a non-Christian equivalent of the Bible could be found, a sacred writing that would thus provide the authoritative key to the beliefs, practices, and institutions of the religion under consideration. Gradually, however, it became apparent that sacred scriptures play very different roles in different religious cultures. Somewhat later in developing were studies of the Buddhist canon in Pāli (an ancient Indian language), which, through the work of such scholars as the English Orientalist T.W. Rhys Davids (1843–1922) and of the Pāli Text Society, which he founded, had a remarkable impact in revealing to the West the full range of Theravādin (southern Buddhist) religious literature; it tended to make Western scholars look upon the Theravāda as the earlier, "purer" form of Buddhism; but the editing of early Mahāyāna ("Greater Vehicle," or northern Buddhist) texts and the recognition of the different strata in the Pāli canon have modified this view. Buddhist studies were enhanced by the growth of Tibetan, Chinese, and Japanese studies. Some of the more important modern scholars of Zen Buddhism (a Mahāyāna sect) have been Japanese, notably the philosopher D.T. Suzuki (1870–1966), sometimes called

the apostle of Zen Buddhism to America, whose editions and interpretations have been widely influential.

The productivity of the study of religious literature of the late 19th century was immense, for it was not confined to the foregoing literary and archaeological activities but to the investigation of the Chinese Classics and the roots of Chinese civilization as well. Thus, by the early 20th century, Western scholars were in a position to study the main range of non-Western literary cultures. The wave of interest in these texts and the freeing of their dissemination from some of their traditional constraints (*e.g.,* the restriction of Vedic revelation to the upper classes of the Indian caste system) contributed to the revival of other religious cultures—notably Hinduism and Buddhism, under the stimulus of the Western challenge. Modern scholarship thus provided the basis for a new self-understanding among such religious traditions.

Interest in textual studies

Meanwhile, the texts of Zoroastrianism, an Iranian religion originating in the 6th century BC, were being discovered and edited (from 1850 onward). The disentangling of different layers of varying antiquity indicated the complex ways in which the religion of Zoroaster had developed.

During the latter part of the 19th and early part of the 20th century, there was a remarkable flowering of ancient Middle Eastern studies. Archaeology contributed to the unravelling of non-Jewish and Jewish religious history. The discovery of the *Epic of Gilgamesh,* a major work of Mesopotamian religious literature, and other materials brought a whole new perspective to the development of ideas in Mesopotamia; and in Egypt archaeological and papyrological studies brought to light the famous and revealing Egyptian funerary text, the Book of the Dead. These various ancient Middle Eastern discoveries have thrown light on the evolution of Judaism, and Semitic studies have likewise illuminated the origins and background of Islām. Furthermore, classical and European studies assembled data about the pre-Christian religions of the West so that scholars might gain a more detailed and scientific understanding of them. Compilations such as the *Corpus Inscriptionum Graecorum* and the *Corpus Inscriptionum Latinarum,* assembled in the 19th century, and the publication of Germanic, Celtic, and Scandinavian texts provided the tools for a reappraisal of these older traditions. Throughout the period intense researches into the composition and milieu of the Old and New Testaments reflected a new and "scientific" spirit of enquiry—which was, however, not without its controversial elements, sometimes because of the intimate tie between religious positions and evaluations of the Bible and sometimes because of the application of speculative patterns in the history of (non-Christian) religions to the New Testament. Meanwhile, the assemblage of materials extended forward into Christian history through the application of classical philological methods to patristic texts (the writing of the early Church Fathers) and to the corpus of Reformation writings.

The significance of 19th-century archaeological discoveries

Archaeological studies. The great archaeological discoveries of Heinrich Schliemann, the German excavator of Troy; the English archaeologists Arthur Evans in Crete and Wm. M. Flinders Petrie in Egypt; the French archaeologist Jacques de Morgan in Elam; the German Orientalist Hugo Winckler in Boğazköy (Anatolia); the French archaeologists Claude Schaeffer and C. Virolleaud in Ras Shamra (Ugarit); and other archaeologists greatly enlightened modern knowledge of the Greco-Roman and ancient Middle Eastern worlds. Biblical archaeology, culminating perhaps in the discovery of Masada, the Judaean hill fortress where the Jews made their last stand against the Romans in the revolt of AD 66–73 and that was mainly excavated in 1963, has given a new perspective to Old Testament, intertestamental, and later studies of ancient Judaism. The spectacular discovery by the English archaeologist John Marshall and others of the Indus Valley civilization pushed back knowledge of Indian prehistory to about 3500 BC and called into question the earlier theory of the primacy of Vedic culture in the formation of the Indian tradition, many features of which appear to have their first manifestation in the Indus Valley cities.

Archaeology made another profound impact on the study of religion when in 1841 the discovery of prehistoric human artifacts and later finds gave clues to early man's magico-religious beliefs and practices. These discoveries, notably the cave paintings in the Dordogne, northern and eastern Spain, and elsewhere, gave scholars encouragement to work out the course of man's religious evolution from earliest times. Spectacular as prehistoric archaeology was proving to be, however, it could only yield fragments of a whole that is difficult to reconstruct. Even the famous cave paintings of Les Trois Frères, in the Dordogne, for example, which portray among other things a dancing human with antlers on his head and a stallion's tail decorating his rear, does not yield an unambiguous interpretation: is the dancing figure a sorcerer, a priest, or what? He very likely is a priest presenting himself as a divine figure connected with animal fertility and hunting rites—but this remains as only an educated guess. Hence, it became attractive to many scholars of religion to try to supplement ancient archaeological evidence with data drawn from contemporary primitive peoples—*i.e.,* to interpret the prehistoric Stone Age through present-day stone age cultures. This procedure has several pitfalls—partly because contemporary "primitives" are themselves the product of a long historical process and because their culture may have changed over the millennia in many and various ways.

The work of the archaeologists has not merely stimulated new thinking about the early stages of religious history but it has also been a factor in drawing attention to the roles of buildings and art objects in religion. During the present century, spectacular religious monuments of the past, such as Angkor Wat (Cambodia), Borobudur (Indonesia), Ellora and Ajantā (India), and the Acropolis (Athens), have been officially preserved for scholarly and public viewing. Though iconography (the study of content and meaning in visual arts) has been better developed among art historians, students of religion are now paying increased attention to the religious decipherment of the visual arts. By contrast, very little has been done in the sphere of music, despite the considerable role it plays in so many religions. This is a further way in which the study of texts and ideas needs to be supplemented by knowledge of the milieu in which they have their meaning.

Anthropological approaches to the study of religion. *Theories concerning the origins of religion.* To draw a clear line between anthropology and sociology is difficult, and the two disciplines are divided more by tradition than by the scholarly methods they employ. Anthropology, however, has tended to be chiefly concerned with nonliterate and technologically primitive cultures and thus has stressed a certain range of techniques, such as the use of participant observation. Much anthropological investigation, however, has been carried out recently in more complex societies, such as in various Hindu areas of India, where there are different layers of society, ranging from an educated elite to illiterate workers who carry out the traditional menial tasks of the lowest castes and the outcastes. Because of the anthropologists' interest in tribal and "primitive" societies, it has not been unnatural for them to try to use the data gained in the study of such societies to speculate about the genesis and functions of religion.

An early attempt to combine archaeological evidence of prehistoric peoples, on the one hand, and anthropological evidence of primitive peoples, on the other, was that of the English anthropologist John Lubbock (1834–1913). His book, *The Origin of Civilization and the Primitive Condition of Man,* outlined an evolutionary scheme, beginning with atheism (the absence of religious ideas) and continuing with fetishism, nature worship, and totemism (a system of belief involving the relationship of specific animals to clans), shamanism (a system of belief centring on the shaman, a religious personage having curative and psychic powers), anthropomorphism, monotheism (belief in one god), and, finally, ethical monotheism. Lubbock recognized a point later made by the German theologian and philosopher Rudolf Otto (1869–1937) in distinguishing between the unique holiness (separateness) of God and his ethical characteristics. Unfortunately, much of his information was unreliable, and his schematism was open

Evolutionary theories about religion

to question; he foreshadowed, nevertheless, other forms of evolutionism, which were to become popular both in sociology and anthropology. The English ethnologist E.B. Tylor (1832–1917), who is commonly considered the father of modern anthropology, expounded, in his book *Primitive Culture,* the thesis that animism is the earliest and most basic religious form. Out of this evolves fetishism, belief in demons, polytheism, and, finally, monotheism, which derives from the exaltation of a great god, such as the sky god, in a polytheistic context. A somewhat similar system was advanced by Herbert Spencer (1820–1903) in his *Principles of Sociology,* though he stresses ancestor worship rather than animism as the basic consideration.

The classifications of religion—polytheism, henotheism (*i.e.,* the worship of one god as supreme without necessarily excluding the possibility of other groups' gods), and monotheism—begin from concern with gods and often imply the superiority of monotheism over other forms of belief. Naturally, the anthropologists of the 19th century were deeply influenced by the presuppositions of Western society.

The English anthropologist R.R. Marett (1866–1943), in contrast to Tylor, viewed what he termed animatism as of basic importance. He took his clue from such ideas as mana, mulungu, orenda, and so on (concepts found in the Pacific, Africa, and America, respectively), referring to a supernatural power (a kind of supernatural "electricity") that does not necessarily have the personal connotation of animistic entities and that becomes especially present in certain men, spirits, or natural objects. Marett criticized Tylor for an overly intellectual approach, as though primitive men used personal forces as explanatory hypotheses to account for dreams, natural events, and other phenomena. For Marett, primitive religion is "not so much thought out as danced out," and its primary emotional attitude is not so much fear as awe (in this he is close to Otto, whom he influenced).

Theories of Sir James Frazer

Another important figure in the development of theories of religion was the British folklorist Sir James Frazer (1854–1941), in whose major work, *The Golden Bough,* is set forth a mass of evidence to establish the thesis that men must have begun with magic and progressed to religion and from that to science. He owes much to Tylor but places magic in a phase anterior to belief in supernatural powers that have to be propitiated—this belief being the core of religion. Because of the realization that magical rituals do not in fact work, primitive man then turns, according to Frazer, to reliance on supernatural beings outside his control, beings who need to be treated well if they are to cooperate with human purposes. With further scientific discoveries and theories, such as the mechanistic view of the operation of the universe, religious explanations gave way to scientific ones. Frazer's scheme is reminiscent of that of the French "father of sociology," Auguste Comte.

These and other evolutionary schemes came in for criticism, however, in the light of certain facts about the religions of primitive peoples. Thus, the Scottish folklorist Andrew Lang (1844–1912) discovered from anthropological reports that various primitive tribes believed in a high god—a creator and often legislator of the moral order. Marett and other anthropologists contended that Lang's attempt to argue for an *Urmonotheismus* (primordial monotheism) was contrary both to evolutionary ideas and to the established view of the lack of sophistication and half-animal status of the so-called savage. Since Lang was more of a brilliant journalist than an anthropologist, his view was not taken with as much seriousness as it should have been.

The German Roman Catholic priest and ethnologist Wilhelm Schmidt (1868–1954), however, brought anthropological expertise to bear in a series of investigations of such primitive societies as those of the Tierra del Fuegians (South America), the Negrillos of Rwanda (Africa), and the Andaman Islanders (Indian Ocean). The results were assembled in his *Der Ursprung der Gottesidee* ("The Origin of the Idea of God"), which appeared in 12 volumes from 1912 to 1955. Not surprisingly, Schmidt and his collaborators saw in the high gods, for whose cultural existence they produced ample evidence from a wide variety of unconnected societies, a sign of a primordial monotheistic revelation that later became overlaid with other elements (this was an echo of earlier Christian theories invoking the Fall to similar effect). The interpretation is controversial, but at least Lang and Schmidt produced grounds for rejecting the earlier rather naïve theory of evolutionism.

Theories of primordial monotheism

Modern scholars do not, on the whole, accept Schmidt's scheme. Some, such as the Italian anthropologist Raffaele Pettazzoni (1883–1959), have stressed merely that a sky god has a certain natural preeminence; others emphasize that the high god is often a *deus otiosus* ("idle god")—*i.e.,* not active in the world and hence not the recipient of a functioning cult. In any event, it is a very long jump from the premise that primitive tribes have high gods to the conclusion that the earliest men were monotheists.

Others who have looked at religions from an anthropological point of view have emphasized the importance, in a number of cultures, of the mother goddess (as distinct from the male sky god). A pioneer work in this direction was that of the Swiss anthropologist and jurist J.J. Bachofen (1815–87), whose *Das Mutterrecht* ("The Mother Right") unravelled some puzzles in ancient law, mythology, and art in terms of a matriarchal society.

Functional and structural studies of religion. The search for a tidy account of the genesis of religion in prehistory by reference to primitive societies was hardly likely to yield decisive results. Thus, anthropologists became more concerned with functional and structural accounts of religion in society and relinquished the apparently futile search for origins.

Notable among these accounts was the theory of the French sociologist Émile Durkheim (1858–1917). According to Durkheim, totemism was fundamentally significant (he wrongly supposed it to be virtually universal), and in this he shared the view of some other 19th-century savants, notably Salomon Reinach (1858–1932) and Robertson Smith (1846–94), not to mention Sigmund Freud (1856–1939). Because Durkheim treated the totem as symbolic of the god, he inferred that the god is a personification of the clan. This conclusion, if generalized, suggested that all the objects of religious worship symbolize social relationships and, indeed, play an important role in the continuance of the social group.

Various forms of functionalism in anthropology—which understood social patterns and institutions in terms of their function in the larger cultural context—proved illuminating for religion, such as in the stimulus to discover interrelations between differing aspects of religion. The Polish-British anthropologist Bronisław Malinowski (1884–1942), for instance, emphasized in his work on the Trobriand Islanders (New Guinea) the close relationship between myth and ritual—a point also made emphatically by the "myth and ritual" school of the history of religions (see below *Other studies and emphases*). Furthermore, many anthropologists, notably Paul Radin (1883–1959), moved away from earlier categorizations of so-called primitive thought and pointed to the crucial role of creative individuals in the process of mythmaking.

Analyses of social patterns and institutions

A rather different approach to myths was made by the 20th-century French anthropologist Claude Lévi-Strauss, whose rather formalistic structuralism tended to reinforce analogies between "primitive" and sophisticated thinking and also provided a new method of analyzing myths and stories. His views had wide influence, though they are by no means universally accepted by anthropologists.

Specialized studies. The impact of Western culture, including missionary Christianity, and technology upon a wide variety of primitive and tribal societies has had profound effects and represents a specialized area of study closely related to religious anthropology. One pioneering work is *Religions of the Oppressed* by the Italian anthropologist and historian of religion Vittorio Lanternari. What is striking is the way in which similar types of reaction, creating new religious movements, occur at different points across the world. There are, thus, many possibilities of a comparative treatment.

Among a number of contemporary anthropologists, including the American Clifford Geertz, there is a concern,

after a period of functionalism, with exploring more deeply and concretely the symbolism of cultures. The English social anthropologist E.E. Evans-Pritchard (1902–73), noted among other things for his work on the religion of Nuer people (who live in The Sudan), produced in his *Theories of Primitive Religion* a penetrating critique of many of the earlier anthropological stances. Though it has always been difficult to confirm theories in view of the complexity of the data, a statistical approach has been attempted—*e.g.*, by G. Swanson in his *Birth of the Gods*, which attempts to exhibit correlations between types of social arrangement and religious beliefs, such as the caste system and belief in reincarnation.

Because of the nature of the societies that typically have come under the scrutiny of anthropology, the discipline has necessarily had to come to terms with religion. In terms of the methods used, the anthropological approach is of considerable interest to historians of religion and is a corrective to overintellectual, text-based accounts of religions. Also, the present concerns for comparative studies and symbolic analysis coincide with existing concerns in the phenomenology of religion (see below *History and phenomenology of religion*).

Sociological studies of religion. *Theories of stages.* Auguste Comte (1798–1857) is usually considered the founder of modern sociology. His general theory hinged substantially on a particular view of religion, and this view has somewhat influenced the sociology of religion since that time. In his *Cours de philosophie positive* (*The Positive Philosophy of Auguste Comte*) Comte expounded a naturalistic Positivism and sketched out the following stages in the evolution of thought. First, there is what he called the theological stage, in which events are explained by reference to supernatural beings; next, there is the metaphysical stage, in which more abstract unseen forces are invoked; finally, in the positivistic stage, men seek causes in a scientific and practical manner. To seek for scientific laws governing human morality and society is as necessary, in this view, as to search for those in physics and biology—hence Comte's role in advocating a science of society, namely sociology. Among the leading figures in the development of sociological theories were Spencer and Durkheim (see above *Anthropological approaches to the study of religion*).

Marxist theories of religion

A rather separate tradition was created by the German economic theorist Karl Marx (1818–83). A number of Marxists, notably Lenin (1870–1924) and K. Kautsky (1854–1938), have developed social interpretations of religion based on the theory of the class struggle. Whereas sociological functionalists posited the existence in a society of some religion or a substitute for it (Comte, incidentally, propounded a positivistic religion, somewhat in the spirit of the French Revolution), the Marxists implied the disappearance of religion in a classless society. Thus, in their view religion in man's primordial communist condition, at the dawn of the historical dialectic, reflects ignorance of natural causes, which are explained animistically. The formation of classes leads, through alienation, to a projection of the need for liberation from this world into the transcendental or heavenly sphere. Religion, both consciously and unconsciously, thus becomes an instrument of exploitation. In the words of the young Marx, religion is "the generalized theory of the world . . . , its logic in popular form." The modern intellectualist accounts of religion, tending to ignore the rituals, experiences, and institutions but concentrating rather on the doctrines and myths, have proved something of a problem for later Marxist applications of their theory. Since the theory was a product of a rather early and unsophisticated stage of theorizing about religion, it was not adapted particularly well to deal with other cultures—hence a considerable debate in modern China on the status of Chinese religion in the light of Marxism, some holding that Marx's critique did not, for example, fit Buddhism.

Comparative studies. One of the most influential theoreticians of the sociology of religion was the German scholar Max Weber (1864–1920). He observed that there is an apparent connection between Protestantism and the rise of capitalism, and in *The Protestant Ethic and the Spirit of Capitalism* he accounted for the connection in terms of Calvinism's inculcating a this-worldly asceticism—which created a rational discipline and work ethic, together with a drive to accumulate savings that could be used for further investment. Weber noted, however, that such a thesis ought to be tested; and a major contribution of his thinking was his systematic exploration of other cultural traditions from a sociological point of view. He wrote influentially about Islām, Judaism, and Indian and Chinese religions and, in so doing, elaborated a set of categories, such as types of prophecy, the idea of charisma (spiritual power), routinization, and other categories, which became tools to deal with the comparative material; he was thus the real founder of comparative sociology. Because of his special interest in religion, he can also be reckoned a major figure in the comparative study of religion (though he is not usually reckoned so in most accounts of the history of religions). Though he made significant contributions to the study of religion, his judgments on Indian and other religions are not all or mostly accepted now—since he necessarily based his views on secondary sources—and some of his categorial distinctions are open to debate, such as his rather broad use of the category of prophet.

Theories of Max Weber

Weber's comparative method in the scientific sociology of religion introduced an analogue to experimentation (*i.e.*, looking at similar patterns in independent cultures with varying contextual conditions). Since the 1950s there has been considerable emphasis on statistical methods, side by side with the more theoretical discussions arising from classical sociology. Typical of the trend is the American sociologist Gerhard Lenski's *Religious Factor*, which delineates the relations between religious allegiance and other factors in a large city in the United States.

Other sociological studies. An extensive literature on religious sects and similar groups has also developed. To some extent this has been influenced by the German theologian Ernst Troeltsch in his distinction between church and sect (see below *Theological studies*). Notable among modern investigators of sectarianism is the British scholar Bryan Wilson. Church organizations also have attempted to use the insights of sociology in the work of evangelism and other church-related activities—a use of the discipline that is sometimes called "religious sociology" to distinguish it from the more theoretical and "objective" sociology of religion.

Coordination between sociology and the history of religions is not usually very close, since the two disciplines operate as separate departments in most universities and often in different faculties. From the sociological end, Weber represents one kind of synthesis; from the history-of-religions end, the writings of the German-American scholar Joachim Wach (see below *The "Chicago school"*) were quite influential. In his book *Sociology of Religion* he attempted to exhibit the ways in which the community institutions of religion express certain attitudes and experiences. This view was in accordance with his insistence on the practical and existential side of religion, over against the intellectualist tendency to treat the correlate of the group as being a system of beliefs.

Recent sociological studies

Among the more recent theorists of the sociology of religion is the influential and eclectic American scholar Peter Berger. In *The Sacred Canopy* he draws on elements from Marx, Durkheim, Weber, and others, creating a lively theoretical synthesis. One problem is raised by his method, however; he espouses what he calls "methodological atheism" in his work, which appears to presuppose a view about religion. Despite Berger's sympathy in dealing with religious phenomena, the methodological stance adopted in this book seems to imply a reductionist position—namely, one in which religious beliefs are explained by reference to basically nonreligious sentiments, sociopsychological circumstances, and other factors. In itself, this is a theory having possibilities, for the study of religion cannot rule out a priori the thesis that religion is a projection—*e.g.*, that it rests upon an illusion—or other such theses; but the question arises as to whether or not the methods espoused in the scientific study of religion have already secretly prejudged the issue.

On the whole, modern sociology is largely geared to dealing with Western religious institutions and practices, though some notable work has been done, especially since World War II, in Asian sociology of religion. Emphasis has been placed upon the process of secularization in a number of Western sociological studies (which have had some impact on the formation of modern Christian theology), notably in *The Secular City* of the American theologian Harvey Cox. There are indications that the process of secularization does not occur in the same degree or occurs in a different manner in non-Western cultures.

In general, the main question of the sociology of religion concerns the effectiveness with which it can relate to other studies of religion. This question is posed in *The Scientific Study of Religion,* by the American sociologist J. Milton Yinger. A similar tendency is noted in the synthesis between the history and the sociology of religion in a new-style evolutionism propounded by another American scholar, Robert Bellah.

The psychology of religion. The study of religious psychology involves both the gathering and classification of data and the building and testing of various (usually rather wide-ranging) explanations. The former activity overlaps with the phenomenology of religion, so it is to some extent an arbitrary decision under which head one should include descriptive studies of religious experience and related subjects.

Psychological studies. Notable among investigations by psychologists was *The Varieties of Religious Experience,* by the American philosopher and psychologist William James (1842–1910), in which he attempted to account for experiences such as conversion through the concept of invasions from the unconscious. Because of the clarity of his style and his philosophical distinction, the work has had a lasting influence, though it is dated in a number of ways and his examples come from a relatively narrow selection of individuals, largely within the ambit of Protestant Christianity. This points to a recurring problem—that of relating individual psychology to the institutions and symbols of different cultures and traditions.

Investigations of religious experience

More radical, but drawing from a rather larger range of examples, was the American psychologist J.H. Leuba (1868–1946). In *A Psychological Study of Religion* he attempted to account for mystical experience psychologically and physiologically, pointing to analogies with certain drug-induced experiences. Leuba argued forcibly for a naturalistic treatment of religion, which he considered to be necessary if religious psychology was to be looked at scientifically. Others, however, have argued that psychology is in principle neutral, neither confirming nor ruling out belief in the transcendent. Most scholars would, however, consider the problem to be a complex philosophical one, which goes beyond psychology as such.

Among those who have attempted a fairly detailed classification of mystical experience, but not necessarily from a scientific-psychological point of view, mention should be made of the English scholar Evelyn Underhill (1875–1941), drawing on examples from the Jewish, Christian, and Islāmic traditions. Recently, systematic explorations (taking into account Eastern mysticism as well) have been undertaken. Rudolf Otto was important in elucidating the nature of numinous experience, and there has also been a certain amount of scholarly work performed in the description and classification of types of shamanism, spirit possession, and similar phenomena.

Psychoanalytical studies. More influential than James and Leuba and others in that tradition were the psychoanalysts. Freud gave explanations of the genesis of religion in various of his writings. In *Totem and Taboo* he applied the idea of the Oedipus complex (involving unresolved sexual feelings of, for example, a son toward his mother and hostility toward his father) and postulated its emergence in the primordial stage of human development. This stage he conceived to be one in which there were small groups, each dominated by a father. According to Freud's reconstruction of primordial society, the father is displaced by a son (probably violently), and further attempts to displace the new leader bring about a truce in which incest taboos (proscriptions against intrafamily sexual relations) are formed. The slaying of a suitable animal, symbolic of the deposed and dead father, connected totemism with taboo. In *Moses and Monotheism* Freud reconstructed biblical history in accord with his general theory, but biblical scholars and historians would not accept his account since it was in opposition to the point of view of the accepted criteria of historical evidence. His ideas were also developed in *The Future of an Illusion.* Freud's view of the idea of God as being a version of the father image and his thesis that religious belief is at bottom infantile and neurotic do not depend upon the speculative accounts of prehistory and biblical history with which Freud dressed up his version of the origin and nature of religion. The theory can still stand as an account of the way in which religion operates in individual psychology, though of course it has also attracted criticism on grounds other than historical ones (*e.g.,* Buddhism does not have a father figure to worship).

Theories of Freud and Jung

A considerable literature has developed around the relationship of psychoanalysis and religion. Some argue, despite the atheistic mood of Freud's writing and his critique of religious belief, that the main theory is compatible with faith—on the grounds, for instance, that the theory describes certain mechanisms operative in people's religious psychology that represent modes in which people respond to the challenge of religious truth. Even if this position can be sustained, it is clear, nevertheless, that acceptance of Freudian insights makes a considerable difference to the way in which religious experience and behaviour are viewed. Questions have arisen about the range of applicability of Freud's ideas—*e.g.,* whether or not his theories apply outside the Western milieu, such as in Theravāda Buddhism, which does not possess a father figure or worship a god. Various attempts have been made to test Freud's theory of religion empirically, but the results have been ambiguous.

The Swiss psychoanalyst C.G. Jung (1875–1961) adopted a very different posture, one that was more sympathetic to religion and more concerned with a positive appreciation of religious symbolism. Jung considered the question of the existence of God to be unanswerable by the psychologist and adopted a kind of agnosticism. Yet he considered the spiritual realm to possess a psychological reality that cannot be explained away, and certainly not in the manner suggested by Freud. Jung postulated, in addition to the personal unconscious (roughly as in Freud), the collective unconscious, which is the repository of human experience and which contains "archetypes" (*i.e.,* basic images that are universal in that they recur in independent cultures). The irruption of these images from the unconscious into the realm of consciousness he viewed as the basis of religious experience and often of artistic creativity. Religion can thus help men, who stand in need of the mysterious and symbolic, in the process of individuation—of becoming individual selves. Some of Jung's writings have been devoted to elucidating some of the archetypal symbols, and his work in comparative mythology, the history of alchemy, and other similar areas of concern has proved greatly influential in stimulating the investigations of other interested scholars. Thus, the Eranos circle, a group of scholars meeting around the leadership of Jung, contributed considerably to the history of religions. Associated with this circle of scholars have been Mircea Eliade, the eminent Romanian-French historian of religion, and the Hungarian-Swiss historian of religion Károly Kerényi (1897–1973). This movement has been one of the main factors in the modern revival of interest in the analysis of myth.

Among other psychoanalytic interpreters of religion, the American scholar Erich Fromm (1900–80) modified Freudian theory and produced a more complex account of the functions of religion. Part of the modification is viewing the Oedipus complex as based not so much on sexuality as on a "much more profound desire"—namely, the childish desire to remain attached to protecting figures. The right religion, in Fromm's estimation, can, in principle, foster an individual's highest potentialities, but religion in practice tends to relapse into being neurotic.

Authoritarian religion, according to Freud, is dysfunctional and alienates man from himself.

Other studies. Apart from Jung's work, there have been various attempts to relate psychoanalytic theory to comparative material. Thus, the English anthropologist Meyer Fortes, in his *Oedipus and Job in West African Religion,* combined elements from Freud and Durkheim, and G.M. Carstairs (a British psychologist), in *The Twice Born,* investigated in depth the inhabitants of an Indian town from a psychoanalytic point of view and with special reference to their religious beliefs and practices. Among the more systematic attempts to evaluate the evidences of the various theories is *Religious Behaviour,* by Michael Argyle, another British psychologist.

A certain amount of empirical work in relation to the effects of meditation and mystical experience—and also in relation to drug-induced "higher" states of consciousness—has also been carried on. Investigation of religious responses as correlated with various personality types is another area of enquiry; and developmental psychology of religion, largely under the influence of the French psychologist Jean Piaget (1896–1980), has played a prominent part in educational theory in the teaching of religion. Most scholars agree, however, that more needs to be done to make results in the psychology of religion more precise; and also, for reasons that are unclear, very few people recently have concerned themselves with the field, which thus is in a state of suspension after a flurry of activity in the late 19th and early 20th centuries.

Trends in the philosophy of religion

Philosophy of religion. *The concerns of the philosophy of religion.* The scope of the philosophy of religion has changed somewhat in the last century and a half—that is, in the time since it came to be recognized as a separate branch of philosophy. Its nature is, as is typically the case in philosophy, open to debate. Three main trends, however, can be noted: (1) the attempt to analyze and describe the nature of religion in the framework of a general view of the world; (2) the effort to defend or attack various religious positions in terms of philosophy; and (3) the attempt to analyze religious language. Philosophical materials are also often incorporated into theologies—a modern example being the use of Existentialism in the theology of Rudolf Bultmann, the German New Testament scholar (see below *Neo-orthodoxy and demythologization*), and others; an older example is the medieval theologian Thomas Aquinas' use of Aristotle and of his (Aquinas' own) insights in the service of a systematic Christian theology. The different activities mentioned above overlap substantially. The second of them is usually taken to include the exploration of natural theology (*i.e.,* the truths about God that can be known, as it is claimed, by the aid of reasoning and insight, independently of the truths vouchsafed by revelation). Metaphysical systems (concerning the nature of reality) sometimes function as analogues to natural theology and thus provide a kind of support for a revealed religious belief system. Thus, much of philosophy of religion is concerned with questions not so much of the description of religion (historically and otherwise) as with the truth of religious claims. For this reason philosophy can easily become an adjunct of theology or of antireligious positions. To this extent, philosophy lies outside the main disciplines concerned with the descriptive study of religion; thus, it is often difficult to disentangle descriptive problems from those bearing on the truth of the content of what is being described. Feuerbach's "projection" theory of religion, for example, possessed a metaphysical framework, but it also included empirical claims about the nature of religion. The following brief account of philosophical trends is necessarily selective, leaning toward those philosophical theories that have a stronger content of, or relevance to, descriptive claims about religion.

Theories of Schleiermacher and Hegel. Immanuel Kant's powerful critique of traditional natural theology appeared to rob religion of its basis in reason and to make it an adjunct to morality. But Kant's system depended on drawing certain distinctions, such as that between pure and practical reason, which were open to challenge. One reaction that attempted to place religion in a more realistic position (*i.e.,* as neither primarily to do with pure nor with practical reason) was that of the German theologian and philosopher Friedrich Schleiermacher (1768–1834) in his *On Religion: Speeches to Its Cultured Despisers.* He attempted there to carve out a separate territory for religious experience, as distinct from both science and morality. For him the central attitude in religion is "the feeling of absolute dependence." In drawing attention to the affective and experiential side of religion, usually neglected in preceding philosophical discussions, Schleiermacher set in motion the modern concern to explore the subjective or inner aspect of religion. Schleiermacher's main goal, however, was not the exploration of religion as such but rather the construction of a new type of theology—the "theology of consciousness." In so doing he relegated doctrines to a secondary role, their function being to express and articulate the deliverances of religious consciousness. Thus, incidentally, it became important for New Testament historians who were influenced by Schleiermacher to penetrate the religious consciousness of Jesus—this becoming, in effect, the reputed locus of his divinity.

Philosophical approach to religious experience

G.W.F. Hegel had, as noted above, a profound effect upon the development of historical and other studies. His own system, the system of the Absolute, contained a view of the place of religion in human life. According to this notion, religion arises as the relation between man and the Absolute (the spiritual reality that undergirds and includes the whole universe), in which the truth is expressed symbolically, and so conveyed personally and emotionally to the individual. As the same truth is known at a higher—that is, more abstract—level in philosophy, religion is, for all its importance, ultimately inferior to philosophy. The relationship between abstract and concrete truth was, incidentally, taken up in the 19th-century Hindu renascence as a parallel to the doctrine of the Absolute—the Advaita (nondualism), the dominant expression of Hindu metaphysics—held by the 8th-century Hindu philosopher Śaṅkara. The Hegelian account of religion was worked out in the context of the dialectical view of history, according to which opposites united in a synthesis, which in turn produced its opposite, and so on. Hegel was influential in the interpretation of Christian history: Jesus as thesis, Paul as antithesis, and early Catholicism as the synthesis, the latter becoming a new thesis that would elicit a new antithesis, Protestantism.

Hegel attracted some radical criticism, however. One such was that of the aforementioned German philosopher Ludwig Feuerbach (1804–72), whose ideas have been sketched above. Another was that of the Danish philosopher and theologian Søren Kierkegaard (1813–55), sometimes regarded as the father of modern Existentialism, who reacted against the metaphysical and "rational" approach to Christianity in Hegel's thought. Kierkegaard's penetrating psychological insights were put to the service of philosophy and theology and threw new light on the nature of religious experience and its relation to features of man's inner life, such as dread and despair. Kierkegaard's main concern, however, was prophetic rather than descriptive. From a very different standpoint (*i.e.,* that of liberal Protestantism), the German theologian Albrecht Ritschl (1822–89) made an apologetic defense of Christianity in his attempt to analyze theological utterances as essentially affirming value judgments.

Schleiermacher's delineation of religious experience was complemented by attempts among the Romantics and by the German philosopher Ernst Cassirer (1874–1945) to exhibit the nature of symbolic thinking and in particular the special character of religious symbolism. This was some distance from the rationalism of Kant, though Cassirer was nevertheless influenced by the Neo-Kantian tradition.

Empiricism and Pragmatism. The Hegelian school, very influential in the 19th century, entered a period of rapid decline in the early part of the 20th. The common sense and scientifically oriented philosophy of the English scholars G.E. Moore (1873–1958) and Bertrand Russell (1872–1970) introduced a period of Empiricism in Britain, while William James's Pragmatism had a similar effect in America. Theologically, there was an antimetaphysical revolution during and after World War I. On the conti-

Influence of Existentialism and Linguistic Analysis

nent of Europe, the increasing influence of Existentialism was hostile to the old type of metaphysics. British Empiricism was expressed very strongly in Logical Positivism (maintaining the exclusive value of scientific knowledge and the denial of traditional metaphysical doctrines) and its linguistic aftermath. This stimulated the analysis of religious language, and the movement was complicated by the transformation in the thought of the Austrian-English philosopher Ludwig Wittgenstein (1889–1951), who in his later thought was very far removed from his early, rather formalistic treatment of language.

Theoretically, the Analytic attempt to exhibit the nature of religious language could have been a chiefly descriptive task, but, in fact, most analyses have occurred in the context of questions of truth—thus some scholars have been concerned with exhibiting how it is possible to hold religious beliefs in an Empiricist framework, and others with showing the meaninglessness or incoherence of belief. A landmark was the publication, in 1955, of *New Essays in Philosophical Theology,* edited by the English philosophers A.G.N. Flew and A. MacIntyre. Though Wittgenstein stressed the idea of "forms of life," according to which the meaning of religious beliefs would have to be given a practical and living contextualization, little has been done to pursue the idea empirically. The discovery by the English philosopher J.L. Austin (1911–60) and others of performative uses of language has stimulated some enquiry in this direction. On the whole, however, the Analytic philosophy of religion has been pursued rather independently of the descriptive study and history of religion.

Modern Existentialist and Phenomenological studies. Since linguistic philosophy tends to be considered by its proponents to be a method or a group of methods, internal diversity within the area of concern is not surprising. Similarly, Existentialism, which is less of an "-ism" than an attitude, expresses itself in a variety of ways. The most influential modern Existentialists have been the German philosopher Martin Heidegger (1889–1976) and the French philosopher, dramatist, and novelist Jean-Paul Sartre (1905–80); the former was especially important in the development of modern continental theology, particularly for the use made of some of his ideas by Rudolf Bultmann.

Influence of Heidegger and Sartre

According to Heidegger, man's existence is characterized as "care." This care is shown first in possibility: man makes things instrumental to his concerns and so projects forward. Secondly, there is his facticity, for he exists as a finite entity with particular limitations (his "thrownness"). Thirdly, man seeks to avoid the anxiety of his limitations and thus seeks inauthentic existence. Authenticity, on the other hand, involves a kind of stoicism (positive attitude toward life and suffering) in which death is taken up as a possibility and man faces the "nothing." The structure of man's world as analyzed by Heidegger is revealed, in a sense, affectively—*i.e.,* through care, anxiety, and other existential attitudes and feelings.

Sartre's thought has had less direct impact on the study of religion, partly because his account of human existence represents an explicit alternative to traditional religious belief. Sartre's analysis begins, however, from the human desire to be God: but God is, on Sartre's analysis, a self-contradictory notion, for nothing can contain the ground of its own being. In searching for an essence man fails to see the nature of his freedom, which is to go beyond definitions, whether laid down by God or by other human beings.

The French philosopher Gabriel Marcel (1889–1973) is not individualistic like Sartre (or at least the early Sartre, whose thinking was modified by Marxism); instead, he stresses the communal character of human existence—the highest virtue being fidelity. Marcel also emphasizes the mysterious (as distinguished from the empirically problematic) character of love, evil, hope, freedom, and, above all, being. His work provides a rich analysis and interpretation of the religious dimensions of human experience and thus is a philosophical basis for the study of religious experience.

The Existentialist approach attempts to describe and evoke the way human beings are and thus can lay claim to be phenomenological. It is clear, however, from the divergencies among Existentialists, that they contain speculative and idiosyncratic elements, and one question raised about the general applicability of their characterizations is how far they are bounded by the product of a particular mood in Western culture.

Influence of Phenomenology

The German philosopher Edmund Husserl (1859–1938) has had, as the main exponent of Phenomenology, a wide effect on the study of religion. His program of describing experience and "bracketing" the objects of experience, in the pursuit of essences of types of experience, was in part taken up in the phenomenology of religion. Husserl distinguished Phenomenology from psychology, however, because, in his view, the latter concerns facts in a spatio-temporal setting, whereas Phenomenology uncovers timeless essences. This aspect of Husserl's thinking has not always or wholly been accepted by phenomenologists of religion, who have been much more oriented toward facts, though Husserl's emphasis on essences often has tended to make religious phenomenology lean toward a static typology.

Relationship between Western and non-Western philosophy in regard to religion. Western philosophy has thus had a significant influence on the study of religion. It has also come into contact with non-Western traditions and has thus stimulated concern with the problem of the nature of religious truth in a world perspective. The most influential product of this interplay has most likely been the neo-Advaitin philosophy (a new version of Advaita, or nonduality) espoused by a number of modern Indians, such as Swami Vivekananda (1863–1902), who made a sensational appearance at the Parliament of Religions in Chicago in 1893, and the Indian philosopher Sarvepalli Radhakrishnan (1888–1975). Both of these thinkers attempted to reveal the underlying unity in the great religions—a unity described from a point of view drawing on the thought of Śaṅkara.

The idea of a world faith

The U.S. philosopher William Ernest Hocking (1873–1966) pursued similar interests in the construction of a world faith that he considered might come about through the mutual modification of, and interchange between, the great religious traditions. These concerns have raised important questions about the criteria of truth between religions, the tests of whether one religion is truer than others, and the extent to which valid identifications of belief can be made between one faith and another. The various elements of the philosophical traditions of the last two centuries have thus had a bearing on religious questions, and most scholars consider that though the philosophy of religion tends to be normative rather than descriptive, it is a necessary adjunct to descriptive studies. Philosophical insights and expertise are of significant relevance to the numerous questions of method that arise in the study of religion. (See also PHILOSOPHIES OF THE BRANCHES OF KNOWLEDGE: *Philosophy of religion.*)

Theological studies. *Historical-critical studies.* The major feature in the development of Christian theology during the 19th and 20th centuries has been the impact of historical enquiry on the biblical sources of belief (there has also been a similar effect on Jewish and other theologies, but Christian theology has been the most influential in the development of Western culture). A pioneer in the attempt to understand the mythological elements in the New Testament was the German theologian David F. Strauss (1808–74), whose controversial *Life of Jesus* (published in German, 1835–36) was an attempt to sift out the historical Jesus from the overlay of myth created by the poetic imagination of the early church. Similarly, the German church historian Adolf von Harnack (1851–1930), influenced by Albrecht Ritschl, intended to penetrate the accretions of dogma attached to the historical Jesus. Such attempts were later to come under radical criticism from, among others, the Alsatian philosopher-theologian and Nobel laureate Albert Schweitzer (1875–1965) for describing the alleged Jesus of history in terms tailored to fit the presuppositions of liberal Protestantism. Thus was raised an important methodological question on how to deal with such material as the Gospels.

Important in trying to spell out principles for dealing

Principles of historical criticism

with the material was Ernst Troeltsch, who argued that history has to be written in accordance with the following principles: first, the principle of criticism—*i.e.,* the sifting of the evidences and testing of conclusions (thus historical certainty about much in the ancient witnesses to Jesus is impossible); second, the principle of analogy—*i.e.,* in the absence of firsthand experience, scholars must treat reports of miraculous events with skepticism since people do not encounter such events in their own experiences (here Troeltsch adopts the position of David Hume); and third, the principle of correlation—*i.e.,* events in history are continuous with one another in a causal nexus, which rules out irruptions into the causal order by God: if he works in history he is immanently in all of it. Troeltsch, it may be noted, had some effect on the sociology of religion—*e.g.,* in his distinction between church-type and sect-type organizations in the history of Christianity, a distinction that has formed the starting point of considerable researches in recent times, as noted above. The implications of Troeltsch's historical treatment of religion seemed to be relativistic. Christianity, at any rate, is viewed as a part of religious history as a whole, a point that had not always been clearly recognized by theologians. Troeltsch thereby raised some important questions about the relationship between Christianity and other religions and showed how Christian theology was beginning to take a more realistic view of mankind's religious experience and history, in distinction to the earlier rather simplistic dichotomies between special (*i.e.,* Judeo-Christian) and general (*i.e.,* natural) revelation.

Discoveries about ancient Middle Eastern religions were also bound to affect biblical studies, and a well-defined school developed in Germany—the Religionsgeschichtliche Schule (History of Religions school)—which was critical of the rather unhistorical treatment of Jesus by Ritschl and others. This school emphasized the degree to which biblical ideas were the product of the ancient cultural milieu. Important in this line of development was Albert Schweitzer, in whose *Quest of the Historical Jesus* the eschatological teachings (statements about the "last times," or end of the world as it is now understood) of Jesus are emphasized, together with the dissimilarity of his thought world from our own. Criticism of Harnack also came from a different direction. The French theologian Alfred Loisy (1857–1940), from a Roman Catholic point of view but taking into account the work of Protestant biblical critics, found the essence of Christianity in the faith of the developed church, which could not be found simply by trying to discover the nature of the historical Jesus. The founder, in effect, of Modernism within the Roman Catholic Church, Loisy was excommunicated; and this was a main factor in discouraging some of the livelier Roman Catholic studies of the New Testament until after the epochal ecumenical second Vatican Council (1962–65).

Neo-orthodoxy and demythologization. Liberal Protestantism of the Harnack type was severely criticized by Karl Barth, the founder of Neo-orthodoxy; liberalism's optimism, in any event, came under a cloud through the outbreak of World War I. Barth's *Epistle to the Romans* and his later *Church Dogmatics* became highly influential. His theology depended in part on a distinction between the Word (*i.e.,* God's self-revelation as concretely manifested in Christ and in preaching) and religion. The latter, according to Barth, is the product of human culture and aspirations and is not to be identified with saving revelation (for salvation cannot come from mankind, only from God). This rather uncompromising view made use of the projectionist theory of religion expressed by Feuerbach and others. Barth's conclusion was challenged somewhat by another Swiss theologian, Emil Brunner (1889–1966), who allowed a modicum of insight for fallen man into God's nature. The concession was, however, a slight one. The Dutch theologian Hendrik Kraemer (1888–1965) applied the doctrine of the theology of the Word to non-Christian religions in *The Christian Message in a Non-Christian World,* which had a wide impact on the overseas mission field. Since man's religions are cultural products and since each system of belief is organic and particular, there are, according to Kraemer, no points of contact between them and the Gospel (even Christianity as an empirical religion must be distinguished from it: its only advantage is to have been continuously under the judgment and influence of the Gospel). Kraemer's position has come under some criticism from students of comparative religion; one of the theological problems it poses is that it seems to shut off the possibilities of dialogue between religions.

Influence of Barth and Bultmann

After Barth, the most influential theologian in the 20th century has been Rudolf Bultmann (1884–1976). Though he was mainly concerned with the presentation of the faith, his project of "demythologization" has had a wide significance for the historian of religions, for it involves a theory of myth. Bultmann came to the New Testament material partly as a historian and partly as a theologian influenced by the Existentialism of Heidegger. He centred his interest on the difference between the style of thinking in the early church, as expressed in the New Testament writings, and modern thought. Modern man, he held, cannot think in the mythological terms employed in the New Testament presentation of the Gospel. Therefore, it is necessary to demythologize the New Testament message. For Bultmann, the mythological elements are belief in the pre-existence of Christ, the three-layer universe (heaven, earth, and hell), miracles, ascension into heaven, demonology, and various other elements of the Judeo-Christian-Hellenistic world view. The inner meaning of the myths, he claimed, must be explicated in existential terms and purged of the objectifications that they contain. Thus, his theory contains an empirical claim, namely about the original function of myths (expressing existential attitudes through objectified representations). Bultmann's theory, however, has not yet been brought together with anthropological and other theories of myth.

A follower of Bultmann, Fritz Buri, considers Bultmann's stance to be insufficiently radical, for Bultmann differentiated between the kerygma (the essential proclamation of the early church) and the myths, desiring to retain the former, but not the latter. Buri has attempted to overcome this distinction. Authentic existence is not, according to Buri, distinctively Christian, and he has been led to a position not altogether different in principle from that of Troeltsch. Buri's views have also led him into considering in some depth the significance of other religions.

The relationship of Western Christianity to other religions. Since World War II, Western Christianity has found it difficult, from a cultural point of view, to ignore the challenge of other religions; and the mood has changed somewhat from the more rigorous climate in which the theology of the Word (*i.e.,* Barth's position) was dominant. The "theology of religions" (analogous to the "history of religions") has moved in the direction of dialogue, which sometimes simply refers to mutual acquaintance in charity so that people of differing faiths can come to understand more deeply the meaning of each other's religions. More significantly, it means a kind of mutual theologizing. Among the more prominent writers who have been involved one way or another in the process of dialogue have been the Jewish philosopher Martin Buber (1878–1965), the English Islāmic scholar Kenneth Cragg, and the Canadian Islāmic scholar Wilfred Cantwell Smith. In effect, modern dialogue continues an earlier tradition that emphasized some continuities between religions, notably the work of the British theologian John Oman (1860–1939), who was influenced both by Schleiermacher and Otto, though critical of the latter. Oman contrasted prophetic and mystical religion and considered that the former had the highest conception of the supernatural. There are analogies between his position and that of the important Swedish theologian, historian of religion, and archbishop Nathan Söderblom (1866–1931).

Inter-religious dialogue

A rather different theory of myth and symbolism from that of Bultmann was expressed by Paul Tillich, who viewed religion as having to do with what concerns man ultimately. He taught that symbolic and mythological language, used by all religions, points beyond itself to the being in which the symbols participate. Tillich used the term being in an existential sense (one related directly to human experience and commitment) rather than a strictly metaphysical sense. Also, he claimed that it is not possible

to dispense with the symbolic, which is essential to the task of speaking about ultimate reality, but the myths are to be "broken"—that is, they are to be seen as not being literally true.

Christian theology, in the 19th and 20th centuries, has been more concerned with intellectual and social challenges, however, than with the analysis of religion, which has been secondary to that concern.

History and phenomenology of religion. The history of religions and the phenomenology of religion are generally understood by scholars to be nonnormative—that is, they attempt to delineate facts, whether historical or structural, without judging them from a Christian or other standpoint. At any rate, their tasks are considered to be different from that of articulating and systematizing a faith. The same, in principle, is true for the comparative study of religion, though this sometimes is thought to cover the theology of other religions, such as the Christian appraisal of Hindu history. Needless to say, the fact that a discipline aims to be nonnormative does not mean that it will succeed in being so. Also, the history and phenomenology of religion tend to raise essentially philosophical questions of explanation, where the issues are often debatable.

Modern origin and development of the history and phenomenology of religion. The history of religions on a cross-cultural basis, though it has quite an ancient pedigree, came into its own in a modern sense from about the time of Max Müller. During the latter part of the 19th century an attempt was made to place comparative methodology on a systematic basis (often called the Science of Religion), and in this connection the work of the Dutch theologians P.D. Chantepie de la Saussaye (1848–1920) and C.P. Tiele (1830–1902) was important. During this period, various lectureships and chairs in the subject were instituted. In The Netherlands, following the reform of the theological faculties in 1876, four chairs in the history of religions were founded. In 1879 a chair was founded at the Collège de France (followed by others elsewhere in France), while a number were created in Switzerland. The subject also spread to Great Britain (where chairs at Manchester and London were instituted), the United States (at Harvard and Chicago), and elsewhere in the Western world. In Germany, on the other hand, there was strong resistance, notably from Adolf von Harnack, who thought that theology should avoid what he regarded as dilettantism and that the subject was sufficiently covered in the study of biblical religion.

The first congress of *Religionswissenschaft* (Science of Religion) took place in Stockholm in 1897, and a similar one in the history of religions at Paris in 1900. Later, the International Association for the History of Religions, dedicated to a mainly nonnormative and nontheological approach, was formed. Also important was the compilation of encyclopaedias, notably Hastings' *Encyclopaedia of Religion and Ethics,* with many distinguished contributions. Thus, there were development and progress in the new subject in the latter part of the 19th and early part of the 20th century. In the 1960s came the next major burst of expansion.

A great amount of the work of scholars in the field has been devoted to exploring particular histories—piecing together, for instance, the history of Gnosticism (a Hellenistic-Christian heretical sect that emphasized dualism) or of early Buddhism. In principle, Christianity is considered from the same point of view, but much significant work has also been comparative and structural. This can range from the attempt to establish rather particular comparisons, such as Otto's comparison (in his *Mysticism East and West*) of the medieval German mystic Meister Eckehart and the medieval Hindu philosopher Śaṅkara, to a systematic typology, as in *Religion in Essence and Manifestation* by the Dutch historian of religion Gerardus van der Leeuw.

There have been many significant scholars in the history and phenomenology of religion since Max Müller. Rudolf Otto (1869–1937) made a profound impression on the scholarly world with the publication of *The Idea of the Holy* (in its German edition of 1917), which showed the influence of Schleiermacher, Marett, Edmund Husserl, and the Neo-Kantianism of Jakob Fries (1773–1843). More important than the philosophical side of his enterprise, however, was the excellent delineation of a central experience and sentiment and the elucidation of the concept of the Holy. The central experience Otto refers to is the numinous (Latin *numen,* "spirit") in which the Other (*i.e.,* the transcendent) appears as a *mysterium tremendum et fascinans*—that is, a mystery before which man both trembles and is fascinated, is both repelled and attracted. Thus, God can appear both as wrathful or awe inspiring, on the one hand, and as gracious and lovable, on the other. The sense of the numinous, according to Otto, is *sui generis,* though it may have psychological analogies, and it gives an access to reality, which is categorized as holy. Otto stresses what he calls the nonrational character of the numinous, but he does not deny that rational attributes may be applied to God (or the gods or other numinous powers), such as goodness and personality. The impact of Otto's work, however, does not depend on the now rather curious Neo-Kantian scheme into which he presses his data. Not all scholars would agree that the numinous is universal as a central element in religion, as Otto seems to have supposed: early Jainism and Theravāda Buddhism, for example, have other central values. Otto's treatment of mysticism, which is central to Buddhism, wavers somewhat, and the notions of the "wholly Other" and of the *tremendum* do not easily apply to the experience of Nirvāṇa (the state of bliss) or to other deliverances of the contemplative mystical consciousness.

The concept of the Holy

Friedrich Heiler (1892–1967), like Otto a professor at Marburg (Germany), was a strong proponent of the phenomenological and comparative method, as in his major work on prayer. Heiler, however, went beyond the scientific study of religion in attempting to promote interreligious fellowship, partly through the Religiöser Menschheitsbund (Union of Religious Persons), which he helped to found. Heiler believed in the essential unity of religions—a recurring theme in various guises in the period, though open to question because of the widely apparent divergences between prophetic and other religions, such as Theravāda Buddhism and Jainism, which do not believe in a supreme personal being.

The phenomenologist of religion who probably has had the greatest influence after Otto, partly because he is fairly explicit about method, is Gerardus van der Leeuw (1890–1950), who was somewhat influenced by the French anthropologist Lucien Lévy-Bruhl (1857–1939) and his notion of prelogical mentality, which he applied to primitive cultures to distinguish them from civilized cultures. Van der Leeuw emphasized power as being the basic religious conception. His major work, *Religion in Essence and Manifestation,* is an ambitious and wide-ranging typology of religious phenomena, including the kinds of sacrifice, types of holy men, categories of religious experience, and other types of religious phenomena. The work has been criticized, however, as being unhistorical. Partly because of his philosophical presuppositions, borrowed chiefly from Husserl, van der Leeuw held the disputable doctrine that Phenomenology knows nothing of the historical development of religion: it picks out timeless essences of religious phenomena. Apparently it is not necessary, however, to hold this doctrine, since one could as well classify types of religious change (*i.e.,* temporal sequences), as indeed Max Weber attempted to do. Classificatory and historical techniques and conclusions are not incompatible, however. Thus, the work of Nathan Söderblom, who, as well as being a historian of religions, was prominent in the ecumenical movement, combined the two aspects in his *Living God.*

The "Chicago school." The phenomenological method was brought to the United States primarily by the German-American historian of religions Joachim Wach (1898–1955), who established *Religionswissenschaft* (Science of Religion) in Chicago and was thus the founder of the modern "Chicago school" (though his successor, Mircea Eliade, has a rather different slant). Wach was concerned with emphasizing three aspects of religion—the theoretical (or mental; *i.e.,* religious ideas and images), the practical (or behavioral), and the institutional (or social); and be-

Emphases of Joachim Wach and Mircea Eliade

cause of his concern for the study of religious experience, he interested himself in the sociology of religion, attempting to indicate how religious values tended to shape the institutions that expressed them. Wach, however, was not committed to a religious neutralism in his use of the idea of a "science of religion." For him, *Religionswissenschaft* deepens the sense of the numinous and strengthens, rather than paralyzes, religious impulses.

Mircea Eliade (1907–), a Romanian scholar who emigrated to the United States after World War II, has had a wide influence, partly because of his substantive studies on *yoga* (a Hindu meditation technique) and on shamanism (both these major works are now regarded as classical studies of their subjects) and partly because of his later writings, which attempt to synthesize data from a wide variety of cultures. The synthesis incorporates a theory of myth and history. Eliade was also a founder of the journal *History of Religions*, which expresses the "Chicago school" viewpoint. Eliade has been somewhat influenced by Jung, both in his psychological interpretations of certain religious experiences (such as those attained in the practice of *yoga*) and more importantly in his attempt to give an interpretation in depth to the mythic material over which he ranges so widely. He also affirms strongly the importance of the history of religions in the intellectual world and is thus concerned to emphasize its unique and positive role in providing a "creative hermeneutics" (critical interpretive method) of man's religious and existential condition. Two important elements in the theory of Eliade are, first, that the distinction between the sacred and the profane is fundamental to religious thinking and is to be interpreted existentially (the symbols of religion are, typically, profane in literal interpretation but are of cosmic significance when viewed as signs of the sacred); and, second, that archaic religion is to be contrasted with the linear, historical view of the world. The latter essentially comes from biblical religion; the former viewpoint tends to treat time cyclically and mythically—referring to foundational events, such as the creation, the beginning of the human race, and the Fall of man, on to *illud tempus* (the sacred primordial time), which is re-enacted in the repetitions of the ritual and in the retelling of the myth. Though Christianity has contained archaic elements, in essence it is linear and historical. Thus, faith in Christianity involves a kind of fall from archaic timelessness, and secularization—in which the overt symbolism of religion is driven underground into the unconscious—is a second fall. Eliade is not very explicit about his meaning beyond this point. Not only is he concerned with descriptive phenomenology, in which context his analysis of the religious functions of time and space is most illuminating, but also with a kind of metaphysical speculation (as exemplified in his idea of the "fall").

The myth and ritual school

Other studies and emphases. Though not always giving a detailed account of the correlation between myth and ritual, Eliade is indebted to the so-called myth and ritual school, which has influenced thinking in the history of religions and which was important in the 1930s, especially in the interpretation of Middle Eastern mythology. Thus, the *Enuma elish*, the Babylonian creation epic, was discovered to be no mere set of stories but rather a mythic drama re-enacted every year at the spring festival, at which time the foundation of the world is ritually renewed. More generally, it was seen that for a wide range of sacred stories it was important to discover the ritual context. The most influential statement of the school's position is to be found in *Myth and Ritual* (1933), edited by the English biblical scholar and Orientalist Samuel Hooke.

Meanwhile, the categorization of types of religion (*e.g.*, as polytheism, henotheism, or other) continued to stimulate attempts at a deeper understanding of the emergence of monotheism. To some extent scholars remained under the influence of the older evolutionism. An important work in this connection was *Dio: Formazione e sviluppo del monoteismo nella storia delle religioni* ("God: Formation and Development of Monotheism in the History of Religions"), by the Italian historian of religion Raffaele Pettazzoni (1883–1959), who emphasized the importance of the divinized sky in the development of monotheism. He was critical of the *Urmonotheismus* of Wilhelm Schmidt, considering that the latter's theory of an original monotheism went very far beyond the evidence. At best, the facts could only support the conclusion that primitive peoples believed in a supreme celestial being. Pettazzoni, in his concern for problems of method, was critical of the sharp division between phenomenology and history. He considered that the former cannot exist without the historical sciences—*e.g.*, history, philology, and archaeology—but that it supplies scholars in the latter fields a sense of the religious significance of what they discover. This point of view has also been more vigorously espoused by the Swedish scholar Geo Widengren (1907–), who has specialized mainly in Iranian religions. The need to integrate historical and structural studies has caused some debate in recent years; and there has also been some contrast made between historical approaches and contemporary sociological and (essentially theological) dialogical approaches to religion. To some extent, such debates represent different ideals of scholarship; but it is difficult to note where the essential incompatibilities lie. For many scholars, the multidisciplinary way of studying religion is difficult to comprehend.

Meanwhile, the longstanding interest in the Indo-European group of religions was given a new impetus in the work of the French comparative philologist and mythologist Georges Dumézil (1898–), who broke away from an etymological (analysis of word derivations) approach and sought instead the thematic traits of the gods in the mythical material. This approach, pioneered by others before Dumézil, also was skeptical of the easy identification of gods with natural forces and emphasized the sociological functions of the divinities—without, however, holding to a reductionist theory. Dumézil's theory was partly stimulated by discoveries in the Middle East, notably that of Boğazköy (Turkey), which revealed a similarity between some of the chief gods of the Indo-European Mitannians and those of the Aryans of the Indian Vedic tradition. His theory correlated the functions of the gods with the tripartite division of Indo-European societies—namely the priestly regal, the nobility, and the producers (agriculturalists, craftsmen). Though his work has been controversial (there are, for instance, some difficulties about its application to ancient Greece, despite the fact that the analysis seems to apply to the threefold division of society into philosophers, warriors, and producers in Plato's *Republic*), there is no doubt that the search for correlated functions of the kind Dumézil postulated has been significant in the area of Indo-European mythology.

Studies on the sociological functions of deities

Dumézil's work is one example of a thematic, comparative study. The interest in such studies has grown since World War II. Examples can be found in the writings of such thinkers as the English scholar S.G.F. Brandon (1907–71) in his treatment of ideas such as creation and time in different religions, but with special reference to the ancient Middle East, and the English Indo-Iranian scholar R.C. Zaehner (1913–74), notably in his work on mysticism, as in his *Mysticism Sacred and Profane*. Zaehner's was a definitely Christian approach rather than a scientific-descriptive one; and his concern was to distinguish between theistic and other forms of mysticism, such as monistic mysticism as found, according to him, in Yoga, Advaita, and even Theravāda Buddhism.

Apart from the comparative, phenomenological studies, there has also been a strong growth of historical work in regard to particular religions. This has been most obvious in Indian religions—in Hinduism and Buddhism especially. In part, this is the result of a general growth in non-Christian religions in the post-World War II era and of the need to come to terms with Asian and African cultures after the demise of European hegemony.

PROBLEMS AND DIRECTIONS

The foregoing, a necessarily rather selective account of some of the principal developments and scholars in the various disciplines related to the descriptive, analytical study of religion, emphasizes the artificiality of some of the divisions between traditional disciplines. Thus, Dumézil's work could as easily fall under sociology or anthropology

as under the history of religions; and there are obvious connections between philosophy and sociology in, for example, Marxist interpretations of religion. Again, the description and typology of religious experience belong as much to psychology as to the phenomenology of religion, and the analysis of the nature of symbolism requires a variety of disciplinary approaches. To some extent, the study of religion has suffered from the barriers between disciplines, and this fact is increasingly recognized in the formulations, notably in the United States, of the idea of religion as a subject that should be institutionalized in a university department or program in which historians, phenomenologists, and members of other disciplines work together. There are some, however, who consider that there are dangers in such an arrangement; thus Eliade prefers to work rather tightly within the framework of the history of religions, concerned lest the social sciences overwhelm and distract the interpreter of religious meanings. Similarly, the theological tradition in the West remains powerfully operative (quite legitimately) in regard to the articulation of the Christian faith and sometimes resists any attempt to treat Christianity itself in the manner dictated by the history and phenomenology of religion. Thus, the history of religions and the comparative study of religion still tend to mean in practice "the study of religions other than Judaism and Christianity." Educational and social pressures have arisen, however, within a secularistic, increasingly pluralistic society and (in effect) a shrunken world, increasing the tendency toward a pluralism in the study of religion that expands the viewpoints of traditional faculties and departments of theology, both in universities and theological seminaries.

A further problem about the multidisciplinary study of religion is that little has been done to explore the problem of the people to whom religions are interpreted—the clientele for the subject. Hitherto, the main assumption has been that the study is for Westerners, though a number of distinguished Asian and African scholars are working in the field. Until recently, owing to the unequal cultural and political relationship between Western and non-Western religions, however, some of the most vital contributions have been primarily attempts to articulate (for the new apologetic situation) the old traditions. This has been a main concern of scholars of Asian religions such as Sarvepalli Radhakrishnan, T.R.V. Murti, and K.N. Jayatilleke. The prospect is, however, that an intellectual community will be the clientele of the subject. To this extent the study of religions will most likely involve, as it does already to some extent, a complex dialogue between religions.

Another problem is the need to elucidate the basis of a dynamic typology of religion in which phenomenology and history are properly brought together. The tendency toward a rift between the historians and phenomenologists is unnecessary and causes harm to the pursuit of the subject.

Meanwhile, some emergent tendencies within the various disciplines can be perceived. There is an increased concern in anthropological theory for the content of religious symbolism, such as in the work of the English anthropologist Mary Douglas; and the sociology of religion is, in a sense, returning to the method of Max Weber in stressing the comparison of cultures. The important development of Oriental and African studies since World War II has made this task easier—American sociologists have, for example, examined in some detail Japanese culture and religion. The interest in symbolism and mythology coincides with developments in the philosophy of religion, which, under the influence of Wittgenstein (in his later, more open phase), is concerned with explicating different functions of language. One area of the study of religion that is seriously underdeveloped at the present time—other than in respect to the psychoanalytic approaches—is the psychology of religion, although current interest in mysticism and other forms of religious experience has stimulated the collection and interpretation of data. One of the difficult problems to be solved is the extent to which cultural conditioning exerts an influence on the actual content of such experience.

In many ways the present position promises well for an expanding multidisciplinary approach to problems in the study of religion. Historians of religion are recognizing some of the contributions to be made by modern sociology, and sociologists—partly because of the development of the sociology of knowledge—have become more aware of the need for accounting for the particular systems of meaning in religion. An area that may very well exhibit the new synthesis is the study of new religious movements.

After a period of relative unconcern, Christian theology is increasingly aware of the challenge of other religious beliefs, so that there are greater impulses toward blending Christian and other studies—often kept rather artificially apart, though biblical studies, especially Old Testament studies, have usually been quite closely related to the history of the relevant religions of the ancient Middle East.

Meanwhile, in a number of Western countries (chiefly in Europe, but also to some extent in the United States), the study of religion on a pluralistic and multidisciplinary basis is being increasingly viewed as an important element in the education of secondary school students. This, together with the popularity of the subject in universities, may ensure that the study of religion will increase in significance.

(N.Sm./Ed.)

The classification of forms and phenomena

The classification of religions, the attempt to systematize and bring order to a vast range of knowledge about man's religious beliefs, practices, and institutions, has been the goal of students of religion for many centuries but especially so with the increased knowledge of the world's religions and the advent of modern methods of scientific inquiry in the last two centuries.

The classification of religions involves: (1) the effort to establish groupings among historical religious communities having certain elements in common or, (2) the endeavour to group similar religious phenomena in categories that serve to reveal the structure of human religious experience as a whole.

FUNCTION AND SIGNIFICANCE

Classification as a tool in understanding religion

The many schemes suggested for classifying religious communities and religious phenomena all have one purpose in common; *i.e.*, to bring order, system, and intelligibility to the vast range of knowledge about human religious experience. Classification is basic to all science as a preliminary step in reducing data to manageable proportions and in moving toward a systematic understanding of a subject matter. Like the zoologist who must distinguish and describe the various orders of animal life as an indispensable stage in the broad attempt to understand the character of such life as a whole, the student of religion also must use the tool of classification in his outreach toward a scientific account of man's religious experience. The growth of scientific interest in religion in Western universities over the past 130 years has compelled most leading students of religion to discuss the problem of classification or to develop classifications of their own.

The difficulty of classifying religions is accounted for by the immensity of religious diversity that history exhibits. As far as scholars have discovered, there has never existed any people, anywhere, at any time, who were not in some sense religious. The individual who embarks upon the arduous task of trying to understand religion as a whole confronts an almost inconceivably huge and bewilderingly variegated host of phenomena from every locale and every era. Empirically, what is called religion includes the mythologies of the preliterate peoples on the one hand and the abstruse speculations of the most advanced religious philosophy on the other. Historically, religion, both ancient and modern, embraces both primitive religious practices and the aesthetically and symbolically refined worship of the more technologically progressive and literate human communities. The student of religion does not lack material for his studies; his problem is rather to discover principles that will help him to avoid the confusion of too much information. Classification is precisely the appeal to such principles; it is a device for making the

otherwise unmanageable wealth of religious phenomena intelligible and orderly.

The endeavour to group religions with common characteristics or to discover types of religions and religious phenomena belongs to the systematizing stage of religious study. According to Max Müller,

> All real science rests on classification and only in case we cannot succeed in classifying the various dialects of faith, shall we have to confess that a science of religion is really an impossibility.

PRINCIPLES OF CLASSIFICATION

The criteria employed for the classification of religions are far too numerous to catalogue completely. Virtually every scholar who has considered the matter has evidenced a certain amount of originality in his view of the interrelationships among religious forms. Thus, only some of the more important principles of classification will be discussed.

True and false religion

Normative. Perhaps the most common division of religions—and in many ways the most unsatisfactory—distinguishes true religion from false religion. Such classifications may be discovered in the thought of most major religious groups and are the natural, perhaps inevitable, result of the need to defend particular perspectives against challengers or rivals. Normative classifications, however, have no scientific value, because they are arbitrary and subjective, inasmuch as there is no agreed method for selecting the criteria by which such judgments should be made. But because living religions always feel the need of apologetics (systematic intellectual defenses), normative classifications continue to exist.

Many examples of normative classification might be given. The early Church Fathers (*e.g.,* Clement of Alexandria, 2nd century AD) explained that Christianity's Hellenistic (Greco-Roman culture) rivals were the creations of fallen angels, imperfect plagiarisms of the true religion, or the outcome of divine condescension that took into account the weaknesses of men. The greatest medieval philosopher and theologian, Thomas Aquinas, distinguished natural religion, or that kind of religious truth discoverable by unaided reason, from revealed religion, or religion resting upon divine truth, which he identified exclusively with Christianity. In the 16th century Martin Luther, the great Protestant Reformer, forthrightly labelled the religious views of Muslims, Jews, and Roman Catholic Christians to be false and held the view that the gospel of Christianity understood from the viewpoint of justification by grace through faith was the true standard. In Islām, religions are classified into three groups: the wholly true, the partially true, and the wholly false, corresponding with Islām, the Peoples of the Book (Jews, Christians, and Zoroastrians), and polytheism. The classification is of particular interest because, being based in the Qurʾān, (the Islāmic sacred scripture), it is an integral part of Islāmic teaching, and also because it has legal implications for Muslim treatment of followers of other religions.

Although scientific approaches to religion in the 19th century discouraged use of normative categories, elements of normative judgment were, nonetheless, hidden in certain of the new scientific classifications that had emerged. Many evolutionary schemes developed by anthropologists and other scholars, for example, ranked religions according to their places on a scale of development from the simplest to the most sophisticated, thus expressing an implicit judgment on the religious forms discussed. Such schemes more or less clearly assume the superiority of the religions that were ranked higher (*i.e.,* later and more complex); or, conversely, they serve as a subtle attack on all religion by demonstrating that its origins lie in some of humanity's basest superstitions, believed to come from an early, crude stage. A normative element is also indicated in classification schemes that preserve theological distinctions, such as that between natural and revealed religion. In short, the normative factor still has an important place in the classification of religions and will doubtless always have, since it is extraordinarily difficult to draw precise lines between disciplines primarily devoted to the normative exposition of religion, such as theology and philosophy of religion, and disciplines devoted to its description or scientific study.

Geographical categories

Geographical. A common and relatively simple type of classification is based upon the geographical distribution of religious communities. Those religions found in a single region of the earth are grouped together. Such classifications are found in many textbooks on comparative religion, and they offer a convenient framework for presenting man's religious history. The categories most often used are: (1) Middle Eastern religions, including Judaism, Christianity, Islām, Zoroastrianism, and a variety of ancient cults; (2) Far Eastern religions, comprising the religious communities of China, Japan, and Korea, and consisting of Confucianism, Taoism, Mahāyāna ("Greater Vehicle") Buddhism, and Shintō; (3) Indian religions, including early Buddhism, Hinduism, Jainism, and Sikhism, and sometimes also Theravāda Buddhism and the Hindu- and Buddhist-inspired religions of South and Southeast Asia; (4) African religions, or the cults of the tribal peoples of black Africa, but excluding ancient Egyptian religion, which is considered to belong to the ancient Middle East; (5) American religions, consisting of the beliefs and practices of the Indian peoples indigenous to the two American continents; (6) Oceanic religions—*i.e.,* the religious systems of the peoples of the Pacific islands, Australia, and New Zealand; (7) classical religions of ancient Greece and Rome and their Hellenistic descendants. The extent and complexity of a geographical classification is limited only by the classifier's knowledge of geography and his desire to seek detail and comprehensiveness in his classification scheme. Relatively crude geographical schemes that distinguish Western religions (usually equivalent to Christianity and Judaism) from Eastern religions are quite common.

Although religions centred in a particular area often have much in common because of historical or genetic connections, geographical classifications present obvious inadequacies. Many religions, including some of the greatest historical importance, are not confined to a single region (*e.g.,* Islām), or do not have their greatest strength in the region of their origins (*e.g.,* Christianity, Buddhism). Further, a single region or continent may be the dwelling place of many different religious communities and viewpoints that range from the most archaic to the most sophisticated. At a more profound level, geographical classifications are unacceptable because they have nothing to do with the essential constitutive elements or inner spirit of religion. The physical location of a religious community reveals little of the specific religious life of the group. Though useful for some purposes, geographical classifications contribute minimally to the task of providing a systematic understanding of man's religions and religiousness.

Use of the comparative study of languages

Ethnographic-linguistic. Max Müller, often called the "Father of the history of religions," stated that "Particularly in the early history of the human intellect, there exists the most intimate relationship between language, religion, and nationality." This insight supplies the basis for a genetic classification of religions (associating them by descent from a common origin), which Müller believed the most scientific principle possible. According to this theory, in Asia and Europe dwell three great races, the Turanians (including the Ural-Altaic peoples), the Semites, and the Aryans, to which correspond three great families of languages. Originally, in some remote prehistory, each of these races formed a unity, but with the passage of time they split up into a myriad of peoples with a great number of distinct languages. Through careful investigation, however, the original unity may be discerned, including the unity of religion in each case. Müller's principal resource in developing the resulting classification of religions was the comparative study of languages, from which he sought to demonstrate similarities in the names of deities, the existence of common mythologies, the common occurrence of important terms in religious life, and the likeness of religious ideas and intuitions among the branches of a racial group. His efforts were most successful in the case of the Semites, whose affinities are easy to demonstrate, and probably least successful in the case of the Turanian peoples, whose early origins are hypothetical. Müller's greatest contribution to scholarship, however, lay in his study of

Aryan languages, literatures, and comparative mythology.

Because Müller was a scholar of the first rank and a pioneer in several fields, his ethnographic-linguistic (and genetic) classification of religions has had much influence and has been widely discussed. The classification has value in exhibiting connections that had not been previously observed. Müller (and his followers) discovered affinities existing among the religious perspectives of both the Aryan and Semitic peoples and set numerous scholars on the path of investigating comparative mythology, thus contributing in a most direct way to the store of knowledge about religions.

There are, nevertheless, difficulties with the ethnographic-linguistic classification. To begin with, Müller's evidence was incomplete, a fact that may be overlooked given the state of knowledge in his day. More important is the consideration that peoples of widely differing cultural development and outlook are found within the same racial or linguistic group. Further, the principle of connection among race, language, and religion does not take sufficiently into account the historical element or the possibility of developments that may break this connection, such as the conversion of the Aryan peoples of Europe to a Semitic religion, Christianity.

Other scholars have developed the ethnographic classification of religion to a much higher degree than did Müller. The German scholar Duren J.H. Ward, for example, in *The Classification of Religions* (1909) accepted the premise of the connection between race and religion but appealed to a much more detailed scheme of ethnological relationship. He says that "religion gets its character from the people or race who develop or adopt it" and further that

> the same influences, forces, and isolated circumstances which developed a special race developed at the same time a special religion, which is a necessary constituent element or part of a race.

In order to study religion in its fullness and to bring out with clarity the historical and genetic connections between religious groups, the ethnographic element must thus have adequate treatment. Ward devised a comprehensive "Ethnographico-historical Classification of the Human Races to facilitate the Study of Religions—in five divisions." These major divisions were (1) the Oceanic races, (2) the African races, (3) the American races, (4) the Mongolian races, and (5) the Mediterranean races, each of which has its own peculiar religion. The largest branch, the Mediterranean races, he subdivided into primeval Semites and primeval Aryans, in order to demonstrate in turn how the various Semitic, Indo-Aryan, and European races descended from these original stocks.

Philosophical. The past 150 years have also produced several classifications of religion based on speculative and abstract concepts that serve the purposes of philosophy. The principal example of these is the scheme of G.W.F. Hegel, a seminal German philosopher, in his famous *Lectures on the Philosophy of Religion* (1832). In general, Hegel's understanding of religion coincided with his philosophical thought; he viewed the whole of human history as a vast dialectical movement toward the realization of freedom. The reality of history, he held, is Spirit, and the story of religion is the process by which Spirit—true to its own internal logical character and following the dialectical pattern of thesis, antithesis, and synthesis (the reconciliation of the tension of opposite positions in a new unity that forms the basis of a further tension)—comes to full consciousness of itself. Individual religions thus represent stages in a process of evolution (*i.e.,* progressive steps in the unfolding of Spirit) directed toward the great goal at which all history aims.

Hegelian dialectics

Hegel classified religions according to the role that they have played in the self-realization of Spirit. The historical religions fall into three great divisions, corresponding with the stages of the dialectical progression. At the lowest level of development, according to Hegel, are the religions of nature, or religions based principally upon the immediate consciousness deriving from sense experience. They include: immediate religion or magic at the lowest level; religions, such as those of China and India plus Buddhism, that represent a division of consciousness within itself; and others, such as the religions of ancient Persia, Syria, and Egypt, that form a transition to the next type. At an intermediate level are the religions of spiritual individuality, among which Hegel placed Judaism (the religion of sublimity), ancient Greek religion (the religion of beauty), and ancient Roman religion (the religion of utility). At the highest level is absolute religion, or the religion of complete spirituality, which Hegel identified with Christianity. The progression thus proceeds from man immersed in nature and functioning only at the level of sensual consciousness, to man becoming conscious of himself in his individuality as distinct from nature, and beyond that to a grand awareness in which the opposition of individuality and nature is overcome in the realization of Absolute Spirit.

Many criticisms have been offered of Hegel's classification. An immediately noticeable shortcoming is the failure to make a place for Islām, one of the major historical religious communities. The classification is also questionable for its assumption of continuous development in history. The notion of perpetual progress is not only doubtful in itself but is also compromised as a principle of classification because of its value implications.

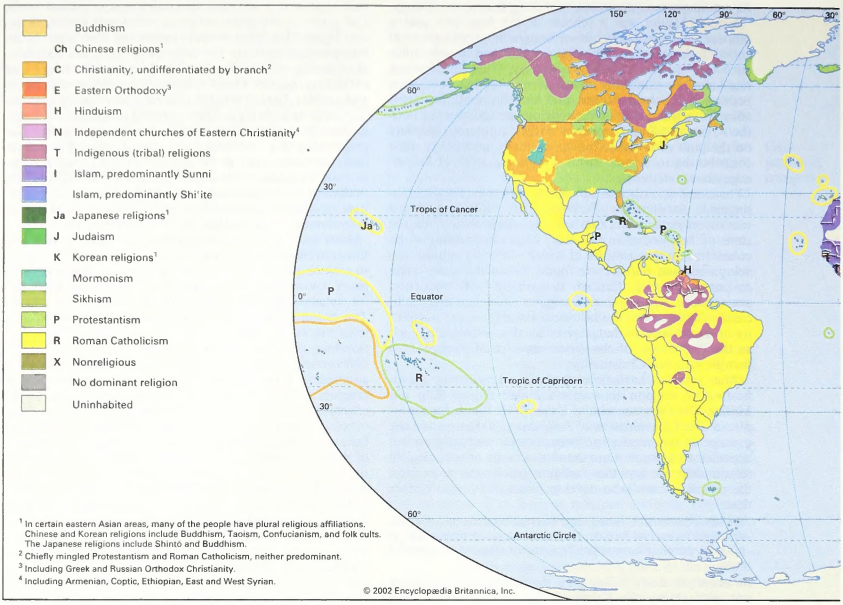
Nevertheless, Hegel's scheme was influential and was adapted and modified by a generation of philosophers of religion in the Idealist tradition. Departure from Hegel's scheme, however, may be seen in the works of Otto Pfleiderer, a German theologian of the 19th century. Pfleiderer believed it impossible to achieve a significant grouping of religions unless, as a necessary preliminary condition, the essence of religion were first isolated and clearly understood. Essence is a philosophical concept, however, not a historical one. Pfleiderer considered it indispensable to have conceptual clarity about the underlying and underived basis of religion from which all else in religious life follows. In *Die Religion, ihr Wesen und ihre Geschichte* ("Religion, Its Essence and History"), Pfleiderer held that the essence of religious consciousness exhibits two elements, or moments, perpetually in tension with one another: one of freedom and one of dependence, with a number of different kinds of relationships between these two. One or the other may predominate, or they may be mixed in varying degrees.

Religions of freedom and dependence

Pfleiderer derived his classification of religions from the relationships between these basic elements. He distinguished one great group of religions that exhibits extreme partiality for one over against the other. The religions in which the sense of dependence is virtually exclusive are those of the ancient Semites, the Egyptians, and the Chinese. Opposite these are the early Indian, Germanic, and Greek and Roman religions, in which the sense of freedom prevails. The religion of this group may also be seen in a different way, as nature religions in the less-developed cultures or as culture or humanitarian religions in the more advanced. A second group of religions exhibits a recognition of both elements of religion, but gives them unequal value. These religions are called supernatural religions. Among them Zoroastrianism gives more weight to freedom as a factor in its piety, and Brahmanism and Buddhism are judged to have a stronger sense of dependence. The last group of religions is the monotheistic religions: Islām, Judaism, and Christianity, which are divided again into two sub-groups, *i.e.,* those that achieve an exact balance of the elements of religion and those that achieve a blending and merging of the elements. Both Judaism and Islām grant the importance of the two poles of piety, though there is a slight tendency in Islām toward the element of dependence and in Judaism toward freedom. It is Christianity alone, he claimed, that accomplishes the blending of the two, realizing both together in their fullness, the one through the other.

The intellectual heritage that lies behind this classification will be immediately apparent. The classification reflects its time (19th century) and place (western Europe) of conception in the sense that the study of religion was not yet liberated from its ties to the philosophy of religion and theology.

Morphological. Considerable progress toward more scientific classifications of religions was marked by the



Geographical distribution of the religions of the world.

emergence of morphological schemes, which assume that religion in its history has passed through a series of discernible stages of development, each having readily identifiable characteristics and each constituting an advance beyond the former stage. So essential is the notion of progressive development to morphological schemes that they might also be called evolutionary classifications. Trends in the comparative study of religions have retained the interest in morphology but have decisively rejected the almost universal 19th-century assumption of unitary evolution in the history of religion. The crude expression of evolutionary categories such as the division of religions into lower and higher or primitive and higher religions has been subjected to especially severe criticism.

Influence of E.B. Tylor

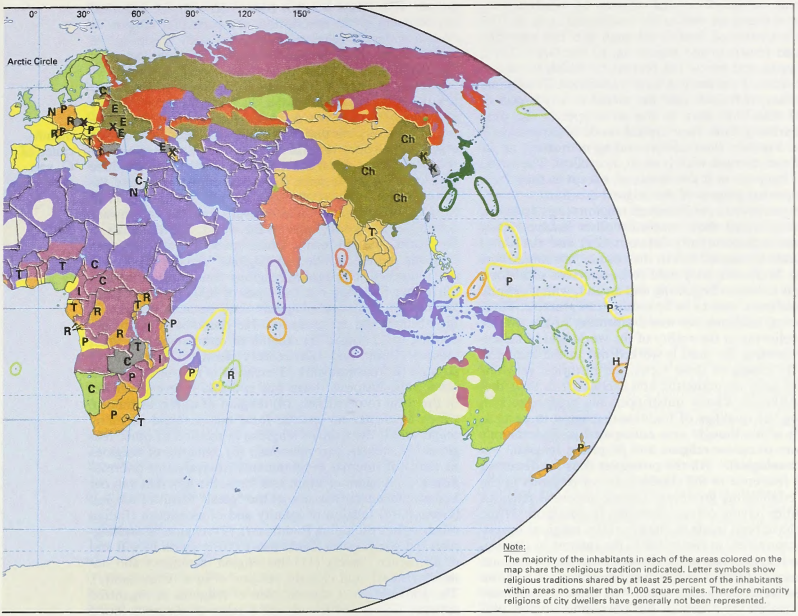
The pioneer of morphological classifications was E.B. Tylor, a British anthropologist, whose *Primitive Culture* (1871) is among the most influential books ever written in its field. Tylor developed the thesis of animism, a view that the essential element in all religion is belief in spiritual beings. According to Tylor, the belief arises naturally from elements universal in human experience (*e.g.,* death, sleep, dreams, trances, and hallucinations) and leads through processes of primitive logic to the belief in a spiritual reality distinct from the body and capable of existing independently. In the development of the idea, this reality is identified with the breath and the life principle; thus arises the belief in the soul, in phantoms, and in ghosts. At a higher stage, the spiritual principle is attributed to aspects of reality other than man, and all things are believed to possess spirits that are their effective and animating elements; for example, primitive men generally believe that spirits cause sickness and control their destinies.

Of immediate interest is the classification of religions drawn from Tylor's animistic thesis. Ancestor worship, prevalent in preliterate societies, is obeisance to the spirits of the dead. Fetishism, the veneration of objects believed to have magical or supernatural potency, springs from the association of spirits with particular places or things and leads to idolatry, in which the image is viewed as the symbol of a spiritual being or deity. Totemism, the belief in an association between particular groups of people and certain spirits that serve as guardians of those people, arises when the entire world is conceived as peopled by spiritual beings. At a still higher stage, polytheism, the interest in particular deities or spirits disappears and is replaced by concern for a "species" deity who represents an entire class of similar spiritual realities. By a variety of means, polytheism may evolve into monotheism, a belief in a supreme and unique deity. Tylor's theory of the nature of religions and the resultant classification were so logical, convincing, and comprehensive that for a number of years they remained virtually unchallenged.

The morphological classification of religions received more sophisticated expression from C.P. Tiele, a 19th-century Dutch scholar and an important pioneer in the scientific study of religion. His point of departure was a pair of distinctions made by the philosophers of religion Abraham Kuenen and W.D. Whitney. In the Hibbert Lectures for 1882, *National Religions and Universal Religions,* Kuenen had emphasized the difference between religions limited to a particular people and those that have taken root among many peoples and qualitatively aim at becoming universal. Whitney saw the most marked distinction among religions as being between race religions ("the collective product of the wisdom of a community") and individually founded religions. The first are the result of nature's unconscious working through long periods of time, and the latter are characterized by a high degree of ethical awareness. Tiele agreed strongly with Whitney in distinguishing between nature and ethical religions. Ethical religion, in Tiele's views, develops out of nature religion,

> But the substitution of ethical religions for nature-religions is, as a rule, the result of a revolution; or at least of an intentional reform.

Each of these categories (*i.e.,* nature or spiritualistic–ethical) may be further subdivided. At the earliest and lowest



Religions of nature and spiritualistic–ethical religions

stage of spiritual development was polyzoic religion, about which there is no information but which is based on Tiele's theory that man must have regarded natural phenomena as endowed with life and superhuman magical power. The first known stage of the nature religions is called polydaemonistic (many spirits) magical religion, which is dominated by animism and characterized by a confused mythology, a firm faith in magic, and the preeminence of fear above other religious emotions. At a higher stage of nature religions is therianthropic polytheism, in which the deities are normally of mixed animal and human composition. The highest stage of nature religion is anthropomorphic polytheism, in which the deities appear in human form but have superhuman powers. These religions have some ethical elements, but their mythology portrays the deities as indulging in all sorts of shocking acts. None of the polytheistic religions, thus, was able to raise itself to a truly ethical point of view.

Ethical religions fall into two subcategories. First are the national nomistic (legal) religions that are particularistic, limited to the horizon of one people only and based upon a sacred law drawn from sacred books. Above them are the universalistic religions, qualitatively different in kind, aspiring to be accepted by all men, and based upon abstract principles and maxims. In both subtypes, doctrines and teachings are associated with the careers of distinct personalities who play important roles in their origin and formation. Tiele found only three examples of this highest type of religion: Islām, Christianity, and Buddhism.

Influence of Nathan Söderblom

Tiele's classification enjoyed a great vogue and influenced many who came after him. Nathan Söderblom, a Swedish archbishop who devoted much energy to problems of classification, accepted the division of higher religions into two great groups but used a varied terminology that pointed to some of the characteristics of the two types of religion. In addition to natural religion and revealed religion, or religions of nature and religions of revelation, Söderblom spoke of culture religions and founded religions, and of nature religions and historical religions. The highest expression of the first category is the "mysticism of infinity" that is characteristic of the higher aspects of Hindu and Buddhist religious experience. The apex of genuine prophetic religion is reached in the "mysticism of personality." All these distinctions mean the same thing, and all are indebted to Tiele's thought. Söderblom, however, sharply disagreed with Tiele's thesis of continuous development in the history of religion. In Söderblom's view, the line between nature religion and prophetic religion is a deep and unbridgeable chasm, a qualitative difference so enormous that one type could never evolve by natural historical processes into the other. Prophetic religion can be explained only as a radical and utterly new incursion into history. As Söderblom was a churchman and theologian as well as a distinguished historian of religion, there is without doubt an element of theological judgment influencing his stand on this matter. Söderblom was eager to defend the uniqueness of biblical religion, and he believed that his historical and scientific studies provided an objective basis for asserting not only the uniqueness but also the superiority of Christianity.

Tiele's enduring influence may also be seen in the classification of religions advanced by Mircea Eliade, a Romanian-American scholar who was one of the most prolific contemporary students of religion. Eliade, who in other respects might be considered among the phenomenologists of religion, was interested in uncovering the "structures" or "patterns" of religious life. The basic division that Eliade recognized is between traditional religions—including primitive religions and the archaic cults of the ancient civilizations of Asia, Europe, and America—and historical religions. The distinction is better revealed, however, in the terms cosmic religion and historical religion. In Eliade's estimation, all of traditional religion shares a common outlook upon the world—chiefly, the deprecation of history and the rejection of profane, mundane time. Religiously, traditional man is not interested in the unique and specific but rather exclusively in those things and actions that repeat and restore transcendental models. Only those things that participate in and reflect the eternal archetypes

or the great pattern of original creation by which cosmos came out of chaos are real in the traditional outlook. The religious activities of traditional man are the recurring attempts to return to the beginning, to the Great Time, to trace again and renew the process by which the structure and order of the cosmos were established. Traditional religions may, therefore, find the sacred in any aspect of the world that links man to the archetypes of the time in the beginning; thus, their typical mode of expression is repetitive. Further, their understanding of history, as far as they are concerned with it at all, is cyclical. The world and what happens in it are devalued, except as they show forth the eternal pattern of the original creation.

Modern, postarchaic, or historical religions (*e.g.,* Judaism, Christianity, Islām) show markedly other features. They tend to see a discontinuity between God and the world and to locate the sacred not in the cosmos but somewhere beyond it. Moreover, they hold to linear views of history, believing it to have a beginning and an end, with a definite goal as its climax, and to be by nature unrepeatable. Thus, the historical religions are world affirming in the double sense of believing in the reality of the world and of believing that meaning for man is worked out in the historical process. By reason of these views, the historical religions alone have been monotheistic and exclusivist in their theologies. Although Eliade outstripped his predecessors in delineating the qualities of traditional religion in particular, much of his thought was anticipated in Söderblom's descriptions of nature religion and prophetic religion.

Phenomenological. All the principles thus far discussed have had reference to the classification of religions in the sense of establishing groupings among historical religious communities having certain elements in common. While attempts have been made to classify entire religions or religious communities, in recent times the interest in classifying entire religions has markedly declined, partly because of an emerging interest in the phenomenology of religion.

This new trend in studies, which has come to dominate the field, claims its origin in the phenomenological philosophy of Edmund Husserl, a German Jewish–Lutheran scholar, and has found its greatest exponents in The Netherlands. Phenomenology of religion has at least two aspects. It is first of all an effort at devising a taxonomic (classificatory) scheme that will permit the comprehensive cataloging and classifying of religious phenomena across the lines of religious communities, but it is also a method that aims at revealing the self-interpretation by religious men of their own religious responses. Phenomenology of religion thus rejects any overview of religion that would interpret religion's development as a whole, confining itself rather to the phenomena and the unfolding of their meaning for religious men. Phenomenologists are especially vigorous in repudiating the evolutionary schemes of past scholars, whom they accuse of imposing arbitrary semiphilosophical concepts in their interpretation of the history of religion. Phenomenologists also have little interest in history for its own sake, except as a preliminary stage of material gathering for the hermeneutical (critical–interpretive) task that is to follow.

Purposes of the phenomenology of religion

One of the earliest Dutch phenomenologists, W. Brede Kristensen (1867–1953), spoke of his work as follows:

> Phenomenology of Religion attempts to understand religious phenomena by classifying them into groups . . . we must group the phenomena according to characteristics which correspond as far as possible to the essential and typical elements of religion.

The material with which phenomenology is concerned is all the different types of religious thinking and action, ideas about divinity, and cultic acts. Kristensen's systematic organization of religious phenomena may be seen in the table of contents of his *Meaning of Religion* in which he divides his presentation of material into discussions of (1) cosmology, which includes worship of nature in the form of sky and earth deities, animal worship, totemism, and animism, (2) anthropology, made up of a variety of considerations on the nature of man, his life, and his associations in society, (3) cultus, which involves consideration of sacred places, sacred times, and sacred images, and (4) cultic acts, such as prayer, oaths and curses, and ordeals. Kristensen was not concerned with the historical development or the description of a particular religion or even a series of religions but rather with grouping the typical elements of the entire religious life, irrespective of the community in which they might occur.

Probably the best known phenomenologist is G. van der Leeuw, another Dutch scholar. In his *Religion in Essence and Manifestation,* van der Leeuw categorized the material of religious life under the following headings: (1) the object of religion, or that which evokes the religious response, (2) the subject of religion, in which there are three divisions: the sacred man, the sacred community, and the sacred within man, or the soul, (3) object and subject in their reciprocal operation as outward reaction and inward action, (4) the world, ways to the world, and the goals of the world, and (5) forms, which must take into account religions and the founders of religions. Van der Leeuw was not interested in grouping religious communities as such but rather in laying out the types of religious expression. He discussed distinct religions only because religion in the abstract has no existence. He classified religions according to 12 forms: (1) religion of remoteness and flight (ancient China and 18th-century deism), (2) religion of struggle (Zoroastrianism), (3) religion of repose, which has no specific historical form but is found in every religion in the form of mysticism, (4) religion of unrest or theism, which again has no specific form but is found in many religions, (5) dynamic of religions in relation to other religions (syncretism and missions), (6) dynamic of religions in terms of internal developments (revivals and reformations), (7) religion of strain and form, the first that van der Leeuw characterizes as one of the "great" forms of religion (Greece), (8) religion of infinity and of asceticism (Indian religions but excluding Buddhism), (9) religion of nothingness and compassion (Buddhism), (10) religion of will and of obedience (Israel), (11) the religion of majesty and humility (Islām), and (12) the religion of love (Christianity). The above is not a classification of religions as organized systems. Categories 3, 4, 5, and 6 relate to elements found in many if not all historical religious communities, and the categories from 7 onward are not classifications but attempts to characterize particular communities by short phrases that express what van der Leeuw considered to be their essential spirit. The "primitive" religions of less-developed peoples are not classified.

Significance of G. van der Leeuw

Other principles. William James, the American philosopher and psychologist, in his book *The Varieties of Religious Experience,* differentiated two types of religion according to the attitude toward life—the religion of healthy-mindedness, which minimizes or ignores the evil of existence, and that of morbid-mindedness, which considers evil as the very essence of life. Max Weber, a German sociologist, distinguished between religions that express themselves primarily in mythopoeic ways and those that express themselves in rational forms. The distinction comes very close to that between traditional and historical religions, though its emphasis is somewhat different.

Psychological and sociological types

Nathan Söderblom, in his prolific scholarly career, devised several classifications other than the principal one discussed above. In his great work on primitive religions, *Das Werden des Gottesglaubens* ("Development of the Belief in God"), Söderblom divided religions into dynamistic, animistic, and theistic types according to the way primitive peoples apprehend the divine. In other works (*Einführung in die Religionsgeschichte,* or "Introduction to the History of Religion," and *Thieles Kompendium der Religionsgeschichte neu bearbeitet,* or "Tiele's Compendium of the History of Religion Revised") he contended that Christianity is the central point of the entire history of religions and, therefore, classified religions according to the historical order in which they came into contact with Christianity. Similarly, Albert Schweitzer, the French theologian, medical missionary, and Nobel laureate, in *Christianity and the Religions of the World,* grouped religions as rivals or nonrivals of Christianity. Still another scheme may be seen in Söderblom's Gifford Lectures, *The Living God,* in which religions were divided according to their doctrines of the relation between human and divine activity in the achievement of salvation. Thus, among higher religions

there are those in which man alone is responsible for salvation (Buddhism), God alone is responsible (the Bhakti cults of India), or God and man cooperate (Christianity).

The American sociologist Robert Bellah, having in mind the advances of the social sciences in their understanding of religions, offers a refurbished and more highly sophisticated version of an evolutionary scheme that he thinks to be the most satisfactory possible in the present state of scholarly knowledge. He views religion as having passed through five stages, beginning with the primitive and proceeding through the archaic, the historical, and the early modern to the modern stage. The religious complexes that emerge in each stage of this evolution have identifiable characteristics that Bellah studies and differentiates according to the following categories: symbol systems, religious actions, religious organizations, and social implications. Two basic concepts run through Bellah's classification, providing the instruments for the division of religions along the evolutionary scale. The first is that of the increasing complexity of symbolization as one moves from the bottom to the top of the scale, and the second is that of increasing freedom of personality and society from their environing circumstances or, in other words, the growing secularization of the religious field. Bellah's classification is important because of the wide discussion it has awakened among social scientists.

One may find additional classifications based upon the content of religious ideas, the forms of religious teaching, the nature of cultus, the character of piety, the nature of the emotional involvement in religion, the character of the good toward which religions strive, and the relations of religions to the state, to art, to science, and to morality.

CONCLUSION

The classification of religions that will withstand all criticism and serve all the purposes of a general science of religions has not been devised. Each classification presented above has been attacked for its inadequacies or distortions, yet each is useful in bringing to light certain aspects of religion. Even the crudest and most subjective classifications throw into relief various aspects of religious life and thus contribute to the cause of understanding. The most fruitful approach for a student of religion appears to be that of employing a number of diverse classifications, each one for the insight it may yield. Though each may have its shortcomings, each also offers a positive contribution to the store of knowledge and its systematization. The insistence upon the exclusive validity of any single taxonomic effort must be avoided. To confine oneself to a single determined framework of thought about so rich and variegated a subject as religion is to risk the danger of missing much that is important. Classification should be viewed as a method and a tool only.

Criteria for making and judging classifications

Although a perfect classification lies at present beyond scholars' grasp, certain criteria, both positive and negative in nature, may be suggested for building and judging classifications. First, classifications should not be arbitrary, subjective, or provincial. A first principle of the scientific method is that objectivity should be pursued to the extent possible and that findings should be capable of confirmation by other observers. Second, an acceptable classification should deal with the essential and typical in the religious life, not with the accidental and the unimportant. The contribution to understanding that a classification may make is in direct proportion to the penetration of the bases of religious life exhibited in its principles of division. A good classification must concern itself with the fundamentals of religion and with the most typical elements of the units it is seeking to order. Third, a proper classification should be capable of presenting both that which is common to religious forms of a given type and that which is peculiar or unique to each member of the type. Thus, no classification should ignore the concrete historical individuality of religious manifestations in favour of that which is common to them all, nor should it neglect to demonstrate the common factors that are the bases for the very distinction of types of religious experience, manifestations, and forms. Classification of religions involves both the systematic and the historical tasks of the general science of religion. Fourth, it is desirable in a classification that it demonstrate the dynamics of religious life both in the recognition that religions as living systems are constantly changing and in the effort to show, through the categories chosen, how it is possible for one religious form or manifestation to develop into another. Few errors have been more damaging to the understanding of religion than that of viewing religious systems as static and fixed, as, in effect, ahistorical. Adequate classifications should possess the flexibility to come to terms with the flexibility of religion itself. Fifth, a classification must define what exactly is to be classified. If the purpose is to develop types of religions as a whole, the questions of what constitutes a religion and what constitutes various individual religions must be asked. Since no historical manifestation of religion is known that has not exhibited an unvarying process of change, evolution, and development, these questions are far from easily solved. With such criteria in mind it should be possible continuously to construct classification schemes that illuminate man's religious history. (C.J.A.)

BIBLIOGRAPHY. JAN DE VRIES, *The Study of Religion* (1967), a fairly useful and brief historical survey of the development of the subject; H. PINARD DE LA BOULLAYE, *L'Étude comparée des religions,* 2 vol. (1922–25), a thorough and excellent account; J. MILTON YINGER, *The Scientific Study of Religion* (1970), an attempt to indicate the multidisciplinary approach to the study of religion; J. HASTINGS (ed.), *Encyclopaedia of Religion and Ethics,* 13 vol. (1908–26), dated in many respects but still enormously important; PAUL EDWARDS (ed.), *Encyclopedia of Philosophy,* 8 vol. (1967), many entries on world religions, doctrines, and religious thinkers; JOHN MACQUARRIE, *Twentieth Century Religious Thought* (1963), a survey, despite its title, of both 19th- and 20th-century thinkers, including many important in the history and phenomenology of religion (also a good general guide to issues in modern Western theology); G. VAN DER LEEUW, *Phänomenologie der Religion* (1933; Eng. trans., *Religion in Essence and Manifestation,* 1938), the most wide-ranging and ambitious attempt at a systematic and classificatory phenomenology of religion; RUDOLF OTTO, *Das Heilige* (1917; Eng. trans., *The Idea of the Holy,* 2nd ed., 1950), a highly influential classic; J. WACH, *The Comparative Study of Religion* (1958) and *Sociology of Religion* (1962), still useful compendiums; J. HINNELLS (ed.), *The Comparative Study of Religion in Education* (1970); MICHAEL BANTON (ed.), *Anthropological Approaches to Religion* (1966); E. EVANS-PRITCHARD, *Theories of Primitive Religion* (1965), which, with Banton, indicates the main issues about the genesis and function of religion debated by anthropologists; THOMAS O'DEA, *The Sociology of Religion* (1966), a useful survey; and MAX WEBER, *Religionssoziologie* (1922; Eng. trans., *The Sociology of Religion,* ed. by TALCOTT PARSONS, 1963), a good introduction to the thought of Weber. PETER BERGER, *The Sacred Canopy* (also published as *The Social Reality of Religion,* 1969), more speculative but a stimulating example of modern sociological theorizing about religion; V. LANTERNARI, *The Religions of the Oppressed* (1963), an example of comparative sociology of religion; and J. HICK, *The Philosophy of Religion* (1963), a useful survey of issues in the philosophy of religion. MIRCEA ELIADE wrote widely from a standpoint that combines elements drawn from depth psychology, phenomenology, and the history of religions: his *Sacred and the Profane* (1961) and *The Quest* (1969) give an insight into his general approach. (N.Sm.)

Two monographs dealing specifically with the classification of religions, each of which offers a survey of previous classifications in addition to the author's own scheme, are DUREN J.H. WARD, *The Classification of Religions: Different Methods, Their Advantages and Disadvantages* (1909); and FRED LOUIS PARRISH, *The Classification of Religions: Its Relation to the History of Religions* (1941), containing a full survey of classification schemes with brief characterizations of each and the best bibliographical guide for pursuing the subject in depth. Other books for further study are as follows: P.D. CHANTEPIE DE LA SAUSSAYE, *Lehrbuch der Religionsgeschichte,* 2 vol. (1887–89; Eng. trans. of vol. 1, *Manual of the Science of Religion,* 1891), which includes classification problems at the beginning of vol. 1; C.P. TIELE, *Elements of the Science of Religion,* 2 vol. (1897–99), a classic work by an important scholar on this subject; and F. MAX MUELLER, *Introduction to the Science of Religion* (1873), another classic work. Of more recent origin is GUSTAV MENSCHING, *Die Religion: Erscheinungsformen, Strukturtypen und Lebensgesetze* (1959), a popular manual of the history of religions that includes a long section on classification problems. (C.J.A.)

Systems of Religious and Spiritual Belief

Man, it has been said, is incurably religious. This is an epigrammatic and somewhat facetious statement of the most impressive feature of religions—their wide distribution. All through human history and all through present-day societies, there appears some systematic relation to what is regarded as sacred. To cover all these phenomena, however, the definition of religion must be so comprehensive as to become ambiguous, if not meaningless. Each successive broadening of the geographic horizon, moreover, has necessitated a redefinition of religion to include unfamiliar phenomena and to exclude features that had been part of the older definition because of a generalization based on too few forms.

This article examines the structural categories into which religious and spiritual belief systems, and various repudiations of such systems, are conventionally grouped for comparison and contrast. The more particularized elements of belief within such systems and of practice or worship are treated cross-culturally in the *Macropædia* articles DOCTRINES AND DOGMAS, RELIGIOUS; and RITES AND CEREMONIES, SACRED, respectively. The particular religions themselves are discussed in the *Macropædia* in either of two ways: articles on the non-surviving religions of ancient peoples are grouped under the regional headings MIDDLE EASTERN RELIGIONS, ANCIENT; and EUROPEAN RELIGIONS, ANCIENT; and individual articles on the particular extant religions are to be found under their respective names (*e.g.*, CHRISTIANITY; ISLĀM, MUḤAMMAD AND THE RELIGION OF; ZOROASTRIANISM AND PARSIISM). The history and methods of the study of religions are discussed in the *Macropædia* article RELIGIONS, THE STUDY AND CLASSIFICATION OF.

For coverage of related topics in the *Macropædia* and the *Micropædia*, see the *Propædia*, sections 811, 812, and 10/53, and the *Index*.

The article is divided into the following sections:

Nature worship 531
  Nature as a sacred totality 531
  Heaven and earth as sacred spaces, forces, or processes 531
    Heaven
    Earth
  Celestial phenomena as objects of worship or veneration 533
    The sun
    The moon
    Eclipses of the sun and moon
    Stars and constellations
  Elements and forces of nature 535
    Water
    Fire
    Weather
  Worship of animals 537
Animism 537
  Importance in the study of man and religion 537
  Theoretical issues 538
    Tylor's theory of animism
    Counter theories
  Animistic phenomena in their social contexts 539
  The animistic world view 539
Totemism 540
  The nature of totemism 540
    Group totemism
    Individual totemism
  Some examples of totemism 541
    Wiradjuri
    Nor-Papua
    Iban
    Birhor
    Kpelle
  A short history of totemistic theory 542
    From McLennan to Lévi-Strauss
    Present situation and emerging trends
Ancestor worship 544
  Nature and significance 544
  Basic patterns and functions 545
  Ancestor worship in religions of the world 545
    In nonliterate societies
    In Eastern societies
    In ancient Middle Eastern and European societies
  Theories and interpretations 546
Polytheism 547
  The nature of polytheism 547
  Forms of polytheistic powers, gods, and demons 547
    Natural forces and objects
    Vegetation
    Animal and human forms
    Functional deities
  Types of polytheism 549
Pantheism and panentheism 550
  Nature and significance 550
  Diverse views of the relation of God to the world 551
  Pantheism and panentheism in non-Western cultures 551
    Hindu doctrines
    Buddhist doctrines
    Ancient Middle Eastern doctrines
  Pantheism and panentheism in ancient and medieval philosophy 552
    Greco-Roman doctrines
    Medieval doctrines
  Pantheism and panentheism in modern philosophy 553
    Renaissance and post-Renaissance doctrines
    Nineteenth-century doctrines
    Twentieth-century doctrines
  Criticism and evaluation of pantheism and panentheism 555
Religious dualism 555
  Nature and significance 555
  Historical varieties of religious dualism 556
    Among ancient civilizations and peoples
    Among religions of the East
    Among religions of the West
    Among religions of modern nonliterate peoples
  Themes of religious dualism 558
    The sacred and the profane
    Good and evil
    Creation and destruction: life and death
    Polytheistic themes
  Functions of religious dualism 559
    Cosmological and cosmogonic functions
    Anthropological functions
    Sociological functions
Monotheism 560
  The spectrum of views: monotheisms and quasi-monotheisms 560
    The basic monotheistic view
    The extreme positions
    The middle positions
    Alternate positions
  Monotheism in the world religions 562
    Classical monotheism
    Monotheistic elements in ancient Middle Eastern and Mediterranean religions
    Monotheistic elements in Indian and Chinese religions
Theism 563
  Theism in Western thought 563
    God encountered as person
    The existence of God
    The problem of particular knowledge of God
    The nature of God in modern thought
    Theism in Islām
  Theism in Eastern thought 566
Deism 567
  Nature and scope 567
  The historical Deists 567
    The English Deists
    Deists in other countries

Agnosticism 569
Nature and kinds of agnosticism 569
Huxley's nonreligious agnosticism
Religious agnosticism
Historical antecedents of modern agnosticism 570
Antecedents of secular agnosticism
Antecedents of religious agnosticism
Incompatibility with fideism 571
Rejections of the agnostic principle 571
Atheism 572
Atheism as rejection of religious beliefs 572
Atheism and theism 572
Atheism and metaphysical beliefs 573
Atheism and intuitive knowledge 574
Comprehensive definition of atheism 574
Bibliography 576

Nature worship

In the history of religions and cultures, nature worship as a definite and complex system of belief or as a predominant form of religion has not been well documented. Among primitive peoples the concept of nature as a totality is unknown; only individual natural phenomena—*e.g.,* stars, rain, and animals—are comprehended as natural objects or forces that influence them and are thus in some way worthy of being venerated or placated. Nature as an entity in itself, in contrast with man, human society and culture, or even God, is a philosophical or poetic conception that has been developed among advanced civilizations. This concept of nature worship, therefore, is limited primarily to scholars involved in or influenced by the modern (especially Western) study of religion.

NATURE AS A SACRED TOTALITY

To students of religion, the closest example of what may be termed nature worship is perhaps most apparent in ancient or nonliterate cultures in which there is a high god as the lord in heaven who has withdrawn from the immediate details of the governing of the world. This kind of high god—the *Deus otiosus,* hidden, or idle, god—is one who has delegated all work on earth to what are called "nature spirits," which are the forces or personifications of the forces of nature. High gods exist, for example, in such indigenous religions on Africa's west coast as that of the Dyola of Guinea. In such religions the spiritual environment of man is functionally structured by means of personified natural powers, or nature spirits.

Pantheism (a belief system in which God is equated with the forces of the universe) or deism (a belief system based on a non-intervening creator of the universe), as was advocated in the rationalistic philosophy of religion of western Europe of the 16th to 18th centuries, is not appropriate in studies of nature worship in preliterate cultures. Worship of nature as an omnipotent entity, in the pantheistic sense, has not as yet been documented anywhere.

The concepts of mana, orenda, wakan, and manitou

The power or force within nature that has most often been venerated, worshipped, or held in holy awe is mana. Mana, often designated as "impersonal power" or "supernatural power," is a term used by Polynesians and Melanesians that 19th-century Western anthropologists appropriated to apply to that which affected the common processes of nature. Mana was conceptually linked to North American Indian terms that conveyed the same or similar notions—*e.g.,* orenda of the Iroquois, wakan of the Dakotas, and manitou of the Algonkin. Neither the designation "impersonal power" nor "supernatural power" implies what mana really means, however, because mana usually issues from persons or is used by them, and the concept of a supernatural sphere as distinct or separate from a natural sphere is seldom recognized by preliterate peoples.

Thus, a better designation for mana is "super force" or "extraordinary efficiency." A person has mana when he is successful, fortunate, and demonstrates extraordinary skill—*e.g.,* as an artisan, warrior, or chief. Mana can also be obtained from the *atuas* (gods), providing that they themselves possess it. Derived from a root term that has aristocratic connotations, mana corresponds to Polynesian social classifications. The *ariki,* or *alii,* the nobility of Polynesia, have more power (mana), and the area that belongs to them and even the insignia associated with them have mana. Besides areas and symbolic elements that are associated with the *ariki,* many objects and animals having special relationships with chiefs, warriors, or priests have mana.

The concept of *hasina* of the Indonesian Hova (or Merina) on Madagascar is very similar to mana. It demonstrates the same aristocratic root character as the word mana, which is derived from the Indonesian *manang* ("to be influential, superior").

The Iroquoian term orenda, similar to mana, designates a power that is inherent in numerous objects of nature but that does not have essential personification or animistic (soul) elements. Orenda, however, is not a collective omnipotence. Powerful hunters, priests, and shamans have orenda to some degree. The *wakanda,* or wakan, of the Sioux Indians is described similarly, but as Wakan-Tanka it may refer to a collective unity of gods with great power (wakan). The manitou of the Algonkin is not merely an impersonal power, comparable to the wakan, that is inherent in all things of nature but is also the personification of numerous manitous (powers), with a Great Manitou (Kitchi-Manitou) at the head. These manitous may even be designated as protective spirits that are akin to those of other North American Indians, such as the *digi* of the Apaches, *boha* of the Shoshones, and *maxpe* of the Crow, as well as the *sila* of the Eskimos.

The super forces (such as Mulungu, Imana, Jok, and others in Africa) that Western scholars have noted outside of the Austronesian–American circles of peoples are often wrongly interpreted as concepts of God. Only the *barakah* (derived from the pre-Islāmic thought world of the Berbers and Arabs), the contagious superpower (or holiness) of the saints, and the power Nyama in western Sudan that works as a force within large wild animals, certain bush spirits, and physically handicapped people—appearing especially as a contagious power of revenge—may be added with a certain justification to that force of nature that is designated by mana. A striking similarity with mana may also be noted in the concepts of *heil* (good omen), *saell* (fortunate), and *hamingja* (luck) of the Germanic and Scandinavian peoples.

HEAVEN AND EARTH AS SACRED SPACES, FORCES, OR PROCESSES

Heaven and earth as personified powers of nature

Heaven and earth, as personified powers of nature and thus worthy of worship, are evidently not of equal age. Though from earliest times heaven was believed to be the residence of a high being or a prominent god, the earth as a personified entity is much rarer; it probably first occurred among archaic agrarian civilizations, and it continues to occur in some primitive societies in which agriculture is practiced. Gods of heaven, however, are characteristic spiritual beings of early and contemporary hunter and collector cultures and are found in almost all cultures.

Primitive world views generally assume the earth to be simply given (*i.e.,* as continuously existing). Sometimes the earth is believed to have emerged out of chaos or a primal sea or to have come into existence by the act of a heavenly god, transformer, or demiurge (creator). Even in such world views, however, the earth usually remains without a divine owner, unless through agriculture and the cult of the dead the earth is conceived as the underworld or as the source of the renewing powers of nature. The fact that heaven is animated by rain-giving clouds (with lightning and thunder) and by a regular chorus of warming and illuminating celestial bodies (sun, moon, and stars) led to concepts of the personification of heaven from earliest times.

Heaven. Heavenly deities, as the personification of the physical aspects of the sky, appear in variations that are adapted to the types of cultures concerned. The listing offered below does not represent a unilinear development that is applicable everywhere.

The father of the family. The god of heaven is often

viewed as an ever-active father of the family, often called upon but rarely the recipient of sacrifices. He is able to intervene in human and natural affairs without the aid of an intermediary (*e.g.,* priest, medicine man, or ancestors).

Relationship of the "high god" to men

As a numinous (spiritual) being, he is closer to man than other spiritual powers. He sends lightning and rain and rules the stars that are at most essential aspects of himself or are members of his family subject to him. He is the creator and the receiver of the dead. Modern scholars have designated such a being as the "high god," "supreme god," the "highest being" of the "original monotheism" (according to the theories of a German scholar, P.W. Schmidt), the idealized god of heaven (according to the views of an Italian historian of religion, Raffaele Pettazzoni), or the familiar father deity (according to the views of a British anthropologist, Andrew Lang). Very human, often comical, or even unethical and repulsive traits of such deities are often represented in myths that also sometimes include legends of animal or human ancestors.

This type of deity is generally found in its most developed form among the old hunter and collector peoples of the temperate and arid areas (*e.g.,* forest Indians of North America, Indians of California and of Tierra del Fuego in South America, Australian Aborigines, and African Bushmen) and of the tropic primeval forests, where he is usually conceived as a storm and thunder being (*e.g.,* Tore of the Ituri Pygmies or Karei of the Semang of the Andaman Islands). He is also worshipped among the pastoral peoples as the "blue" or "white" sky of the wide pastures in the steppes of northeastern Africa (*e.g.,* Waka of the Galla) and of Central and North Asia (*e.g.,* Torem, Num, and Tengri of the Ugrians, Samoyeds, and Mongols).

Among such peoples, heaven is often merged with an old hunting deity, the Lord of the Animals, or it allows the latter to exist as a hypostasis by his side.

The withdrawn god. The god of heaven may be a *Deus otiosus,* who has, after completing the creation, withdrawn into heaven and abandoned the government of the world to the ancestors of men or to nature spirits that are dependent on him and act as mediators between him and men. This type of the god, who is able to intervene directly only when there are widespread existential necessities or needs (*e.g.,* drought, pestilence, or war), can be found primarily where worship of the dead or worship of individual local "earth spirits"—not yet integrated into an all-inclusive earth deity—obscures everything else. This type of god occurs especially in areas of so-called primitive agriculture (*e.g.,* large parts of Africa, Melanesia, and South America).

Relationships of the "high god" to other deities

The first among equals. The god of heaven also may be the head of a pantheon of gods, the first among equals, or the absolute ruler in a hierarchy of gods. This occurs in polytheism (belief in many gods) in its purest form. The deities associated with him are often related to him by family ties (genealogies of gods). Occasionally, the heavenly phenomena are distributed among members of the clan of gods, the god of heaven himself thus becoming rather vague. The divine pair heaven–earth represents only one among many possible combinations—*e.g.,* Dyauspitṛi (= heaven, male) and Pṛthivī (= earth, female) in Vedic India or, with an unusual distribution of the sexes, Nut (= heaven, woman) and Geb (= earth, man) in ancient Egypt.

Occasionally, generations of gods succeed each other (*e.g.,* Greece, western Asia). In such instances, the more universal god of heaven is often replaced by the younger god of thunderstorms (*e.g.,* Zeus of the Greeks, Teshub of the Hittites, or Hadad of the Semites) or is even relegated to the background by a goddess, such as Inanna-Ishtar (the love or fertility goddess in Babylonia) or Amaterasu, the sun goddess of Japan.

In ancient China, Heaven (T'ien, or Shang Ti, the highest lord) ruled over the many more popular gods and was even closely related to the representatives of the Imperial household. Deification of the celestial emperor is a cultic practice that extends from Korea to Annam (part of Vietnam). The roots of the worship of heaven in Asia are probably the beliefs of central and northern Asiatic nomadic peoples in a solitary god of heaven. Gods of heaven, above or behind a pantheon (grouping of gods), probably originated in areas where a theocratic stratified bureaucracy existed or where sacral kingdoms exist or have existed—*e.g.,* in The Sudan or northeastern Africa (Akan-Baule, Dahomey, Yoruba-Benin, Jukun, Buganda, and neighbouring states), western Indonesia, Polynesia and Micronesia, and in the advanced civilizations of pre-Columbian Meso-America and South America (see also *Polytheism* below).

Heaven and earth deities as partners. The god of heaven in many areas is a partner of an earth deity. In such cases, other numina (spirits) are missing or are subject to one of the two as spirits of nature or ancestors. Myths depicting the heaven–earth partnership usually describe the foundations or origins of the partnership in terms of a separation of a primeval chaos into heaven and earth or in terms of a later separation of heaven and earth that originally lay close together, and they describe the impregnation of the earth by the seed of the god (*e.g., Hieros gamos,* or the sacred marriage). This partnership of the god of heaven and the goddess of earth may be found in areas of Africa that have been influenced by advanced civilizations (especially The Sudan and northeastern Africa), in eastern Indonesia, and in some areas of America under the influence of advanced civilizations.

Not infrequently the god of heaven and the goddess of earth are fused into a hermaphroditic higher deity. This accords with certain traits of ancient civilizations which try to show in customs and myths that the dichotomies, for example, of heaven and earth, day and night, or man and woman, need to be surmounted in a kind of bisexual spiritual force. Certain myths express the loss of an original bisexuality of the world and people. In a creation myth found in the Vedas, for example, it was Puruṣa, an androgynous primal man, who separated into man and woman and from whom the world was created with all its contrasts. Another such creation myth is the cosmic egg, which was separated into the male sky and the female earth.

The god of heaven viewed dualistically. In several religions the god of heaven has an antagonistic evil adversary who delights in destroying completely or partially the good creative deeds of the god of heaven. This helps to explain the insecurity of existence and concepts of ethical dualism. In most such cases, the contrasts experienced in the relationship between heaven and earth deities have been re-evaluated along ethical lines by means of exalting the heavenly elements at the expense of the earthly ones (especially in Jewish, Christian, and Islāmic sects in Europe, west central and northern Asia, and certain areas in the northern half of Africa). The figure of an antagonistic trickster or demiurge that has a somewhat ethical component may be the result of diffusion and is rather rare in primitive cultures—*e.g.,* African Bushmen, Australian Aborigines, and North American Indians (see also *Religious dualism* below).

Exaltation of heavenly elements over earthly elements

The god of heaven viewed monotheistically. The god of heaven, viewed in his ethical aspect, is always an active, single god—*e.g.,* Christian, Jewish, and Islāmic monotheism (see also *Monotheism* below).

Earth. Although in polytheistic religions the earth is usually represented as a goddess and associated with the god of heaven as her spouse, only rarely is there an elaborate or intensive cult of earth worship. There are in many religions mother goddesses who have elaborate cults and who have assumed the function of fertility for land and man, but they hardly have a chthonic (earth) basis. Some mother goddesses, such as Inanna-Ishtar, instead have a heavenly, astral origin. There are, however, subordinate figures of various pantheons, such as Nerthus (Jörd) in Germanic religion or Demeter and Persephone (earth mother and corn girl) in Greek religion, who have played greater roles than the world mother (Gaea). Among Indo-Europeans, western Asiatics (despite their various fertility deities), Chinese, and Japanese, the gods of heaven, sun, and thunderstorms have held a paramount interest.

"Mother earth"

When the common people have displayed intensive attention to "mother earth" (such as the practice of laying down newborn babies on the earth and many other rites), this partially reflects older cults that have remained

relatively free from warrior and nation-building peoples with their emphasis on war (as in western Sudan, pre-Aryan India, and the Indian agrarian area of northern Mexico). The Andean earth-mother figure, Pacha-Mama, worshipped by the Peruvians, stands in sharp contrast to the sun religion of the Inca (the conquering lord of the Andes region). Earth deities are most actively venerated in areas in which people are closely bound to ancestors and to the cultivation of grain.

Mountains. Especially prominent mountains are favourite places for cults of high places, particularly when they are isolated as island mountains, mountains with snowcaps, or uninhabited high mountain ranges. The psychological roots of the cults of high places lie in the belief that mountains are close to the sky (as heavenly ladders), that clouds surrounding the mountaintops are givers of rain, and that mountains with volcanoes form approaches to the fiery insides of the earth.

Mountains, therefore, serve as the abodes of the gods, as the centres of the dead who live underground, as burial places for rainmakers (medicine men), and as places of oracles for soothsayers. In cosmogenic (origin of the world) myths, mountains are the first land to emerge from the primeval water. They frequently become the cosmic mountain (*i.e.,* the world conceived as a mountain) that is symbolically represented by a small hill on which a king stands at the inauguration. Pilgrimages to mountain altars or shrines are favourite practices of cults of high places.

The larger mountain ranges and canyons between volcanic mountains—especially in Eurasia from the Pyrenees to the Alps, the Carpathian Mountains, the Caucasus Mountains, the Himalayas, the mountainous areas of northern China and Japan, and the mountainous areas of North and South America (the Rocky Mountains, the Andes)—are most often centres of cults of high places. Elevations of the East African Rift Valley (Kenya, Tanzania, Uganda), volcanic islands of the Pacific Ocean (*e.g.,* Hawaii), and the mountains of the Indian Deccan have also served as centres of the cult of high places.

In the early civilizations, the cult of high places was closely combined with that of the earth; *e.g.,* Olympus of the Greeks, the mountains of Enlil or of the Mountain Mother Cybele in western Asia, and the Meru mountain of India were believed to bring heaven and earth into a close relationship and were often viewed as the middle pillar of the world pillars upholding the sky. Bush and wild spirits (Lord of the Animals) of the cultures of the hunters and collectors were often believed to reside in inaccessible mountainous areas (*e.g.,* the Caucasus).

The high places as the abodes of the gods

In addition to other mountain deities of a more recent date (*e.g.,* God of the Twelve Mountains and the One-Legged Mountain God), the mountain deity Yama-no-kami has been demonstrated to be a deity of the hunt (*i.e.,* god of the forest, Lord of the Animals) of ancient Japan. Through the worship of farmers, Yama-no-kami assumed the elements of a goddess of vegetation and agriculture. The mountain goddesses (earth mothers) of non-Aryan India still incorporate numerous features of hunt deities, and, because of indigenous influences, the Vedic (early Aryan scriptural) gods and their wives (*e.g.,* Pārvatī, Umā, and Durgā) have their abodes on mountains. The isolated mountains of East Africa, surrounded by clouds, are believed to be the dwelling places of the heaven and rain gods, and in the area of Zimbabwe there are pilgrimages to mountain sanctuaries that are viewed as the seats of the gods.

Pre-Islāmic peoples of North Africa and the extinct inhabitants of the Canary Islands (Guanches) associated mountain worship with a cult of goats (and sheep), which, when practiced in rituals, was believed to secure rain and thunderstorms in the often arid landscape. Similar cults are also found on the Balkans and in the valleys of the southern Alps.

Earthquakes. According to the beliefs of many peoples, earthquakes originate in mountains. In areas of Africa where the concept of mana is particularly strong, many believe that the dead in the underworld are the causes of earthquakes, though in the Upper Nile, the Sudan, and East Africa an earth deity is sometimes blamed. In some areas a bearer who holds the world up—a concept that probably came from Arabia, Persia, India—is believed to cause an earthquake when he changes his position or when he moves his burden from one shoulder to the other. In the Arab world, on the east coast of Africa and in North Africa, an ox generally is viewed as the bearer, sometimes standing on a fish in the water. World bearers often are giants or heroes, such as Atlas, but they also may be animals: an elephant (India), a boar (Indonesia), a buffalo (Indonesia), a fish (Arabia, Georgia, and Japan), a turtle (America), or the serpent god Ndengei (Fiji). Generators of earthquakes also may be the gods of the underworld, such as Tuil, the earthquake god of the inhabitants of Kamchatka (a peninsula in eastern Siberia), who rides on a sleigh under the earth. The earthquake is driven away by noise, loud shouting, or poking with the pestle of a mortar. Among peoples with eschatological (last times) views, earthquakes announce the end of the world (Europe, western Asia).

Tides. The view that the tides are caused by the moon can be found over almost all the earth. This regular natural phenomenon seldom gives rise to cults, but the ebb and flow of the coastal waters has stimulated mythological concepts. Not infrequently the moon acquires the status of a water deity because of this phenomenon. The Tlingit (of northwestern America) view the moon as an old woman, the mistress of the tides. The animal hero and trickster Yetl, the raven, is successful in conquering (with the aid of the mink) the seashore from the moon at low tide, and thus an extended area is gained for nourishment with small sea animals.

CELESTIAL PHENOMENA AS OBJECTS OF WORSHIP OR VENERATION

The sun. Generally, the sun is worshipped more in colder regions and the moon in warm regions. Also, the sun is usually considered as male and the moon as female. Exceptions to these generalizations, however, are notable: the prevalent worship of the sun in hot, arid ancient Egypt and in parts of western Asia; the conception of the moon as a man (who frequently is believed to be the cause of menstruation) among primitive hunter peoples (African Bushmen, Australian Aborigines, and hunters of South America) and among certain pastoral and royal cultures of Africa (*e.g.,* the Masai and the Khoikhoi); and the conception of the female sun ruling northern Eurasia from the Northern Sea to Japan and parts of North America.

In many state cults of ancient civilizations, the sun plays a special role, particularly where it has replaced an old god of heaven (*e.g.,* Egypt, Ethiopia, South India, and the Andes) and especially where it is viewed as a marker of time.

Solar religions in Africa and the Orient

The sun as the centre of a state religion. In Africa ancient Egypt was the main centre from which solar deity concepts emanated. The solar religion, promoted by the state, was concerned with the sun god Re (Atum-Re, Amon-Re, Chnum-Re), the sun falcon Horus, the scarab (Chepre), and a divine kingdom that was determined by the sun (*e.g.,* pharaoh Akhenaton's solar monotheism *c.* 1350 BC). The sun religion reached, by way of Meroe (a sun sanctuary until the 6th century AD) and the Upper Nile, as far as western Ethiopia (*e.g.,* the Hego cult in Kefa and the sun kings in Limmu) and Nigeria (*e.g.,* Jukun). In the Orient the sun cult culminated in the religion of Mithra of Persia. Mithra was transported by Roman legionnaires to western Europe and became the Unconquerable Sun of the Roman military emperors. In Japan the Imperial deity in state Shintō is Amaterasu, the sun goddess from whom Jimmu Tennō, the first human emperor, descended. In Indonesia, where the descent of the princes from the sun also is a feature, the sun often replaces the deity of heaven as a partner of the earth. In Peru the ruling Inca was believed to be the sun incarnate (Inti) and his wife the moon. A sun temple in Cuzco contains a representation of Inti as the oldest son of the creator god. The Natchez Indians of southeastern United States, who are culturally connected with Central America, called their king "Great Sun," and the noblemen were called "the Suns."

The sun as a subordinate deity. The sun, within a polytheistic pantheon, often is revered as a special deity

who is subordinate to the highest deity, usually the god of heaven. This may be observed in the great civilizations of ancient Europe and Asia: Helios (Greece); Sol (Rome); Mithra (Persia); Sūrya, Savitṛ, and Mitra (India); Utu (Sumer); and Shamash (Babylonian and other Semitic areas).

The sun not infrequently is considered female (Shams of some Arabs, Shaph of ancient Ugarit in Palestine, Sun of Arinna of the Hittites, as well as the female Sun of the Germans). Siberian people such as the Taimyr Samoyed (whose women pray in spring to the sun goddess in order to receive fertility or a rich calving of the reindeer) or the Tungus worship sun goddesses. They sacrifice to the sun goddess, and her symbols are embroidered on women's clothes.

The sun and moon as a divine pair. A sun god is often related to a moon goddess as one member of a divine pair (in the place of heaven and earth as "world parents"). A sun-moon god exists among the Munda-Dravida in India (Singbonga); a sun-moon (earth) pair, partially seen as bisexual, exists in eastern Indonesia; and Nyambe (the sun) among the Rotse in Zambia is represented as united with the moon goddess as the ruling pair.

The sun as an attribute of the highest being. The sun sometimes is viewed as a coordinated or subordinate attribute, or hypostasis, of the highest being. This may possibly occur because of a partially weakened influence of a stronger solarism in areas of older primitive peoples, such as in The Sudan, Upper Volta, Nigeria, northern East Africa, and Australia.

The sun as a mythical being. The sun, in some religions, is conceived as a purely mythical being, who is cultically recognized in sun dances (*e.g.,* prairie Indians) and in celebrations of the solstice, with jumps over fires, sports festivals, and other events. These rites may be either survivals of an earlier local cult of a sun deity or influences of such a cult.

The moon. The moon is often personified in different ways and worshipped with ritual customs; nevertheless, in contrast to the sun, the moon is less frequently viewed as a powerful deity. It appears to be of great importance as the basis of a lunar calendar but not in the higher agrarian civilizations. The moon, infrequently associated with the highest god, is usually placed below heaven and the sun. When the moon with the sun together (instead of "heaven and earth") constitute an important pair of gods (world parents), it frequently assumes the features of an earth deity. In tropical South America, the sun and moon are usually purely mythical figures.

The moon as a female or a male deity

Between the Tropic of Capricorn and the Tropic of Cancer, the moon is predominantly female. Only some remainders of ancient peoples (hunter peoples) view the moon as a male being (*e.g.,* African Bushmen, Australian Aborigines, Congo Pygmies, Semang, Andamans, Chaco Indians, Ona, and some Brazilian tribes). In the few significant moon gods of the Oriental civilizations (Khons in Egypt, Sin-Nanna in Babylonia, Candra in India)—in contrast with the female Selene and Luna in the Greek–Roman culture, a more ancient substratum may possibly be present. Where the moon is considered as male, he often determines the sexual life of the woman, especially among the Aborigines of Australia.

The phenomenon of the moon that attracts all people is the sequence of its phases. The waxing and waning of the moon crescent is often interpreted as gaining or losing weight (eating, dieting). Thus, the Taulipang in Guayaná believe that the moon is first nourished well and then inadequately by his two wives, Venus and Jupiter. Where the moon is viewed as female, the phases represent pregnancy and delivery. Elsewhere, people see childhood, maturity, and dying as the phases of the moon: the first crescent is thus the rebirth or the replacement of the old by a new moon.

The appearance of the crescent or the full moon is sometimes celebrated by a rest from work, and some attempt to participate in the waxing and waning of the moon by analogous magical rites. Girls with small breasts stand in the full moonlight (in the Salzburg, Austria, area); persons who desire a tumor to decrease point to the waning moon; and newborn children often are exposed to the waning moonlight, or they and everything else that is desired to be healthy and permanent are dyed white; *i.e.,* they are made "moonlike." Nearly everywhere, connections between the moon phases and the rhythm of nature (the tides) and humans (menstruation) are recognized.

Significance of the phases and signs of the moon

The three dark days of the "death" of the moon are believed by many to be dangerous. During this period the moon is believed to be defeated in a battle with monsters who eat and later regurgitate the moon; or the moon is viewed as having been killed by other heavenly beings and later revived. The period is a time in which people, if possible, do not engage in a new enterprise.

The halo of the moon is also viewed as a bad omen among many peoples. Moon spots are regarded as testimonies of a battle with heavenly opponents. In addition to the popular Man in the Moon, there are also other figures represented: "the woman with the basket on her back," "the spinning woman," or "the weaving woman" (in Polynesia the woman who pounds tapa). The most popular animal figure recognized in the features of the moon, the rabbit (from Europe to America), presumably earned this role because of its fertility.

Eclipses of the sun and moon. Eclipses of the sun or the moon—usually interpreted as a battle, as the dying, or the devouring of one of the two heavenly bodies—in many religions are met with anxiety, shouting, drum beating, shooting, and other noises. Many North American Indian tribes, Hottentots in Africa, Ainu in Japan, and the Minangkabau in Sumatra interpret the eclipses as fainting, sickness, or the death of the darkened heavenly body. In Arctic America, Eskimos, Aleuts, and Tlingits believe that the sun and moon have moved from their places in order to see that things are going right on earth. The explanation that heavenly monsters and beasts pursue the stars and attempt to injure and to kill them, however, is a view found over a larger area. Noise and shooting are believed to deter the monsters from their pursuit or to force them to return the celestial bodies if they have already been captured. In China and Thailand the monster is the heavenly dragon; in China, among the Germanic tribes, and among northern American Indians, dogs and wolves (coyotes) are the culprits; in Africa and Indonesia, they are snakes; in South America, the beast is the jaguar; and in India they are the star monsters Rāhu and Ketu. The belief in the darkening of one star by the other in a battle—*e.g.,* between the sun god Lisa and the moon goddess Gleti in Dahomey—is about as widespread. An eclipse may also be interpreted (as in Tahiti) as the lovemaking of sun and moon, who thus beget the stars and obscure each other in the process.

Beliefs and practices centring on eclipses

Stars and constellations. Worship of the stars and constellations in the modern world survives only in a very corrupt or hidden manner. True star worship existed only among some ancient civilizations associated with Mesopotamia, where star worship was practiced. Mesopotamia, where both astronomy and astrology reached a high degree of refinement—especially after a Hellenizing renaissance of astronomy—was the origin of astral religions and myths that affected religions all over the world. Though the view is controversial, Mesopotamian astral worship and influence may have reached as far as Central and Andean America (by way of China or Polynesia). Sumerian, Elamite, and Hurrian contemplation of the stars influenced not only Mesopotamia, Asia Minor, Egypt, Iran, and India but also other areas. Knowledge of the zodiac, the planets, and observation of precession extended from the West to south Asia—*e.g.,* the Pythagoreans and Orphics (mystical philosophers) in the Mediterranean area and astrological mystical thinkers in India, Indonesia, China, and Polynesia. West Sudan, for example, was deeply influenced by the spirit of ancient Mediterranean and Oriental knowledge of the stars.

The influence of Mesopotamian astral worship

Apart from areas in The Sudan, northeast Africa, and Rhodesia (Mwene Matapa, or Monomotapa), not much of Africa has had any considerable knowledge of the stars. That knowledge of the stars is relatively limited among forest peoples, unless old hunting cultures survived, is explained by an Ekoi tribesman in southeast Nigeria, as follows: "Ekoi people do not trouble themselves about the stars, because the trees always hide them." Hunting pyg-

mies likewise have never achieved any significant knowledge of the stars, which the Bushmen on the steppe have.

Knowledge of the stars in the areas of the true primitive peoples rarely leads to a worship of the stars. True star gods are rare, for example, in large parts of Africa. In Polynesia, where significant knowledge of the stars by the seafaring people and fishermen was learned in regular schools of astronomy, there seldom occurred what can be called true religious worship of the stars. Knowledge of the stars, however, is still relatively significant among the hunting peoples in the Southern Hemisphere, especially among the African Bushmen and Australian Aborigines, who were formerly untouched by the high civilizations. Economic considerations connected with the rising and setting of the stars, however, surpasses their mythological significance by far. The stars are usually considered to be living beings, particularly animals that have been transferred to the sky. They evidently are taken seriously primarily because they indicate by their rising and setting the appearance of game to be hunted or fruits to be collected.

Widespread beliefs in the constellation Orion and the planet Venus

The widespread African interpretation of the constellation Orion as a hunter, as game, or as a dog (from East Africa to the lower Congo and in the area of the Niger) is most likely a vestige from an earlier hunting period that has survived in agricultural civilizations. In a different form, the constellation Orion is still known in Europe as a hunter, in north Asia as a hunter of reindeer and elk, and in North America as a hunter of bears. In South America—outside the Andean empires—a whole series of astral beliefs of the ancient hunting culture has been preserved: the concepts of stars and constellations as Lords of the Animals, as helpers of the hunter, or as animals themselves.

Venus, the best known planet, has probably experienced its most significant personification in the figure of the Mesopotamian goddess Inanna-Ishtar. She was viewed as a being, sometimes female and at other times bisexual. Through her identification with the Greek Aphrodite and the Roman Venus, Inanna-Ishtar, the queen of heaven, still survives in Roman Catholic iconography—*e.g.,* as the Virgin Mary standing on the moon. African cultures also have been significantly impressed by this planet, not only in the rare figure of a Zulu heavenly goddess who determines the agricultural work of the women but even more as the evening and morning star, who are the wives of the moon. In the royal culture of Mwene Monomotapa (Rhodesia) and its influences in Buganda and southern Congo, the king is related to the moon, and his wedding with the Venus women is a type of *hieros gamos* ("sacred wedding"). In large areas of Africa the concept of "Venus wives of the moon" is preserved, although the moon is usually considered as the wife (or sister) of the sun. This concept was most likely prevalent at a time when the moon-king ideology was widespread in the eastern half of Africa from the Nile to South Africa, perhaps indicating south Arabian influences.

The Pleiades, Polaris, and the Milky Way

The Pleiades, a group in the constellation of Taurus (six to seven adjacent stars), is viewed in many parts of the world as maidens who are pursued by men (*e.g.,* hunters). The Pleiades is also interpreted as a mother hen with her chickens, especially in Eurasia, where the star Aldebaran, which is located close to the Pleiades, is often included as a part of the constellation. In Africa the Pleiades designates the beginning of the agricultural year. Therefore, in many Bantu languages the verb *kulima* ("to hoe") furnishes the basis for their designation *kilimia,* the Pleiades. In addition to East and South Africa there is still a smaller area in the western Sudan that retains this belief.

Polaris (the north star) enjoys a central significance among the Finno-Ugric and Turkish Tatars as "nail of the world" or "pillar of heaven." Among Altai Tatars Polaris is viewed as the negotiator of the god of heaven Ülgan; in Japan, Polaris is a god of heaven above the ninth layer of clouds.

The Milky Way, depending on a group's economy and life style, is often simply named after hunting or domestic animals: way of the tapir, the donkey, or the camel. It also is called the seam of the heavenly tent or a water stream. As the footsteps of God or the way of God, as the way of the dead, or as a deserted way of the gods, the Milky Way reveals older mythical conceptions, among which is that of the world (cosmic) tree.

Aurora borealis, the northern lights of the polar regions, is frequently interpreted by Arctic and sub-Arctic peoples (*e.g.,* Eskimo, Athapascan, Tlingit) as the reflection of the dance fire of the ghosts or of the peoples further in the north, as the "cooking of meat," or the ball game of these peoples. Northern Germanic tribes saw in it the splendor of the shields of Valkyrie (warrior women).

ELEMENTS AND FORCES OF NATURE

The natural forces of fire and water, which evidently exclude each other, are brought together in a unity of opposites in the worldviews of early archaic civilizations. Both forces are purifying as well as protective, and are viewed by many as being connected with the cosmic powers of the sun and moon. Where they are truly combined, often genetically, fire (as the sun) is usually male, and water (as the moon), female. Where the fire is included more into the chthonic (earthly) sphere it may also receive a feminine character (*e.g.,* fire in the earth, preserved in the womb); where rain is viewed as the semen of heaven, which is usually personified as male, it takes on a male character.

The "living" qualities of water

Water. Many of the qualities of water make it appear to be animated; on this basis it is psychologically understandable that water (*e.g.,* rain, sea, lakes, and rivers) might become a natural phenomenon worthy of worship. Water is always in motion, changes color in the light of the stars, reflects the world and man, "speaks" with murmuring and roaring, brings new life to dried out vegetation, refreshes men and animals, the tired and the ill, and heals. Because it dissolves dirt it is also most suitable for purifying the soul (*e.g.,* after the violation or the commission of a sin of any kind). Under certain circumstances, even pictures (icons) of gods have to be washed. Water also demonstrates destructive forces (seaquakes, floods, and storms). The most important mythical-religious facts symbolized by water are the following: the primal matter; the instrument of the purification and expiation; a vivifying force, a fructifying force; and a revealing and judging instrument.

Water as primal matter. The conception of a primal body of water from which everything is derived is especially prevalent among peoples living close to coasts or in river areas—*e.g.,* the Egyptian Nu (the primordial ocean) and the Mesopotamian Apsu (the primeval watery abyss) and Tiamat (the primeval chaos dragon). The earth may be fished out or emerges from the primeval water; heavenly beings (*e.g.,* Ataentsik, ancestress of the North American Iroquois) appear on the emerged earth; and birds lay an egg that is later divided into two halves (heaven and earth) on the chaotic sea. Thus, water is viewed as the foundation of all things. A survival of the original primeval sea, in such myths, is the water that flows around the earth's disk (*e.g.,* Oceanus).

Water as an instrument for purification and expiation. Water is viewed as an instrument for purification and expiation, especially in arid areas. Cultic acts, in such areas, generally take place only after lustrations, sprinkling, or immersion in water. The same view holds true for entry into new communities or into life (*e.g.,* baptism). Water lustration is especially necessary after touching the dead, and as a purificatory washing for priests and kings. Pictures of the gods also have had water poured over them.

Myths of a great flood (the Deluge) are widespread over Eurasia and America. This flood, which destroys with a few exceptions a disobedient original population, is an expiation by the water, after which a new type of world is created.

Water equated with life forces in the body

Water as a vivifying force. Water is viewed as vivifying, like the heavenly rainwater that moistens the earth. Water also is equated with the flowing life forces of the body (*e.g.,* blood, sweat, and semen). In order to replace the lost liquids, water was added to the mummified dead in Egypt. The African Ashanti designate their patrilinear groups as *ntoro,* which means water, river, and semen, and the Wogeo of Papua call their patrilinear clans *dan; i.e.,* both water and semen. The myth of the Kasuar ancestress of the But of Papua related how Kasuar's blood became sea (and salt).

Water as fructifying. Wherever early archaic culture spread the myth of the world parents heaven and earth, there also was a belief that heaven fructifies the earth with heaven's seed. The springs, pools, and rivers on the earth, therefore, may bring not only healing and expiation, but also fertility. The Scamander River in ancient Greece evidently was so personified; according to Aeschines, a 4th-century-BC Greek orator, girls bathed in it before marrying and said: "Scamander, accept my virginity." Magical rites in which water serves as a substitute for semen or the fertility of men are numerous.

At the corn festival Nsiã of the Bamessing (in Cameroon), which is celebrated in the dry season, the festival begins with the mourning of the dead vegetation. Reminiscent of the Egyptian Osiris and the Mesopotamian Tammuz festivals, the Nsiã festival emphasizes that the god who gave the nourishment has died and is being mourned like a chieftain. The chief, dying symbolically with the god, has to be strengthened with a miraculous "chieftain water," which has to be fetched by virgins of the chieftain's clan. For two weeks the chieftain drinks from the gourds of all the maidens after the women of the tribe have drunk from the holy water place.

Battles of gods and heroes with mythical beings, beasts, and monsters that hold back the fructifying water are widespread in mythology. The liberation of water during the mythical battle is equivalent to the end of the dry season or a drought, to the reviving of vegetation. In Indian mythology Indra slays Urtra; in Syrian and Palestinian mythology Baal battles with Leviathan; and in Huron (North American Indian) mythology Joskeha, the spring hero, kills the frog that attempted to restrain the water from flowing freely.

Water as a revealing or judging instrument. Water also serves as an instrument that reveals and judges. Reflections in the water led to a whole series of oracles originating from an alleged prophetic or divinatory power of water. A visionary look into the water surface was believed to reveal the future as well as past misdeeds. This ancient custom may have been preserved in the use of crystal balls by modern fortune tellers. The custom of water divination is found in ancient Europe, North Africa, the Near East (*e.g.,* Babylonian fortune telling by means of cups), eastern and northern Asia (where the use of metal mirrors by the shamans often replaces the water as a divining means), and in Southeast Asia and Polynesia. Where such means of divination were severely repressed, as in sub-Saharan Africa, these methods of mirror and water gazing were changed into manipulated water ordeals.

Water is also used as a judging element: in ordeals believed to demonstrate the judgment of the gods, water ordeals (*e.g.,* immersion in water), as well as the more frequent fire ordeals, appear. Here also the purifying character of the water plays a role.

Fire. Worship of fire is widespread, especially in areas where the earthly fire is believed to be the image of the heavenly fire.

Fire as a living force or power

Because of various psychological reasons, fire is considered to be a personified animated or living power: it moves vehemently, devours, and becomes hungrier; it spreads fast into a giant blaze and is red like human blood and warm like the human body. It makes the plants that it has devoured suitable for fertilizing the earth; it shines brightly in the night and, by transference, may have "eternal life" or by constant rekindling can be made into a "perpetual fire." In cremation it separates the body from the soul; it drives away predatory animals and insects that cause pestilence.

Its chief functions are similar to those of its main adversary, water: to purify and to ward off evil, especially from home and hearth. Fire magically drives away rain, but with its smoke it also attracts rain clouds during a period of drought. Fire is believed to have both heavenly and earthly origins: it is brought by lightning, and it lives in the volcano of the underworld.

Stories are told of ancestors, heroes, or animals of primeval times who purloined the fire from the higher numina (spiritual powers). Bringers of civilization, similar to the Greek god Prometheus, fetch it—often together with fruits of the field, iron, or musical instruments—from heaven. Like Prometheus, Nommo, the primal being among the Dogon in Mali, brings fire and the first fruits of the field down to the earth. Prometheus steals the fire from the blacksmith Hephaestus, but Nommo himself is the first blacksmith. In both areas this cultural achievement is celebrated with annual torchlight parades (in Greece, called Promethea festivals). Elsewhere, birds, or animals, such as the dog (especially in Africa), who is closely allied to the hearth fire of man, are the bringers of fire. Animals often fetch the fire from the Lord of the Animals in the bush.

Where geysers and volcanos indicate that the oldest fire is beneath the surface of the earth, fire is brought forth by animals and heroes. The Maori hero Maui seizes it from his ancestress Mahuike in the depth of the earth and puts it into a tree. Since that time it has been possible to get fire from the wood of the trees (*e.g.,* the fire borer). In areas practicing a definite ancestor worship, hunters obtained the fire from the subterranean world of the dead (as in East Africa). Previous to the Iron Age (15th century–2nd century BC), the generating of fire with the aid of fire borers, or fire saws, was viewed as a sexual act (male and female fire wood), especially in East and South Africa, India, Indonesia, and Mexico. In the creation myths of the Dayak of Borneo, fire is produced by rubbing a liana (male) on a tree (female) and is interpreted as coitus. The Tlingit (of northwest America) tell a story of the magical conception of a girl by the sawdust of the fire borer. The boring for the new state fire in the Loango empire (West Africa) coincides with the public coitus of a young couple.

The generating of fire as a sexual act

This conceptual framework seems to be a late consequence of earlier ideas of fire in the body of humans, especially of women, as a centre of sexual life. Such views are probably most pronounced among the Aborigines of Papua and Australia. The Marind in New Guinea, who, in their myth of the origin of fire view it as being derived from the sexual act, undertake the new boring of fire in connection with a cultic act in which the raping of a girl is the central rite. Elsewhere in New Guinea, there is a concept that fire lies in the genitals of women, especially of the first wóman.

When iron-smelting techniques by means of fire became common among Neolithic (New Stone Age) peoples of similar mentality, as in Indonesia and Africa, the making of iron in shaft furnaces (considered as female) and bellows (male) has been interpreted as coitus with a subsequent birth (especially among the Bantu).

In archaic civilizations with sacral kings, the sacred perpetual fire (*i.e.,* the state fire) of the residences and temples of the royal ancestors was believed to have a phallic element. It was cared for by virgins, who were viewed as wives of the fire. Vestal virgins of this kind are documented in ancient Rome, Mwene Matapa (Africa), and pre-Columbian America. Among the Maya of Central America, an order of fire caretakers was founded by a deified "virgin of the fire." Extinguishing and rekindling of fire at the inauguration of a prince points to the idea of a spirit of the princes in the state fire and also to the cyclical renewal of the state in the purifying fire, which signifies the beginning of a new era.

Fire gods are found less frequently in primitive cultures than in archaic civilizations. That fire gods are not yet known in some ancient Oriental areas is probably because of the advanced development of a sun deity.

Iranian fire worship was derived from the cult of the god Ātar, but it was made a central act in Zoroastrianism (a religion founded by the 6th-century-BC Iranian prophet Zoroaster). Fire worship continues to be practiced among the Parsis (modern Zoroastrians) of India: in temples the sacred fire is maintained by a priest using sandalwood, while his mouth is bound with a purifying shawl; fire in new temples is kindled from the fire of the old temples; household fires are not permitted to go out and are greeted in the morning by the members of the household and offered sandalwood; and Aryan cremation practices are renounced because the fire would be contaminated, and the dead are thus deposited in the "temples of silence."

Weather. The worship of atmospheric powers can only

with difficulty be separated from the worship of heaven. In most cases the high god in heaven is also the god of thunderstorms and rain. Specific gods of wind and storm are found especially in countries with tornadoes and hurricanes (*e.g.,* the Maya deity Huracan). People (*e.g.,* the Tuareg and Arabs) in arid countries and deserts, dried out by the wind, speak of sand funnel spirits or of a desert god, such as the "boneless Kon" of the Peruvians.

The worship of atmospheric powers

From northern Europe to the tropical forests, thunderstorm deities rule heaven and earth. The most famous group of these numina (spiritual beings) are the Indo-European thunder gods (Thor-Donar of the Germans, Taranis of the Celts, Perkunis of the Slavs, Indra of the Indians, Zeus-Jupiter of the Greeks and Romans), who throw their thunderbolts or bundles of lightning. The Finnish god Ukko and the Basque god Orko probably stem from the same root; these gods still continue in the popular beliefs of Eastern Europe or Latin America today, such as St. Elijah or Santiago. These deities are related to the west Asiatic gods Teshub and Hadad (associated with the steer and with lightning) but also to the thunder god Shango of the Nigerian Yoruba, who is accompanied by a ram (as Thor uses a he-goat for pulling his wagon). Shango, as Yakuta, throws thunderbolts (*i.e.,* stone axes) to the earth, as does the Mayan rain god, Chac.

The goat, the ram, or horses appear as companions of weather gods or as animals that pull the thundering sky vehicle. In other cultures thunderbirds are the companions of the thunder gods or are the lightning itself. The lightning bird Zu, or Imdugud, occurs in ancient Mesopotamia, and the Garuḍa (with Wadjra) in Vedic India. Thunderbirds are represented (sometimes with arrows or spears in their bill or fangs) on archaeological artifacts of the Bronze Age in Dodona, Minussinsk in Siberia, Dong Son in Vietnam, and on pots in north Peru; they are described in myths of the Pueblo and prairie Indians and among East and South Africans.

Rain magic

Where prayers or sacrifices to gods and ancestors in the religious cult are not effective in producing rain, rain magic, which is practiced universally in similar rites, is often able to accomplish it. Trained magicians usually perform such rites, but ancestral priests or "persons holding power" also may do so. In rain magic, sprinkling, spitting, or immersion of people or things is often used to call down heavenly moisture. Smoke clouds to attract the rain accomplish the same purpose. There also must be suitable vestments (fresh greens, skins or pelts of water animals), body painting (representing clouds), or adornment with bird down. The colour black in the clothing or on a killed or exposed animal is believed to be especially effective. Animals held responsible for holding the rain or water back (frogs, snakes, or mythological dragons) must be challenged. The sound of rain or thunder is produced with "bull-roarers," whistling, noise pots, rattles, and chains. If excessive rain is to be stopped, the injunction to perform or refrain from certain acts (*e.g.,* the prohibition of washing, boiling water, burning objects, making noise, and whistling) must be observed.

The rainbow often is considered a being, generally in the form of an animal, who swallows and holds back rain or water. The rainbow serpent (as a double bow also conceived as bisexual) is a figure that is found especially in the tropics of Africa, south Asia, north Australia ("ungud" snake), and Brazil. Elsewhere, the rainbow is viewed as a heavenly bridge that connects the worlds of gods and men: the Bifröst bridge in the Edda (a Norse saga); the bridge of the soul boats in Indonesia or of the creator god in Africa; and the path of the Greek goddess Iris. In Christian iconography, the rainbow is the throne of Christ; among Arabs and some Bantu of Central Africa it is the bow of god, and among the Nandi, Masai, and the Californian Yuki it is the robe of god.

WORSHIP OF ANIMALS

Among the numerous animals that are prominent in religion and magic, the wild animals of the forests, the sea, and the air that are most important for the hunter are the most significant. Hunters and collectors, rooted in the earliest cultures of man, believed that they not only had to kill animals—which were important for their economy as nourishment and raw materials—but also that they had to avoid their revenge. The feeling of a close connection between men and animals that has been lost to the highly civilized people (broadly speaking) led to an anthropomorphizing of animals. They were not only considered as living beings but were humanized to such an extent that the borderline between man and animal became virtually nonexistent and animals were held responsible for crises.

The religious magical attitude by which primitive peoples confront animals may be called animalism, regardless of whether the animal is thought to have life (animatism) or to have a soul (animism). See below, *Animism.*

The best known form of what may be called animal worship is totemism. An anthropologist, Sir James Frazer, defined totemism as "an intimate relation which is supposed to exist between a group of kindred people on one side and a species of natural or artificial objects on the other side, which objects are called the totems of the human group" (*Totemism and Exogamy,* 1910). Frazer added that the mentioned "species" generally is more "natural" than "artificial." In this "intimate relation" the animal has the prerogative and in true totemism this relationship represents a more or less reverent attitude of man toward his animal partner. In totemism the relationship with the animal has not only merged into the realm of religion but has at the same time become a phenomenon of social life. See below, *Totemism.* (Ed.)

Animism

Animism is the belief in innumerable spiritual beings concerned with human affairs and capable of helping or harming men's interests. Animistic beliefs were first competently surveyed by Sir Edward Burnett Tylor in a work, *Primitive Culture* (1871), to which is owed the continued currency of the term. While none of the "great" religions of the world is animistic (though they may contain some animistic elements), most of the "little" religions, those of the tribal (or "primitive") peoples, are. For this reason an ethnographic understanding of animism, based on field studies of the tribal peoples, is no less important than a theoretical one, concerned with the nature or origin of religion.

IMPORTANCE IN THE STUDY OF MAN AND RELIGION

Definition and limits of the term animism

The term animism denotes not a single creed or doctrine but a view of the world consistent with a certain range of religious beliefs and practices, many of which may survive in more complex and hierarchical religions. Modern scholarship's concern with animism is coeval with the problem of rational or scientific understanding of religion itself. After the age of exploration, Europe's best information on the newly discovered peoples of the Americas, Africa, Asia, and Oceania often came from Christian missionaries. While generally unsympathetic to what was regarded as "primitive superstition," some missionaries in the 19th century developed a scholarly interest in beliefs that seemed to represent an early type of religious creed, inferior but ancestral to their own. It is this interest that was crystallized by Tylor in *Primitive Culture,* the greater part of which is given over to the description of exotic religious behaviour. To the intellectuals of that time, profoundly affected by Darwin's new biology, animism seemed a key to the primitive mind—man's intellect at the earliest knowable stage of cultural evolution. Present-day thinkers consider this view to be rooted in a profoundly mistaken premise. All contemporary cultures and religions are regarded as comparable in the sense of reflecting a fully evolved human intelligence capable of learning the arts of the most advanced society. The religious ideas of the "stone-age" hunters interviewed in this century have been far from simple.

Since the "great" religions of the world have all evolved in historic times, it may be assumed that animistic emphases dominated the globe in the prehistoric era. In societies lacking any doctrinal establishment, a closed system of beliefs was less likely to flourish than an open one. There is, however, no ground for supposing that polytheistic

and monotheistic ideas were excluded. But what is plain today—that no historically given creed has an inevitable appeal to the educated mind—had scarcely gained a place in scholarly argument 100 years ago.

THEORETICAL ISSUES

Tylor's theory of animism. For Tylor the concept of animism was an answer to the question, "What is the most rudimentary form of religion which may yet bear that name?" He had learned to doubt scattered reports of peoples "so low in culture as to have no religious conceptions whatever"—he thought religion was present in all cultures, properly observed, and might turn out to be present everywhere. Far from supposing religion of some kind to be a cornerstone of all culture, however, he entertained the idea of a pre-religious stage in the evolution of cultures and believed that a tribe in that stage might be found. To proceed in a systematic study of the problem, he required a "minimum definition of religion" and found it in "the Belief in Spiritual Beings." If it could be shown that no people was devoid of such minimal belief, then it would be known that all of mankind already had passed the threshold into "the religious state of culture."

But, if animism was ushered in as a "minimum definition," it became the springboard for a broad survey. Although anthropology in Tylor's day was mainly an "armchair" science, through wide and critical reading he had developed a good sense for what was credible in the ethnographic sources of his day. He assembled an array of cases and arranged them in series from what seemed to him the simplest or earliest to the most complex or recent. In this way he taught that religion had evolved from a "doctrine of souls" arising from spontaneous reflection upon death, dreams, and apparitions to a wider "doctrine of spirits," which eventually expanded to embrace powerful demons and gods. A fundamental premise was "that the idea of souls, demons, deities, and any other classes of spiritual beings, are conceptions of similar nature throughout, the conceptions of souls being the original ones of the series." Tylor asserted that men everywhere would be impressed by the vividness of dream images and would reason that dreams of dead kin or of distant friends were proof of the existence of souls. The simple belief in these spiritual beings independent of natural bodies would, he thought, expand to include more elaborate religious doctrines, accompanied by rites designed to influence powerful spirits and so control important natural events.

While Tylor offered no special theory for this expansion and so avoided most of the traps of early social evolutionism, he taught that cultures moved, though not along any single path, from simpler to more complex forms. The direction of movement was shown by the survival of animism in muted but recognizable forms (including most "superstitions" and many expressions such as "a spirit of disobedience" or common words like "genius") in the advanced civilization of his own day. This "development theory" he championed against the so-called degradation theory, which held that the religion of remote peoples could only have spread to them from centres of high culture, such as early Egypt, becoming "degraded" in the process of transfer. Tylor showed that animistic beliefs exhibit great variety and often are uniquely suited to the cultures and natural settings in which they are found.

In retrospect, Tylor seems more balanced in his judgments than later writers who constructed the problem of "minimal religion" in a narrower frame. Tylor's greatest limitation was self-imposed, since he narrowed his attention to what may be called the cognitive aspects of animism, leaving aside "the religion of vision and passion." Tylor took animism in its simplest manifestation to be a "crude childlike natural philosophy" that led men to a "doctrine of universal vitality" whereby "sun and stars, trees and rivers, winds and clouds, become personal animate creatures." But his cognitive emphasis led him to understate the urgent practicality of the believer's concern with the supernatural. Tylor's men are "armchair primitives" (the creatures of "armchair anthropologists"), not real men caught in the toils of discord, disease, and fear of perdition.

Counter theories. Tylor thought the idea of the human soul must have been the elementary religious idea and the model for all other supernatural beings. Later scholars, responding to evidence of simpler beliefs that yet entailed a properly religious awe toward the sacred, began to debate the probability of a "pre-animistic stage" of theological evolution. Corresponding to this turn in religious studies was a shift in anthropology toward a concern with "primitive thought" and, in particular, the explanation of religion as intellectual error. Émile Durkheim, a French sociologist, in his *Elementary Forms of the Religious Life* (1915), held that religion originated in totemism, conceiving that identification with a totem animal could result from an irrational projection of men's expectations of security in the bosom of society. He thought such collective projections were more solidly based in the human condition than the "hallucinations" (dreams) that Tylor had supposed must lead to the ideas of soul and supernatural being. Durkheim has been criticized for not seeing totemism as only one animistic cult among many, with special implications as an organizational schema but not "elementary." (See below, *Totemism.*)

The search for a pre-animistic stage

English and German theorists conceived the invention of religion in more pragmatic terms: attempting to extend his control of nature beyond the limits his crude science imposed, man had invented supernatural power—magic. The most prolific scholar of this persuasion was Sir James G. Frazer, who argued in his massive work *The Golden Bough* (1890–1915) that "the magic art" had arisen as a pseudo-science, probably had achieved universality before the emergence of religion, and was more firmly rooted than religion in men's beliefs. He thought intelligent men had become disillusioned with earthly magicians and had invented infallible ones—gods. Frazer's work ranged over classical mythology and savage custom without distinction. Finding parallel traditions everywhere, he compiled a massive testament to the psychic unity of mankind. The myriad structures of both magic and religion that he surveyed all could be reduced to transparent intellectual error. The apparent mystery of religion was virtually explained away, for, if magic had become man-centred religion and religion god-centred magic, there was little to choose between them.

The predominance of human over supernatural spiritual powers

Since Frazer accepted the common notion that the sign of religion is man's humility before the gods and held that magic put man in the ascendancy, it followed that wherever the supernatural beings could be tricked, bribed, or otherwise mastered the system of beliefs was magical. This obscured Tylor's clear "minimum definition of religion" and threw an odd light on what he had called "lower animism," the belief in spirits that men with ritual knowledge can master. Any self-confident ritual act—for example, the Eskimo hunter's ritual control of game spirits or the shaman's cure of a grave affliction—had become magical and so transparently egoistic. The result for an ensuing generation of anthropologists was loss of focus upon the religions encountered in the field. Tylor had found animistic beliefs generally devoid of ethical content even when centred in men's urgent needs. Frazer, interjecting the image of the primitive magician with illusions of unlimited power, made it difficult to grant animistic religion even Tylor's minimum of dignity.

Frazer had identified Melanesia as an area in which, by his terms, magic dominated over religion. An admirer, Bronisław Malinowski, was able to accomplish a thorough ethnographic study of the Trobriand Islanders in Melanesia during World War I, and it was Malinowski who dominated European ideas about the intimate life of "primitive man" in the following decades. Viewing his islanders within the frame of ideas Frazer had provided, Malinowski pictured them as secular in outlook. In numerous works on the Trobriand Islanders, published through the 1920s and 1930s, there was to be scarcely a mention of religion as such. The belief in spirits appeared only as mythical background to magical practices connected with gardening and seafaring and with a ceremonial cycle in which the competition for prestige was dominant. The effect was to reduce religion to its pragmatic and social aspects, thus de-emphasizing the very peculiarities of human belief and

experience that first attracted men such as Tylor to the study of "primitives."

ANIMISTIC PHENOMENA IN THEIR SOCIAL CONTEXTS

While it is futile to seek cases of animism in "pure," "minimal," or "elementary" form, some social contexts are undeniably simpler than others, and it may be tempting to suppose that the religions found in those contexts would follow suit. On that principle, however, nomads such as the Australian Aborigines might be supposed (as they were supposed by Durkheim) to enjoy an uncomplicated religious life, but this is emphatically not the case. What complicates Australian religions is an elaborate ceremonialism not usually found in nomadic societies. Ceremonialism generally can be treated as an emphasis in the area of expressive behaviour, usually consistent with the animistic world view and unlikely to displace it. While it is an emphasis most common among agriculturists, its presence among nomads is by no means confined to Australia. Though there is no reason to suppose that ceremony is of any more recent origin than any other way of expressing man's relation to the spirit world, animistic religions (religious systems in which animism plays an essential role) can be sorted into those with and those without a ceremonial emphasis, and, in this formal sense, the latter are the simpler. The salient characteristic of all animistic religions is their particularism, a quality opposite to the universalism of the "great religions," which conceive man as subject to global powers and personal destiny.

Ceremonialism and particularism

Particularism is evident in the number and variety of spirits recognized and in the peculiar scope attributed to each. The pre-Christian Lapps of Scandinavia have sometimes been called fetishists because they propitiated nature spirits as well as personally named gods and demons. The nature spirits were generally benevolent and always localized. They could be addressed in particular objects, such as stones or posts, which men would set up in likely places. The few personally venerated spirits (gods) were identified with thunder, sun, moon, hunting, childbirth, and the winds. Evil spirits might be incarnate in animal or monstrous forms and could cause disease or other misfortune. In Minnesota–Ontario the Ojibwa world was animated by a great number of eternal spirits (manitous), all of about equal rank, represented in trees, food plants, birds, animals, celestial bodies, winds, and wonders of every description. Besides these esteemed spirits were other categories, which were dreaded: ghosts, monsters, and the windigo (a crazed man-eating ogre), who brought madness (a cannibalistic psychosis). The list of creatures, places, attributes, and events that could be treated as totems in Australia would be quite similar. The Buryat of Lake Baikal in Siberia, living on the fringes of empire (Mongolian and Chinese), developed an elaborate social order and viewed the spirit world as the twin of their own, organized in the same way into noble, commoner, and slave ranks. At death a man passed over to the other world, assuming his proper rank and acquiring fresh power over men, which he might exercise well or ill in accordance with his character in life. Evil men, as it were, became devils and great men gods.

In particularistic religions there is a range of spirits, from sojourning ghosts and mortal witches to perennial beings, whose natures and dispositions to man are attributed by categories (*e.g.*, mermaids and leprechauns are both usually pictured as irresponsible), but in action individual spirits are independent of one another. If some spirits may be called gods, they do not constitute a ruling pantheon, for men do not conceive that any supernaturals enjoy comprehensive control of events. In animism, spirits represent particularistic powers and must be handled accordingly. Typically, men's primary emphasis is on avoidance of trouble, and this is the meaning of the many taboos and propitiatory observances, of an almost mechanical nature, that abound in some societies. When trouble is at last encountered, the responsible witch, demon, or disgruntled spirit must be identified, and this is the task of the diviner. The cure may rely upon ritual cleansing, propitiation, or even the overpowering of the malevolent force through supernatural counteragency—the specialty of the shaman.

The primary emphasis: avoidance of trouble

Judging that an animal will not mind being killed if it is not offended ritually, Eskimos take various precautions before, during, and after the hunt. The rationale lies in the belief that animal spirits exist independent of bodies and are reborn: an offended animal will later lead his companions away so that the hunter may starve. If, in spite of their precautions, men are left without game, a shaman may be called to discover the transgression that has offended an animal spirit—or perhaps he may find that he must do battle with a malevolent being controlled by a rival shaman willing the community harm.

Ceremonialism, when its emphasis is upon feasting, exchange, and display, may be secular in its emphasis, as is the case in much of Melanesia and New Guinea; or if religious, it may be associated with totemic or ancestral cults, as in Australia or Africa, the expressive emphasis of which is on social ties rather than on the quality of relations between men and the supernaturals. Finally, ceremony may be used directly to dramatize the role of the spirits in men's lives, as it is by the Pueblo peoples of Arizona–New Mexico. At their height, the Pueblo ceremonial cycles were as rich as any in the world. Supernaturals were elaborately impersonated by the dancers, and the human condition was portrayed as one of dependency. But, for all this, particularism was not greatly compromised. The supernaturals were many and were represented in a realistic manner emphasizing their differences from ordinary men. The style was that of mummery and conjuring, consciously put on by grown-ups as a sort of morality play. There was no sense of incongruity in the fact that neighbouring pueblos cultivated other sets of spirits. In some pueblos, separate clan societies had complete charge of the ceremonial calendar and formally controlled communication with the supernatural, even to selecting the member who might be curer in case of an illness. But such a step toward ecclesiasticism in a very small community could not greatly affect its animistic premises, and witchcraft prevailed without the blessing of the ceremonial societies.

When the fullness and versatility of all these religions is considered, without any need to press them into simplified categories or evolutionary stages, it can be seen that openness, not narrowness, of doctrine is a general feature of animism. Wherever it is found, it is a grass-roots religion, not a doctrinaire one imposed from above. Ecclesiasticism may coexist with animism, as in China or Burma, where there are no pre-eminent gods whose universal claims presuppose mastery of the whole supernatural world. But the most likely context of animism is an uncentralized social order in which secular power is not developed and each local settlement is at the focus of its own world.

Openness of doctrines

THE ANIMISTIC WORLD VIEW

Part of the conceptual difficulty experienced both in anthropology and in the history of religions, when animism is to be placed among other systems of belief, springs not from the early association of animism with a speculative theory of religious evolution but directly from the huge variety of animistic cults. As a category, Tylor's concept is more general than either polytheism or monotheism, and its meaning is harder to delimit—the word applies broadly to most of the "little religions" but suggests nothing of their varieties. For this reason, much use is made of subordinate labels, such as shamanism, totemism, or ancestor propitiation. These cults do not, in any case, constitute the whole religion of a people. They are, however, institutions that are not bound to one culture area—an Australian totemic cult does bear a "family resemblance" to an African one, though their differences also are many. Shamanism, with its reliance on ecstasy, is found from Greenland to Mysore, and the propitiation of ancestors is not restricted to Africa and the Far East. It has long been recognized that the frequent recurrence of institutions fitting a certain pattern implies that there is a radically limited number of possible patterns, and, in this case, the premises of animism evidently have imposed the limitation. Animism attributes importance to categories of supernatural being whose individual members are attached to particular places and persons or resident in particular creatures and are autonomous in their dealings. In such a system, each human

encounter with the supernatural must work itself out as a distinct episode. Even where ceremonialism emphasizes an enduring moral relationship to certain supernaturals, men are likely to conceive of alternative powers to which they might seek at need. In a crisis, loyalties may shift: in West Africa, gods have been sold to neighbouring villages, and, in Melanesia, a vision of European trade goods has inspired a series of new millenarian cults. The quality of openness lends itself to change and eclecticism, hardly ever to religious chauvinism.

A communication with supernatural beings about practical life needs

Animistic creeds have in common an undertaking on the part of men to communicate with supernatural beings, not about metaphysics or the dilemmas of the moral life but about urgent practicalities: about securing food, curing illness, and averting danger. It is characteristic that genuine worship of a supernatural hardly is found. Creator gods often appear in myth but not in cult. In ancestor cults the most recently dead are the most vividly conceived—the original clan ancestor, for all his symbolic importance, is remote both from men and from godhead. If animistic spirits anywhere exercise authority, they do so in particularistic, even egoistic fashion, sanctioning men for ritual neglect or breaking taboos, not for acts of moral neglect or secular offense. Animistic religions do not readily coalesce with systems of political authority and probably do not favour their development. When it is asked whether the association of animism with smaller and simpler societies proves it the natural (original) religion, the answer can only be that it is not known (and perhaps not knowable) what a prehuman or panhuman religion would be like. The problem is as difficult as reconstructing protohuman speech. If religion is taken as a pattern of serious relations between men and supernaturals, then societies devoid of religion have not been found, and it may perhaps be concluded that religion is usually close to the vital centre of a culture, where the credibility of institutions is determined. The view of all nature as animated by invisible spirits—be it shades, demons, fairies, or fates—with which men could cope in meaningful ways may belong to the past, but philosophies that attribute powers of initiative and responsiveness to nature have not gone out of currency. The lesson of the study of animism is perhaps that religion did not arise, as some of Tylor's successors believed, out of *Urdummheit* ("primal ignorance") or delusions of magical power but out of men's ironic awareness of a good life that they are unable, by earthly means, to grasp and hold. Animistic beliefs have everywhere engaged men's susceptibility to private vision and enabled them to cope with it at the level of accepted meaning. (G.K.P.)

Totemism

Totemism is a system of belief in which man is believed to have kinship with a totem or a mystical relationship is said to exist between a group or an individual and a totem. A totem is an object, such as an animal or plant that serves as the emblem or symbol of a kinship group or a person. The term totemism has been used to characterize a cluster of traits in the religion and in the social organization of many primitive peoples.

Totemism is manifested in various forms and types in different contexts, especially among populations with a mixed economy (farming and hunting) and among hunting communities (especially in Australia); it is also found among tribes who breed cattle. Totemism can in no way be viewed as a general stage in man's cultural development; but totemism has certainly had an effect on the psychological behaviour of ethnic groups, on the manner of their socialization, and on the formation of the human personality.

Origin of the term totem

The term totem is derived from *ototeman* from the language of the Algonkian tribe of the Ojibwa (in the area of the Great Lakes in eastern North America); it originally meant "his brother–sister kin." The grammatical root, *ote,* signifies a blood relationship between brothers and sisters who have the same mother and who may not marry each other. In English, the word totem was introduced in 1791 by a British merchant and translator who gave it a false meaning in the belief that it designated the guardian spirit of an individual, who appeared in the form of an animal—an idea which the Ojibwa clans do indeed portray by their wearing of animal skins. It was reported at the end of the 18th century that the Ojibwa name their clans after those animals that live in the area in which they live and appear to be either friendly or fearful. The first accurate report about totemism in North America was written by a Methodist missionary, Peter Jones, himself an Ojibwa chief, who died in 1856 and whose report was published posthumously. According to Jones, the Great Spirit had given *toodaims* ("totems") to their clans; and because of this act, it should never be forgotten that members of the group are related to one another and on this account may not marry among themselves.

Generally speaking, totemistic forms are based on the psychomental habits of the so-called primitives, on a distinctive "thought style" which is characterized, above all, by an "anthropopsychic" apprehension of nature and natural beings, for instance, ascribing to them a soul like man's. Beasts and the things of nature are again and again thought of as "persons," but mostly as persons with superhuman qualities.

THE NATURE OF TOTEMISM

It is advisable to define totemism as broadly as possible but concretely enough so that some justice can be done to its many forms. Totemism is, then, a complex of varied ideas and ways of behaviour based on a world view drawn from nature. There are ideological, mystical, emotional, reverential, and genealogical relationships of social groups or specific persons with animals or natural objects, the so-called totems. It is necessary to differentiate between group and individual totemism. These forms exhibit common basic characteristics, which occur with different emphases and not always in a complete form. The general characteristics are essentially the following: (1) viewing the totem as a companion, relative, protector, progenitor, or helper—superhuman powers and abilities are ascribed to totems and totems are not only offered respect or occasional veneration but also can become objects of awe and fear; (2) use of special names and emblems to refer to the totem; (3) partial identification with the totem or symbolic assimilation to it; (4) prohibition against killing, eating, or touching the totem, even as a rule to shun it; and (5) totemistic rituals.

General characteristics

Though it is generally agreed that totemism is not a religion, in certain cases it can contain religious elements in varying degrees, just as totemism can appear conjoined with magic. Totemism is frequently mixed with different kinds of other beliefs—the cult of ancestors, ideas of the soul, beliefs in powers and the spirits. Such mixtures make the understanding of particular totemistic forms difficult. The cultic veneration of definite animals and natural things and powers by all those who belong to an ethnic unit do not belong to totemism itself.

Group totemism. Group (social or collective) totemism is the most widely disseminated form of totemism. Though the following characteristics can belong to it, they must not be taken to be part of a whole system: (1) mystic association of animal and plant species, natural phenomena, or created objects with unilineally related groups (lineages, clans, tribes, moieties, phratries) or with local groups and families; (2) hereditary transmission of the totems (patrilineal or matrilineal); (3) names of groups that can be based either directly or indirectly on the totem (the same holds true for personal names used within groups); (4) totemistic emblems, symbols, and taboo formulas are, as a rule, a concern of the entire group, but they can also belong to subdivisions of that group. Taboos and prohibitions can apply to the species itself or they can be limited to parts of animals and plants (partial taboos instead of partial totems). (5) Totems for groups are sometimes connected with a large number of animals and natural objects (multiplex totems) whereby a distinction can be made between principal totems and subsidiary ones (linked totems). Totems are associated or coordinated on the basis of analogies or on the basis of myth or ritual. (Just why particular animals or natural things—which sometimes possess absolutely no recognizable worth for

Characteristics of group totemism

the communities concerned—were selected as totems is often hard to fathom and may be based on eventful and decisive moments in a people's past which are no longer known.) (6) Accounts of the nature of totems and the origin of the societies in question are informative, even if they are sometimes valuable only as supplementary rationalizations; they are especially informative with regard to their presuppositions. If, for example, one group supposes that it is derived directly or indirectly from the totem, this may be recounted (as a rationalization) that an animal progenitor was changed into a human being who then became the founder of the group or that the ancestral lord of the group was descended from a conjugal union between a man and a representative of the animal species. Groups of men and species of animals and plants can also have progenitors in common. In other cases, there are traditions that the human progenitor of a kin group had certain favourable or unfavourable experiences with an animal or natural object and then ordered that his descendants had to respect the whole species of that animal.

Group totemism is now found especially among peoples in Africa, India, Oceania (especially in Melanesia), North America, and parts of South America who farm rather than simply gather food from nature. Peoples with hunting and partly harvesting economies who exhibit this form of totemism include, among others, the Australian Aborigines (hunters who occupy a special position due to the many forms of totemism among them), the African Pygmies, and various tribes of North America—such as those on the northwest coast (predominantly fishermen), in parts of California, and in northeast North America. Moreover, group totemism is represented in a distinctive form among the Ugrians and west Siberians (hunters and fishermen who also breed reindeer) as well as among tribes of herdsmen in north and Central Asia.

Individual totemism. Individual totemism is expressed in an intimate relationship of friendship and protection between a person and a particular animal or a natural object (sometimes between a person and a species of animal); the natural object can grant special power to its owner. Frequently connected with individual totemism are definite ideas about the human soul (or souls) and conceptions derived from them, such as the idea of an alter ego and nagualism—from the Spanish form of the Aztec word *naualli,* "something hidden or veiled"—which means that a kind of simultaneous existence is assumed between an animal or a natural object and a person; *i.e.,* a mutual, close bond of life and fate exist in such a way that in case of the injury, sickness, or death of one partner, the same fate would befall the other member of the relationship. Consequently, such totems became most strongly tabooed; above all, they were connected with family or group leaders, chiefs, medicine men, shamans, and other socially significant persons. In shamanism, an earlier trait of individual totemism is often ascertained: the animalistic protective spirits can sometimes be derived from individual totems. To some extent, there also exists a tendency to pass on an individual totem as hereditary or to make taboo the entire species of animal to which the individual totem belongs. In this can perhaps be seen the beginning of the development of totems that belong to a group. Many tales about the origins of the group totem could, perhaps, point in this direction.

Nagualism

Individual totemism is widely disseminated. It is found not only among the tribes of hunters and harvesters but also among farmers and herdsmen. Individual totemism is especially emphasized among the Australian Aborigines.

SOME EXAMPLES OF TOTEMISM

Wiradjuri. (New South Wales, Australia) Totem clans are divided among two subgroups and corresponding matrilineal moieties. The group totem, named "flesh," is transmitted from the mother. In contrast to this, individual totems belong only to the medicine men and are passed on patrilineally. Such an individual totem is named *bala,* "spirit companion," or *jarawaijewa,* "the meat (totem) that is within him." There is a strict prohibition against eating the totem. Breach of the taboo carries with it sickness or death. It is said: "To eat your *jarawaijewa* is the same as if you were to eat your very own flesh or that of your father." The medicine man identifies himself with his personal totem. Every offense or injury against the totem has its automatic effect upon the man who commits it. It is a duty of the totem to guard the ritualist and the medicine man while he is asleep. In the case of danger or the arrival of strangers, the animal goes back into the body of the medicine man and informs him. After the death of the medicine man, the animal stands watch as a bright flickering light near the grave. The individual totem is also a helper of the medicine man. The medicine man emits the totem in his sleep or in a trance so that it can collect information for him. Finally, black magic (sorcery) is also practiced by the medicine man; by singing, for instance, the medicine man sends out his totem. To kill an enemy, the totem enters the chest of the enemy and devours his viscera. The transmission of the individual totem to novices is done through the father or the grandfather who, of course, himself is also a medicine man. While the candidate lies on his back, the totem is "sung into" him. The blood relative who is transmitting the totem takes a small animal and places it on the chest of the youngster. During the singing, the animal supposedly sinks slowly into his body and finally disappears into it. The candidate is then instructed on how he has to treat the animal that is his comrade, and he is further instructed in song and the ritual concentration that is necessary to dispatch the totem from his body.

Individual totems

By courtesy of the Australian News and Information Bureau, New York



Australian Aborigines dancing a corroboree. In dancing a corroboree the Australian natives imitate birds, animals, fish, and men, and also the movements of storms and floods.

Nor-Papua. (Murik Lakes, west of the mouth of the Sepik River, north New Guinea) Patrilineal, exogamous groups (consanguineous sibs) are spread over several villages and are associated with animals, especially fish. They believe that they are born from totems and they make them taboo. Children are given an opportunity to decide during their initiation whether they will respect the paternal or maternal totem. Each group of relatives has a holy place to which the totem animal brings the souls of the dead and from which the souls of children are also believed to come. Totem animals are represented in various manifestations, as spirit creatures in sacred flutes, in disguises, and in figures preserved in each man's house. At the end of initiation ceremonies, the totems are mimicked by the members of the group.

Iban. (Sungai Dayak, Sarawak, Malaysia) Among these peoples, individual totemism is clearly discernible. Particular persons dream of a spirit of an ancestor or a dead relative; this spirit appears in a human form, presents himself as a helper and protector, and names an animal (sometimes one of the natural objects) in which he is manifested. The Iban then observe the mannerisms of animals and recognize in the behaviour of the animals the embodiment of their protector spirit (*ngarong*). Sometimes, members of the tribe also carry with them a part of

such an animal. Not only the particular animal, but the whole species of the animal is given due respect. Meals and blood offerings are also presented to the spirit animal. Young men, who wish to obtain such a protector spirit for themselves, sleep on the graves of prominent persons or seek out solitude and fast so that they may dream of a helper spirit. Actually, only a few persons can name such animals as their very own. Individuals with protector spirits have also attempted to require from their descendants the respect and the taboo given the animal representing the spirit. As a rule, such descendants do not expect special help from the protector spirit, but they observe the totemistic regulations anyway. Thus it can be concluded that particular families or groups of relatives of the Iban represent totem communities.

Birhor. (Munda-speaking hunter and harvester tribe that resides in the jungle of Chotanāgpur Plateau, northeast Deccan, India) The Birhor are organized into patrilineal, exogamous totem groups. According to one imperfect list of 37 clans, 12 are based on animals, 10 on plants, eight on Hindu castes and localities, and the rest on objects. The totems are passed on within the group, but no account of their origin is available. From tales about the tribe's origins, it appears that the totem had a fortuitous connection with the birth of the ancestor of the clan. The Birhor think that there is a temperamental or physical similarity between the members of the clan and their totems. Prohibitions with regard to taboo are sometimes cultivated to an extreme degree. In regard to eating, killing, or destroying them, the clan totems are regarded as if they were human members of the group. Moreover, it is believed that an offense against the totems through a breach of taboo will produce a corresponding decrease in the size of the clan. If a person comes upon a dead totem animal, he must smear his forehead with oil or a red dye, but he must not actually mourn over it; he also does not bury it. The close and vital relationship between the totem and the clan is shown in a definite ceremony: the yearly offering to the chief spirit of the ancestral hill. Each Birhor has a tradition of an old settlement—thought to be located on a hill in the area. Once a year, the men of each clan come together at an open place. The elder of the clan functions as the priest who gives the offering. A diagram with four sections is drawn on the ground with rice flour. In one of these, the elder sits while gazing in the direction of the ancestral hill. The emblem of the particular totem is placed in one of the other places of the diagram; depending on the circumstances, this emblem could be a flower, a piece of horn or skin, a wing, or a twig. This emblem represents the clan as a whole. If an animal is needed for such a ceremony, it is provided by the members of another clan who do not hold it as a totem. The Birhor show great fear of the spirits of the ancestral hill and avoid these places as far as possible.

Group totems

Kpelle. (Liberia, West Africa) In this society, there is not only group totemism but individual totemism as well. Both categories have the same designations, namely, "thing of possession," "thing of birth," "thing of the back of men." These phrases express the idea that the totem always accompanies man, belongs to him, and stands behind him as a guide and warner of dangers. The totem also punishes the breach of any taboo. The totems are animals, plants, and natural phenomena. The kin groups that live in several villages were matrilineal at an earlier time, but they are beginning to exhibit patrilineal tendencies. The group totems, especially the animal totems, are considered as the residence of the ancestors; they are respected and are given offerings. Moreover, a great role is played by individual totems that, in addition to being taboo, are also given offerings. Animal personal totems can be transmitted from father to son or from mother to daughter; on the other hand, individual plant totems are assigned at birth (plants of a tree of life for the child) or later. The totem also communicates magical powers. It is even believed possible to alter one's own totem animal; further, it is considered an alter ego. Persons with the same individual totem prefer to be united in communities. The well-known leopard confederation, a secret association, seems to have grown out of such desires. Entirely different groups produce patrilineal taboo communities which are supposedly related by blood; they comprise persons of several tribes. The animals, plants, and actions made taboo by these groups are not considered as totems. In a certain respect, the individual totems in this community seem to be the basis of group totemism.

A SHORT HISTORY OF TOTEMISTIC THEORY

From McLennan to Lévi-Strauss. There are a number of theories or hypotheses concerning totemism. Many of them are marked by methodological deficiencies, preconceived ideas, and a prejudiced selection of source documents; nevertheless, some of these theories contain points of view that deserve consideration.

The first theory was proposed by the Scottish ethnologist John Ferguson McLennan. Following the vogue of 19th-century research, he wanted to comprehend totemism in a broad perspective, and in his study "The Worship of Animals and Plants" (1869, 1870) he did not seek to explain the specific origin of the totemistic phenomenon but sought to indicate that all of the human race had in ancient times gone through a totemistic stage.

In 1899 McLennan's theories were criticized by E.B. Tylor, an English anthropologist who rejected the confusion of totemism with mere worship of animals and plants. Tylor claimed to find in totemism the tendency of the human spirit to classify the world and its things. He thus viewed totemism as a relationship between one type of animal and a clan. But he was opposed to the idea of seeing totems as the basis of religion.

Views of E.B. Tylor, Andrew Lang, and James G. Frazer

Another Scottish scholar, Andrew Lang, early in the 20th century advocated a nominalistic meaning for totemism, namely that local groups, clans, or phratries, in selecting totem names from the realm of nature, were reacting to a need to be differentiated. If the origin of the names was forgotten, there followed a mystical relationship between the objects—from which the names were once derived—and the groups that bore these names. Lang wanted to explain the relationship through nature myths according to which animals and natural objects were considered as the relatives, patrons, or ancestors of the respective social units. Thoughts by the tribes on these matters led eventually to taboos. Group exogamy first originated in the formation of totemistic associations.

The first comprehensive work on totemism was *Totemism and Exogamy,* published in 1910 in four volumes by the British anthropologist Sir James George Frazer. It presented a meritorious compilation of the then known worldwide data on the subject.

Basing his view on research done among primitives in Australia and Melanesia, Frazer saw the origin of totemism as one possibility in the primitive interpretation of the conception and birth of children ("conceptionalism"). According to this primitive idea, women become impregnated when a spirit of an animal or a spiritual fruit enters into their wombs. Since the children therefore participate in the nature of the animal or plant, these plants or animals take on significance. These ideas were hereditary and resulted in the beginning of totem clans derived from a particular natural creature.

A Russian-American ethnologist, Alexander Goldenweiser, subjected totemistic phenomena to sharp criticism. This critical work had lasting importance, especially in the United States, where it engendered a skeptical attitude concerning totemism. Goldenweiser saw in totemism three phenomena that could exist singly and actually coincided only in the rarest of cases. These phenomena were: (1) clan organization; (2) clans taking animal or plant names or having "emblems" obtained from nature; and (3) belief in a relationship between groups and their totems. Goldenweiser did not perceive these phenomena as a unity, since any of them could exist apart from the others.

In another treatise published in 1910, a German ethnologist, Richard Thurnwald, claimed to recognize in totemism the expression of a specific way of thinking among the primitives. Primitives judge the natural environment according to its external appearance without analyzing it any closer and assume that there are sympathetic connections and combinations of natural things; from these

ideas come lasting rules of behaviour (like taboos, respect, and social relationships). For the psychology of totemism, Thurnwald later (1917–18) put forth a detailed, systematic presentation; by means of concrete examples, he also raised questions about the connections of totemism with ancestor worship, notions of souls, belief in power, magic, offerings, and oracles.

Views of Émile Durkheim

The founder of a French school of sociology, Émile Durkheim, in a general work concerning the elementary forms of religion (1912), also examined totemism from a sociological and theological point of view. Durkheim hoped to discover a pure religion in very ancient forms and generally claimed to see the origin of religion in totemism. For Durkheim, the sphere of the sacred is a reflection of the emotions that underlie social activities, and the totem was, in this view, a reflection of the group (or clan) consciousness, based on the conception of an impersonal power. The totemistic principle was then the clan itself, and it was permeated with sanctity. Such a religion reflects the collective consciousness that is manifested through the identification of the individuals of the group with an animal or plant species; it is expressed outwardly in taboos, symbols, and rituals that are based on this identification.

In further contributions, Goldenweiser in 1915–16 and 1918 criticized Lang, Frazer, and Durkheim and insisted that totemism had nothing to do with religion; that man in no way viewed his totem as superior to himself or as a deified being but viewed it as his friend and equal. Goldenweiser also rejected Frazer's thesis of "conceptionalism" as an explanation of totemism. On the other hand, Goldenweiser was of the opinion that all totemistic manifestations do have at least something of a kind of religion, but he was not inclined to include the guardian spirit conception within totemism.

In 1916, an American ethnologist, Franz Boas, posited a theory of totemism as an "artificial" unity, existing only in the thinking of ethnologists. For Boas, totemism exhibited no single psychological or historical origin; since totemistic features can be connected with individuals and all possible social organizations, and they appear in different cultural contexts, it would be impossible to fit totemistic phenomena into a single category. Boas was against systematizing and thought it senseless to ask questions about the origins of totemism.

Views of F. Graebner and B. Ankermann

The first theoretician of the Vienna school of ethnology, Fritz Graebner, attempted to explain the forms of both individual totemism and group totemism and designated them as a moderately creedal or semireligious complex of ideas according to which individual members or subgroups of a society are thought to be in an especially close (but not cultic) relationship to natural objects. According to Graebner, with the help of the cultural–historical method, one can establish (1) the extent to which totemistic forms belong to one definite cultural complex, (2) which forms are "older" or "younger," and (3) the extent to which forms belong together genetically. Graebner tried to work out a "totemistic" complex (a "culture circle") for the South Seas. This complex entailed a patrilineal group totemism as well as the material, economic, and religious elements that, in his opinion, appear to be combined with the totemism in that area.

Another member of the same school, Bernhard Ankermann, in 1915–16 championed the view that all totemisms, regardless of where they are found, contained a common kernel around which new characteristics are built. As seen from the standpoint of what was found in Africa, this kernel appeared to him to be the belief in a specific relationship between social groups and natural things—in a feeling of unity between both—a relationship he believed to be spread throughout the world, even if only in a modified or diminished form. Magical and animalistic ideas and rites are merged with totemism in a strong inseparable unity. The genesis of this type of relationship presupposes a state of mind that makes no distinction between man and beast. Although magic can be closely connected with totemism, the feeling of unity between man and beast has nothing to do with magic, which was connected with it only later. According to Ankermann, the totems are not something perilous, something to be shunned, but, on the contrary, totems are something friendly; and since this is directly due to kinship, a totem is thought to be like a brother and is to be treated as such. The totemistic taboo is believed to be due to the fact that the totem is a relative. Ankermann was inclined to see the formation of totemism in a "lower" form of hunting, in an emotional animal–man relationship, in animalistic behaviour. Men of early times, he thought, might have imitated those animals that attracted their attention most of all. According to Ankermann, pretension and reality, however, for the "primitives" are blurred into one thing. Primitive man identifies himself with the animal while he is imitating it; the habit of so doing could lead to a continuing identification. Early man imitated all animals that interested him, but he imitated those that shared his place of habitation above all.

Views of Wilhelm Schmidt

In 1915–16 Wilhelm Schmidt, then the leader of the Vienna school of Ethnology, viewed totemism strictly according to the then-existing schemes of culture circles (today long abandoned); because totemism was disseminated throughout the world, he thought of it as a closed cultural complex in spite of local differences. He maintained that the differences in totemism shown by earlier theories are exaggerations and could, moreover, be due to the lack of particular elements of totemism, to the loss of certain forms of totemism, to incursions from the outside, or to different stages of the development of totemism, none of which would exclude a unified origin for all of totemism. Schmidt believed that the cultural–historical school of ethnology had produced proof that the older, genuine totemism occurred as an integral part of a culture located in a definite area and that it was "organically" connected with definite forms of technology, economy, art, and world view. From a "pure" totemism, Schmidt wanted to separate similar forms, such as sex and individual totemism. Moreover, though he did not designate totemism as a religion, he saw that it did have some sort of religious meaning. Schmidt (in opposition to Ankermann) wanted to regard the higher form of hunting as the economic basis for the totemistic "culture circle."

Views of A.R. Radcliffe-Brown and Bronisław Malinowski

The leading representative of British social anthropology, A.R. Radcliffe-Brown, took a totally different view on the totemistic problem. Like Boas, he was skeptical of the reality of totemism. In this he opposed the other pioneer of social anthropology in England, Bronisław Malinowski, who wanted to admit the reality of totemism in some way and looked at it more from a biological and psychological point of view than from an ethnological one. According to Malinowski, totemism was not a cultural phenomenon but was the result of trying to satisfy basic human needs within the natural world. As far as Radcliffe-Brown was concerned, totemism was composed of elements that were taken from different areas and institutions, and what they have in common is a general tendency to characterize segments of the community through a connection with a portion of nature. In opposition to Durkheim's theory of sacralization, Radcliffe-Brown took the point of view that nature is introduced into the social order rather than secondary to it. At first, he shared with Malinowski the opinion that an animal becomes totemistic when it is "good to eat." He later came to oppose the usefulness of this viewpoint since many totems—such as crocodiles and flies—are dangerous and unpleasant.

In 1952, when Radcliffe-Brown rethought the problem, he found that the similarities and differences between species of animals are to a certain degree translated into ideas of friendship and conflict, or close relationships and opposition among people. The natural world is represented in the form of social relationships to the extent that these social relationships become valid in primitive societies. The structural principle which Radcliffe-Brown believed he had discovered at the end of his comparative study is based on the fusion of the two contrary ideas of friendship and animosity. Thus totemism speaks in its own way of interrelationships and antitheses, ideas that are also found in moieties. So totemism is formulated as a general problem in which the contrasts in nature serve to create an integral whole. Thinking in terms of opposing

things is, according to Radcliffe-Brown, an essential structural principle for evaluating totemism.

Views of Claude Lévi-Strauss

The most incisive critique of totemistic phenomena, one that denied the reality of totemism, was supplied by the French ethnologist Claude Lévi-Strauss in *Le Totémisme aujourd'hui* (English translation, *Totemism*, 1963). As a chief representative of modern structuralism, Lévi-Strauss was especially stimulated by Radcliffe-Brown, whose views he further attempted to expand. Lévi-Strauss believed that he was to approach the apparent, acknowledged difficulties in the study of totemism from the viewpoint of a study of structure. In order to study the structure of totemism, Lévi-Strauss devised a scheme to illustrate the abstract polarities that he saw in totemism as a phenomenon in human culture. This scheme was implemented in a table of oppositions or polarities, or mutual relationships. The basic opposition, or relationship, was between nature and culture. On the one hand, there were in nature certain natural realities such as species of animals or plants and specific animals or plants. On the other hand, there were in culture various groups and individuals who identified themselves with particular species or with specific animals or plants. Lévi-Strauss distinguished four kinds of opposition, or relationship, between nature and culture within totemism: (1) a species of animal or plant identified with a particular group, (2) a species of animal or plant identified with an individual, (3) a particular animal or plant identified with an individual, and (4) a particular animal or plant identified with a group.

According to Lévi-Strauss, each of these four combinations corresponds to the phenomena that are to be observed in one people or another. The first antithesis holds good, for example, for the Australians, for whom natural things are associated with cultural groups (moieties, sections, subsections, phratries, clans, or the association of persons from the same sex). As an example of the second combination, there is the individual totemism of North American Indians, in which a person is correlated with a species of nature. For the third type of combination, Mota in the Banks Islands of Melanesia is cited: the individual child is thought of as the incarnation of a particular animal, plant, or natural creature that was found and consumed by the mother at the time that she was conscious of her pregnancy. For the fourth type of correlation, Lévi-Strauss cited examples from Polynesia and Africa where definite individual animals formed the object of group patronage and veneration.

Lévi-Strauss also critiqued the findings of A.P. Elkin, a specialist on Australia, where totemism had already played a special role in the formation of theories and where it exhibits an abundance of forms for expressing totemism. Elkin had differentiated the following forms: (1) individual totemism; (2) social totemism—*i.e.*, totemism that is in a family, moiety, section, subsection, patrilineal clan, or matrilineal clan; (3) cultic totemism with a religious content that is patrilineal and "conceptional" in form; (4) dream totemism—totemistic content in dreams—found in social or individual totemism. Elkin denied the unity of totemism, but (according to Lévi-Strauss) wanted to preserve its reality on the condition that he might trace it back to a multiplicity of types. For Elkin, there is no longer "one" totemism but many totemisms, each in itself a single irreducible whole.

In connection with the Australian material, Lévi-Strauss argued that matrilineal clan totemism—that passed on the "flesh" or "blood"—and the patrilineal clan totemism—based on dreaming—were in no way heterogeneous but were to be thought of as being mutually complementary. They were different means of connecting the material and spiritual world; they were two different, but correlative, types that express the relationship between nature and society.

Lévi-Strauss concluded that not the similarities but the dissimilarities correspond to the so-called totemism. Such a pattern was clearly expressed in the basic model of the contrasts of the natural with the cultural (that were outlined above). Depending on the ideas of Radcliffe-Brown, Lévi-Strauss claimed to perceive antithetical thinking as a crucial structural principle in totemism and believed that the similarity among totemistic ideas in various cultures lay in similarities between both systems of differences—those documented in the natural sphere and those in the culturally defined social groups. Lévi-Strauss concluded that the distinction between the classes of man and animal serves as the conceptual basis for social differences. For Lévi-Strauss, totemism is therefore an "illusion" reduced to a form of thinking, and this so-called totemism is connected with understanding the demands that it answers as well as the way in which it seeks to satisfy those demands. Since it is a "logic that classifies," totemism in this sense has nothing of the archaic itself. Its picture is projected onto the material (the natural phenomena), not taken from it. It does not take its substance from without.

Totemism as "illusion"

Present situation and emerging trends. From the publications of Lévi-Strauss and the contributions of his predecessors, it is obvious that difficulties stand in the way of an adequate interpretation of the intricate profusion of totemistic phenomena. But it seems fair to many authorities to ask whether it is possible to dispose of totemism simply as an illusion, whether the very abstract structural interpretation of the facts is actually legitimate. To those who question the position, it seems clear that even though all totemistic forms of expression can hardly be seen under one common denominator, reality cannot be totally denied to totemism. A specific relationship between man and nature, one that serves as a basic scheme of classification, seems to be at the basis of all the various forms of totemism. Indeed, this can be regarded as the prevailing characteristic of totemism in the form in which it manifests itself. A special problem, however, must be taken into consideration: since totemism can be connected with different ideas and practices, of religious, magical, or ideological natures, it is difficult to decide what is "totemistic" and what is "nontotemistic." (Jo.H.)

Ancestor worship

The term ancestor worship describes, in a broad and loose sense, a variety of religious beliefs and practices concerned with the spirits of dead persons regarded as relatives, some of whom may be mythical. Although far from universal, ancestor worship exists or formerly existed in societies at every level of cultural development.

NATURE AND SIGNIFICANCE

The core of ancestor worship is the belief in the continuing existence of the dead and in a close relation between the living and the dead, who continue to influence the affairs of the living. Beliefs in a surviving element of the human person (*e.g.*, the soul) and in an afterlife have been held in all kinds of societies. Attitudes toward the spirits of the dead vary from love, respect, and trust, mingled with special feelings of reverence, to outright fear; the attitudes are sometimes ambivalent. The spirits of the dead are often thought to help the living, but they often are thought to do harm if they are not propitiated. All societies give ritual attention to death or to the souls of the dead, but not all of these practices may appropriately be called ancestor worship. If death itself, rather than the ancestral relationship, is the focus of attention, the name death cults is more appropriate. The deification of dead heroes is similarly poorly distinguished from ancestor worship. Death cults, the worship of dead heroes who may or may not be regarded as ancestors, and clearly distinguishable rites of ancestor worship may all exist in the same society.

Ancestors venerated by elaborate rites are those persons who in their lifetimes held positions of importance, such as heads of families, lineages, clans, tribes, kingdoms, and other social groups. Depending on the manner in which kin are organized into social groups, ancestral spirits that are worshipped may be limited to one sex, or may include both sexes. Among primitive societies that trace descent only through male lines, for example, the titular positions of prestige are held by males, and only male ancestors are significant.

Ancestral spirits that are worshipped also vary in nearness or remoteness in time from the living. In some societies only the spirits of the recently deceased are given atten-

tion; in others, all ancestors, near and remote in time, are included. In still other societies, one ancestor, real, honorary, or mythical, may be the focus of attention, and he is often regarded as a hero.

BASIC PATTERNS AND FUNCTIONS

Importance of kinship

The presence or absence of ancestor worship relates in a general way to the importance of kinship in the societies concerned. In the primitive world society is ordered and life made possible through bonds of kinship, though intergenerational continuity of kinship extending to deceased forebears may not be regarded as important. Among societies of higher levels of cultural complexity, the importance of kinship and the size of actively functioning kin groups decreases. Where continuity of kinship and inheritance of property are very important, elders are characteristically regarded with respect, and the persistence of bonds of affinity with ancestors is favoured, as was the case in traditional China and Japan. Societies in which the only important kin group is the nuclear family composed of parents and immature children, and where economic support as well as emotional well-being among adults do not depend upon kinship, are ill-suited for the development or maintenance of ancestor worship. Illustrative examples are modern China and Japan, where sociocultural changes, brought about by adaptation to Western civilization and modern technology, have included a great decline in the importance of kinship and the size of kin groups, and where traditional practices of ancestor worship have correspondingly declined.

Ancestor worship includes all of the attitudes and acts usually associated with the worship of nonancestral gods and spirits. According to some scholars and theorists, ancestral spirits are anthropocentric conceptions similar to other supernatural beings; that is, the spirits have the qualities of personality and the capabilities of man, to which supernatural potency is added. The spirits see, hear, feel, understand, and communicate with the living; they make moral judgments; they are wishful, willful, joyful, angry, stern, permissive, kind, cruel, and sometimes capricious; and they have all the other emotions and traits of human beings. All of the behaviour and practices that are customary with regard to other kinds of supernatural beings are found in rites of ancestral worship—veneration and propitiation in the forms of prayers, offerings, sacrifices, the maintenance of moral standards, and festivals of honour that may include pageantry, music, dance, and other forms of art. Where ancestral spirits directly control the affairs of the living, their continued favour is sought by established periodic rites, and their special aid may be requested at times of crisis. Perhaps the only truly distinctive ritual acts of ancestor worship are commemorative ceremonies, held annually or at other fixed intervals, and tendance of graves, monuments, or other symbols commemorating them (see also RITES AND CEREMONIES: *Death rites and customs*).

Motives for acts of piety toward ancestors are diverse, and they differ from devotional acts toward other gods or spirits principally in reflecting the idea that the spirits continue in some measure to be kin and are active participants in the life of the community. Rituals directed toward ancestral spirits maintain communion with them in ways that reflect human regard for the deceased elders and desires to aid them in their spiritual existence. These rites and devotional acts also seek to gain spiritual and practical benefits for the living. The powers the ancestral spirits are believed to possess vary greatly from society to society, as do the powers of other supernatural beings. Their powers may be weak or strong, generalized or specific. In some societies, their supernaturalistic roles include that of being intermediaries between living relatives and the gods. Where neglected ancestral spirits are thought to be harmful to the living, the goals of ritual observances may include or emphasize the desire for protection from them. Whether ancestral spirits are themselves gods with powers or are intermediaries, communion with them is a form of transcendence of ordinary states of existence, which may be a conscious or unconscious goal of the acts of devotion.

ANCESTOR WORSHIP IN RELIGIONS OF THE WORLD

Until the 19th and 20th centuries ancestor worship in various forms and of varying importance in total religious complexes was widely but irregularly distributed throughout the world. In most societies, however, it was only one element of a complex of supernaturalism, and seldom a dominant feature. The spread of European culture weakened, displaced, or otherwise put an end to ancestor worship in most nonliterate societies, and technological, social, and ideological changes discouraged its continuation in culturally advanced societies.

In nonliterate societies. Among nonliterate societies, well-developed ancestor cults are limited principally to peoples of sub-Saharan Africa, Melanesia, and some tribal groups of India and adjacent parts of Asia. The greatest development was in Africa, where ancestral spirits are commonly an important part of the roster of supernatural beings. In the aboriginal kingdoms and near-kingdoms of sub-Saharan Africa, the spirits of kings and paramount chiefs were often regarded as generalized ancestors and were venerated by all members of society. Spirits of the heads of clans, often a mythical couple, were also worshipped, as were the spirits of founders of lineages and of deceased heads of individual families. Ancestral spirits of kings and high chiefs often were believed to have power over matters of concern to the entire society, such as rain and the growth of crops and cattle, whereas spirits of heads of families, lineages, and clans influenced matters of immediate concern to the particular social groups. Acts of piety were numerous and included sacrifice, prayers, and hospitable celebrations that honoured the spirits by storytelling and other forms of entertainment. The spirits were generally regarded as very helpful to their living descendants and were propitiated in established cyclic ceremonies as well as at times of crisis when help was needed.

Ancestral spirits in Africa and Melanesia

In Melanesia the spirits of the dead generally were held to be important, and in some societies were the focus of much attention. An outstanding example is the ancestor cult of the Manus of the Bismarck Archipelago, where Sir Ghost, the spirit of the living male head of the household, was the tutelary god of the family and supervised the behaviour of its members. Only spirits of the newly dead were worshipped, and when the head of a household died, the old tutelary god was discarded. The skull of the deceased household head was placed above the entrance in the dwelling, where it watched the conduct of all within, giving rewards and punishments in accordance with their deeds, and protected the family from the malign influences of the guardian spirits of other families.

Among the Trobriand Islanders near New Guinea, the dead had two spirits, one of which was a harmless ghost that vanished a few days after death. The other spirit, *baloma*, had an eternal existence as an ancestral spirit abiding in another world. Death rites were clearly distinguishable from ceremonies directed to the *baloma*. Communion was maintained with these ancestral spirits, who returned to their villages when annual feasts were held by their living relatives, and also appeared in dreams and trances. Seers with special powers of communication with the supernatural world brought news of the spirits by visiting them in person in the land of the dead. The *baloma* were propitiated but were never feared. They were invoked and believed to provide supernatural aid in numerous acts of magic, especially in connection with the raising of the crops upon which the livelihood of the people depended.

Elsewhere in nonliterate societies, ancestral spirits sometimes were important, but nowhere were they the lone or primary supernatural beings. In aboriginal Polynesia, where people of high social status were regarded as descendants of the gods, the spirits of kings and high chiefs had power to help men, but they were not the objects of worship to any great extent. Indian tribes of North and South America seldom gave much ritual attention to ancestors. Among tribes of the present-day United States, for example, the greatest ritual elaboration was among the Pueblo tribes of the Southwest, whose complex ceremonial calendar included impressive rites honouring generalized ancestral spirits. These ancestors were impersonated in

ritual and appealed to for aid in providing rain, bountiful crops, and general well-being.

Ancestor worship in China and Japan

In Eastern societies. Among the civilizations of Asia, the classic examples of ancestor worship have been China and Japan. In both societies, however, reverence for, rather than worship of, ancestors is a more nearly accurate description of the beliefs and practices. In China the ancestor cult is extremely ancient and emphasized continuity of familial lines. Reverence for elders was an act of filial piety strongly supported by the teachings of the sage Confucius of the 6th–5th centuries BC. The family was viewed as a closely united group of living and dead relatives rather than a group of individuals. Unity of the larger kin group was also stressed through devotional acts at clan temples that honoured all ancestral spirits. Rites of reverence were held in the home, at temples, and in graveyards. Ancestral shrines containing tablets bearing the names of recent ancestors and especially notable forebears were maintained in the homes, and rites were observed before them. Temple rites were also observed; funerals and commemorative ceremonies were elaborate, and custom also called for pilgrimages to graves. Motives for performing rites involved concern for the welfare of the ancestors, who were thought to require solicitous care and, since the kin group was an indivisible unit with common goals and fortunes, a means of continuing to receive the aid and cooperation of the deceased relatives.

The early background of Japanese ancestral veneration is obscure, but most of the historically known practices are adaptations of Chinese customs. Some sort of ancestor cult may have existed in the native Shintō religion before the diffusion of Buddhism from China in the 6th century. When Buddhism came to Japan it was a comprehensive religious system. With the passage of time and in coexistence with the Shintō religion, Japanese Buddhism began to emphasize death rites and commemorative ceremonies, and Shintō became more concerned with matters of daily life. Confucianism was never an organized religion in Japan, but quasi-religious Confucian ideals of filial piety were very important and were sometimes incorporated in the teachings of Japanese Buddhist sects, thereby reinforcing ancestral veneration. Japanese rites, like those of China, consisted of elaborate funerals and many commemorative rites at the home, the temple, and the grave. A great annual ceremony honours all spirits of the dead, who return to their homes at that time. Until recent years Shintō rites of passage at death also were conducted in the home.

The state of ancestor worship in modern China is unclear, but it may be disappearing. In modern Japan ancestors have declined in importance, and Buddhist ritual tends to emphasize funerals, giving less attention than formerly to later commemorative ceremonies.

Ancestor worship in India

In India the vast, locally variable, and unorganized complex of theology and rites of the Hindu religion conspicuously includes ceremonies honouring ancestors, but the cult of ancestors is nevertheless a relatively small part of the full religious system. Characteristically, funeral rites are very elaborate and have many motifs of supernaturalism, among which is attention to ancestral spirits. The practices relating to ancestors reflect ideas concerning reincarnation and the system of castes, which are in turn closely intertwined. One of the manifest goals of the funeral rites is to guide the spirits of the deceased during the perilous time between death and rebirth. The idea that one's moral behaviour in life determines his fate in the next life is also connected with caste affiliation. Sinful behaviour will bring rebirth in a low caste. Among some castes and in some regions of India, annual rites are performed in honour of the spirits of the ancestors of the male heads of household, who are believed to give spiritual aid in promoting the growth of crops and in other important matters of everyday life.

In ancient Middle Eastern and European societies. Ancestor worship in various forms existed among the ancient civilization of the Mediterranean, where cults of the dead sometimes also existed, and among later European peoples. Ancient Egyptian religion featured a cult of the dead but gave little attention to ancestral spirits except to those of royalty, which were venerated by the people and especially honoured in rites observed by their royal descendants. In ancient Babylonia a cult of the dead also existed, and among the members of the ruling class ancestral spirits were honoured by festive rites and sometimes deified. Beliefs and practices of late Zoroastrianism (a religion founded by the 6th–7th-century-BC Iranian prophet Zoroaster) included rites for the spirits of the dead, who were believed to have power in the affairs of the living. In ancient Greece ancestor worship overlapped with hero worship. Some ritual attention was given to spirits of household heads and political leaders, and the spirits of men whose deeds were heroic were sometimes elevated to immortality and made the objects of rites of reverence. In ancient Rome ancestor worship was a familial cult activity. Ancestral spirits were believed to have influence on mortal life and to return to visit their relatives, when rites were held in their honour.

Ancestor worship in Greece, Rome, and other ancient European societies

Among various northern and eastern peoples of Europe, ancestral spirits also held some importance. Ancient Celts, Teutons, Vikings, and Slavic groups conducted rites of propitiation and sacrifice.

THEORIES AND INTERPRETATIONS

The 19th-century sociologist Herbert Spencer regarded fear and consequent propitiation of the souls of ancestors as the earliest form of religion, an interpretation that later scholars set aside as unverifiable. Reflecting the decline of ancestor worship among societies of the world, modern scholarship has seldom given much concern to this subject in isolation but has instead followed the trend of the social sciences in considering ancestor cults in relation to other elements of religious complexes, the social order, and the whole of culture. Early writings often expressed the idea that all people fear death, but this idea is questioned by Hindu thought, in which extinction rather than eternal life is the ultimate goal. It is generally held that all peoples have some beliefs of an afterlife. Such beliefs are presumed by anthropologists to be very ancient, as seems evident from Paleolithic burials dating from perhaps as much as 60,000 years ago containing stone tools, shells, and other objects, and the abundance of later prehistoric burials with grave goods.

Modern interpretations of ancestor worship

No recent study of wide comparative scope attempts to interpret the significance of ancestor worship, and modern interpretations of these practices view them as having essentially the same social and psychological value as other beliefs and practices of supernaturalism. Through their symbolic representations of kinship and of the social hierarchy of kin groups, the beliefs and acts of ancestor worship may be seen as establishing and reinforcing ideas of social roles and identities, thereby contributing to psychological well-being and social harmony. Joint rites promote social solidarity, and characteristically display and thereby reinforce the social order. Where ancestral spirits have power in mortal affairs, they are psychologically and socially significant in ways similar to those of gods and other supernatural beings. In the many matters of life over which man lacks secular control, for example, the intervention of ancestral spirits alleviates anxiety. But, quite like various other beliefs of supernaturalism, ideas about ancestors may also be seen sometimes to instill as well as to allay anxiety. In this connection ancestor cults may have an important moral significance by serving as sanctions of social conformity. In many societies, improper behaviour is usually thought to reflect unfavourably upon both living and dead relatives, whether or not ancestor cults exist. Scholars often interpret the primary significance of ancestor cults in some societies as a sanctioning force. Ancestral spirits are viewed as approving or disapproving the behaviour of their descendants, in a generalized way, without exercising specific sanctions, or they may be regarded as vigilant protectors of morality. Among the Manus of Melanesia, the ancestral tutelary spirit was believed to punish all moral defections by removing the "soul stuff" from the wrongdoer, thereby causing illness. When the offense was serious, death followed unless penitential acts were performed. Where, as among the Manus, ancestral wrath might affect any member of the group for

an offense committed by another member, the sanctioning force is viewed as powerful, operating so that all living members monitor the behaviour of each other.

Special value is attributed to the rites of ancestral reverence in promoting familial solidarity, and, to the extent that such rites are emphasized, in promoting the unity of larger kin groups and entire societies. Information is most abundant on traditional practices of familial ancestor worship in China and Japan. In these societies in former times the individual was submerged in the family, and rites of ancestral reverence may thus be viewed as both reflections and reinforcements of the social order. As acts of supernaturalism, they place a special stamp of approval upon familial roles, unity, and continuity. (E.N.)

Polytheism

Polytheism, the belief in many gods, characterizes virtually all religions other than Judaism, Christianity, and Islām, which share a common tradition of monotheism, the belief in one God. Sometimes above the many gods a religion will have a supreme creator and focus of devotion, as in certain phases of Hinduism (there is also the tendency to identify the many gods as so many aspects of the Supreme Being); sometimes the gods are considered as less important than some higher goal, state, or saviour, as in Buddhism; sometimes one god will prove more dominant than the others without attaining overall supremacy, as Zeus in Greek religion. Typically, polytheistic cultures include belief in many demonic and ghostly forces in addition to the gods, and some supernatural beings will be malevolent; even in monotheistic religions there can be belief in many demons, as in New Testament Christianity. Polytheism can bear various relationships to other beliefs. It can be incompatible with some forms of theism, as in the Semitic religions; it can coexist with theism, as in Vaiṣṇavism; it can exist at a lower level of understanding, ultimately to be transcended, as in Mahāyāna Buddhism; it can exist as a tolerated adjunct to belief in transcendental liberation, as in Theravāda Buddhism.

THE NATURE OF POLYTHEISM

Categories of attitudes toward the gods

In the course of analyzing and recording various beliefs connected with the gods, historians of religions have used certain categories to identify different attitudes toward the gods. Thus, in the latter part of the 19th century, the terms henotheism and kathenotheism were used to refer to the exalting of a particular god as exclusively the highest within the framework of a particular hymn or ritual; *e.g.*, in the Vedic hymns (the ancient sacred texts of India). This process often consisted in loading other gods' attributes on the selected focus of worship. Within the framework of another part of the same ritual tradition, another god may be selected as supreme focus. Kathenotheism literally means belief in one god at a time. The term monolatry has a connected but different sense; it refers to the worship of one god as supreme and sole object of the worship of a group while not denying the existence of deities belonging to other groups. The term henotheism is also used to cover this case, or more generally to mean belief in the supremacy of a single god without denying others. This seems to have been the situation for a period in ancient Israel in regard to the cult of Yahweh.

The term animism has been applied to a belief in many *animae* ("spirits") and is often used rather crudely to characterize so-called primitive religions. In evolutionary hypotheses about the development of religion that were particularly fashionable among Western scholars in the latter half of the 19th century, animism was regarded as a stage in which the forces around man were less personalized than in the polytheistic stage. In actual instances of religious belief, however, no such scheme is possible: personal and impersonal aspects of divine forces are interwoven; *e.g.*, Agni, the fire god of the Ṛgveda (the foremost collection of Vedic hymns), is not only personified as an object of worship but is also the mysterious force within the sacrificial fire.

Belief in many divine beings, who typically have to be worshipped or, if malevolent, warded off with appropriate rituals, has been widespread in human cultures. Though a single evolutionary process cannot be postulated, there has been a drift in various traditions toward the unification of sacred forces under a single head, which, in a number of nonliterate "primal" societies, has become embedded in a supreme being. Sometimes this being is a *deus otiosus* (an "indifferent god"), regarded as having withdrawn from immediate concern with men and thought of sometimes as too exalted for men to petition. This observation led Wilhelm Schmidt, an Austrian anthropologist, to postulate in the early 20th century an *Urmonotheismus*, or "original monotheism," which later became overlaid by polytheism. Like all other theories of religious origins, this theory is speculative and unverifiable. More promising are attempts by sociologists and social anthropologists to penetrate to the uses and significance of the gods in particular societies.

Besides the drift toward some unification, there have been other tendencies in human culture that entail a rather sophisticated approach to mythological material—*e.g.*, giving the gods psychological significance, as in the works of the Greek dramatists Aeschylus and Euripides and similarly, but from a diverse angle, in Buddhism. At the popular level there has been, for instance, the reinterpretation of the gods as Christian saints, as in Mexican Catholicism. A fully articulate theory, however, of the ways in which polytheism serves symbolic, social, and other functions in human culture requires clarification of the role of myth, a much debated topic in contemporary anthropology and comparative religion.

FORMS OF POLYTHEISTIC POWERS, GODS, AND DEMONS

Celestial, atmospheric, and earthly gods

Natural forces and objects. A widespread phenomenon in religions is the identification of natural forces and objects as divinities. It is convenient to classify them as celestial, atmospheric, and earthly. This classification itself is explicitly recognized in Indo-Aryan religion: Sūrya, the sun god, is celestial; Indra, associated with storms, rain, and battles, is atmospheric; and Agni, the fire god, operates primarily at the earthly level. Sky gods, however, tend to take on atmospheric roles; *e.g.*, Zeus's use of lightning as his thunderbolt.

In the earliest cultural levels, in which hunting and then pastoralism and agriculture are clearly vital, religion exhibits these identifications in rites connected with fertility. The sun's vitality is seen in the cyclical effects of causing things to grow and wither. Moreover, because of its dominance of the world, the sun is often seen as all-knowing, and thus sky gods of various cultures tend to be highly powerful and knowledgeable, if also sometimes rather remote. The sky is also often associated with creation. By contrast the moon is rarely of the same importance (though in Ur, a city of ancient southern Babylonia, the moon god Sin was supreme). The role of the sky god in ensuring food and in providing light and warmth, over against the chaotic effects of darkness, was a theme of various myths of the cosmic drama and was one main reason for the connection in mythic thought between creation and light.

Heavenly divinities have also been influential in the development of astrology, which assigns a special significance to stars and planets. In the Middle East astrology was important but was weakened by monotheism; and in Indian culture it came to be deeply woven into the fabric of both Hinduism and Buddhism. Astrology was influential in the Greco-Roman world and in the astral religion attached to Gnosticism (dualistic sects that emphasized salvation through esoteric knowledge) and other cults of the early Christian Era. Astrology was also elaborated in Central America, for instance in Aztec religion.

Gods of the sky become especially powerful when they take on an atmospheric guise. The association of gods such as Zeus and Indra with storm, as well as with fertility-bearing rain, makes their connection with warfare fairly natural; thus, Indra is the most perfect example of an Indo-Aryan warrior. Many societies, however, have had separate gods of war. The ambivalence of atmospheric deities is paralleled in female counterparts who are both creative and destructive. The combination of sky and earth and the joining of differing cosmic forces are sometimes

represented in the *hieros gamos* ("sacred marriage"); *e.g.*, between Apsu and Tiamat in Mesopotamia, Śiva and Śakti in India, and Gaea and Uranus in Greece. The forces of water and fire are particularly significant in bridging the gap between the earthly and heavenly realms. Fire is manifested not only in the hearth but also in lightning and the sun; water is sometimes connected with the moon. Thus, earthly fire and water can also be seen at work higher in the cosmos.

Important in the development of fertility religion were the "dying and rising" gods, such as Adonis, Attis, Osiris, and Tammuz. Their cults had a new life in the mystery cults of the Greco-Roman world, where the original agricultural significance of the rites was transformed into more personal and psychological terms.

On earth, besides the divine mother out of whose womb plant life has its birth, there are a host of divinities connected with agricultural and pastoral life. In addition, sacred significance is often attached to features of the particular environment in which a given group finds itself. Thus, sacred mountains, such as Olympus in Greece, have their resident deities; a river, such as the Ganges (Ganga), may be divinized. Underground rivers have special significance in connecting with the underworld, or nether regions, which can be important as the place of repose of the dead but also as the matrix for the re-creation of life. Geographical locations can also have cosmic significance; *e.g.*, Delphi, Greece, was known as the navel of the earth. Further, many cultures have gods and goddesses associated with the sea.

Plants and animals as divine forces

Vegetation. In a number of cultures trees are seen as a primordial form of vegetation and have a symbolic connection both with heaven and earth; sometimes they are held to contain spirits, as the *yakṣa*s of Indian tradition. Particular sorts of trees, such as the *aśvattha,* or pipal (sacred fig), are held in special veneration. Among plant deities, however, probably the most important are those connected with cultivated plants, such as maize in Central America and the vine in the Mediterranean world. Notable is the cult of Dionysus, the ecstatic wine god who became one of the most influential objects of devotion in the classical period. The vine linked agriculture and ecstasy. The connection between vegetation and dying and rising gods has already been noted; to some extent such motifs were carried over into Christianity in the notion that the cross was the tree both of death and of new life. One of the most obvious modern survivals in the West of vegetation cults is the use at the winter solstice of mistletoe, symbolizing fertility and continued life.

Animal and human forms. Just as plants can be seen as divine forces, so can types or species of animals. For instance, the cult of the snake is widespread and is especially important in the Indian tradition. The serpent is vital in the Old Testament story of Adam and Eve and appears in the Babylonian *Epic of Gilgamesh* as one who knows the secret of rejuvenation. The snake has a fertility aspect because of its possible phallic significance and because it lives in holes in the life-giving earth. The cult of the monkey is important in India, having its essence in the figure of Hanumān, half monkey and half human. It is possible that such theriomorphic cults (in which gods are represented by various animal forms) have been assisted by rituals in which priests wear masks representing the relevant divinities, a practice that may in turn explain the hybrid half-human form. Examples of the wide variety of animal and living forms in which gods appear include Huitzlipochtli (hummingbird; Aztec); Cipactli (alligator; Aztec); Viṣṇu's avatars, or incarnations (fish, tortoise, boar, man-lion; Hindu); the Rainbow Snake (Australian Aboriginal); Cernunnos (stag god with antlers; Celtic religion); Nandi (bull; Hindu). A figure partly in animal guise found in Les Trois-Frères cave at Ariège, France, may represent a complex lord of the beasts analogous to the supposed Śiva (the destroyer and re-creator in Hindu mythology) figure found at sites in the Indus Valley; while a bird-man figure at Lascaux, France, may depict a priestly representation of a divine being. Thus, theriomorphism seems to have a very ancient pattern. In brief, various cultures have taken existing species in their environment and woven them into the pantheon—partly because of their essential dependence on the animals and partly for other reasons, such as similarities between animal forms and other sacred forces (*e.g.*, the analogy of the lion to the force behind kingship).

Gods as human and men as divine

Because man can enter into living relationship with the supernatural beings that surround and dominate his life, it has always been natural to model the gods as human beings. Such anthropomorphism is most evident in the Greek tradition, in which the Homeric gods are brilliantly and unashamedly human in their passions and thoughts. The human model has been assisted by the representation of the gods in art; for a statue is not just a symbolic representation of a god but often his place of presence and influence. Thus, in a number of cultures, the images are treated as replete with divinity.

Just as gods can be human in character, so men can be conceived as divine, either by becoming identified with deities (*e.g.*, through descent) or by displaying appropriate power. Thus, divine kingship was a not uncommon feature of the ancient Middle East; it was also found in the Roman world, when the emperors were divinized, and in Japan and China, where the emperor was son of heaven. Culture heroes and other significant humans could be elevated to semidivine status or more; *e.g.*, Kuan Ti and other heroes in the Chinese tradition, Rāma and Kṛṣṇa (Krishna) in India. Strictly, the succession of sages known as buddhas and tīrthaṅkaras in the Buddhist and Jain traditions, respectively, were not conceived as divine but came to be objects of a cult. In the Mahāyāna (Greater Vehicle), celestial buddhas and *bodhisattva*s (those vowed to become buddhas) came to be profoundly important for devotional religion; from a functional point of view, the Mahāyāna has operated as a polytheistic system, united, however, under an overarching doctrine of emptiness, or the void (*śūnya*), according to which all things are said to be empty of the characteristics assigned to them. The Theravāda (Way of the Elders) accepted the principle that virtuous followers of the Buddha could be translated in the next life to a heavenly existence in which they would have godlike status (an impermanent status, however, for gods share the universal transitoriness of all living beings), but such gods were scarcely the objects of a cult.

Functional deities. In addition to the various forces operating in nature, various social and other functions are divinized. Thus, the god Brahmā in the Vedic tradition, besides being creator, contains and expresses in personal form the power implicit in the Brahmin class. Again, there are gods of healing, such as Asclepius in Greece, and of seafaring, agriculture, and so on. The most elaborate reflection of human concerns is, perhaps, to be found in the later Taoist pantheon, which provided a heavenly counterpart to the Chinese Imperial court. In a number of societies gods of war, such as Mars (ancient Rome) and Skanda (India); gods of learning, such as Sarasvatī (India); and gods of love, such as Aphrodite (Greece) and Kāma (India), have been important. Even such abstractions as the directions (north, south, east, and west) have been divinized. The fact that these varied entities and relationships have been taken as gods is, perhaps, partly the result of the mythic style of thinking, in which distinctions between natural forces and social conventions are not clearly perceived.

Of special importance regarding human affairs are the gods concerned with death and judgment after death, such as Osiris in ancient Egypt, Yama in India, Hades in Greece, and Hel in pre-Christian Scandinavian religion. There are also gods associated with cemeteries and more generally with patterns of the disposal of the dead.

Supernatural forces

The various gods must be seen against the background of a whole host of spirits, demons, and other supernatural forces prevalent in the environment of pastoral and agricultural communities. Among entities hostile to man are the antigods, very often older gods, such as the Titans in Greece, who have been displaced by later deities or gods worshipped by a people conquered by a new dominant folk. The warfare between the old and new can be woven into dramatic myths of the fight between good and evil. This is well brought out in the major myth of the

Orphic writings: Zeus's son Dionysus-Zagreus was killed and eaten by the Titans, who in turn were destroyed, burned up by Zeus's lightning flash. Man is made of the ashes, and therefore he is a compound of divinity and titanic evil. Purification from this evil brings redemption and release from the round of reincarnation. Sometimes, however, the ambivalence of good and evil is built into the same deity, so that creation and destruction and good and evil are seen as complementing one another.

TYPES OF POLYTHEISM

Greco-Roman religion. By the time of the establishment of the Roman Empire, the Greek tradition was already exerting considerable influence on the Roman, to the extent that once relatively independent traditions became somewhat fused. Equations between gods were freely made: Zeus became Jupiter, Aphrodite became Venus, and so on. Originally Roman *pietas* (sense of duty to the gods) was a good deal less personalized than the relationship to the anthropomorphic gods of the Homeric pantheon and was directed at spirits called *numina.* In addition, the various philosophical systems, such as Epicureanism and Stoicism, provided a more systematic cosmology and sense of man's destiny than traditional polytheism. Influential in the Hellenistic period were mystery cults—such as those of Isis, Cybele, Mithra, and Demeter—which catered more to personal concerns with salvation than did the official and civic cults. Under the mid-4th-century emperor Julian, a last vigorous attempt was made to revive paganism and to restore the cult of the gods over against the widespread grip of Christianity.

Germanic, Scandinavian, Celtic, and Slavic mythologies. The sources for a reconstruction of northern European religion are far better than those for the south Germanic peoples, but there were evidently similarities between the religions. The three main Scandinavian gods were Odin, Thor, and Freyr: Odin (or Wodan) had great magical power and wisdom and was called All-father; Thor (or Donar) was the warrior god; and Freyr was the god of fertility. It is possible that these gods are a reflection of the tripartite division of Indo-European society—priest, warrior, and cultivator. Among other deities, Balder, the dying god who was killed by a mistletoe branch, had a poignant charm. Nordic mythology also carries with it a sense of final doom of the gods, looking to the point when the world will be burned up, before its eventual re-creation.

The pattern of Celtic cults is not easy to decipher because of lack of written records; but the stag-headed god Cernunnos was highly significant in iconography. There was also a variety of ancestral gods and goddesses, including a "great mother" of the type found in fertility cults of the ancient Middle East. Celtic religion had a special reverence for water in such forms as pools and rivers.

The Slavic religions of eastern Europe and Russia are likewise imperfectly known, but they involved worship of a high god who is both a creator and an atmospheric force. Another important figure in Slavic mythology was the war god Svantovit. Finno-Ugrian pre-Christian religion bears some resemblance to the Scandinavian, possibly indicating some mutual influences, while Baltic cults are of Indo-European type.

Political considerations in Egypt

Egypt and the Middle East. The Egyptian pantheon evolved into a complex form; many deities were theriomorphic but were presided over by such great gods as Re, the sun god, and Nut, the sky goddess. Re's transformation as Horus, with a hawk's head, was connected with the Osiris legend. The pharaoh was identified with him as the "living Horus." Despite the attempt of Akhenaton, pharaoh in the 14th century BC, to exalt Aton as the single god, the Egyptian cult remained essentially polytheistic but highly articulated. With the domination of Egypt by the Ptolemies about 10 centuries later, the worship of Sarapis, a hybrid Greco-Egyptian deity, was instituted as a means of binding together the two groups.

Though in Egypt the cause of the rise and fall of gods was partially the political struggles between the major city-states, the Sumerian religion was much less affected by such "earthly" considerations. An, the god of heaven, remained supreme, and such deities as the water god Enki and the air god Enlil were prominent. In Babylon, partly the successor state of Sumer, the most vital god was Marduk, creator of the world and of mankind, and victor over the primeval Tiamat, or chaos, who all but absorbed the older surrounding gods. His story is recounted in the epic *Enuma elish* ("When on High"). In Assyrian religion Marduk was in effect replaced by Ashur; and Ishtar, the mother goddess, was also important. In general, it can be said that Middle Eastern religion stemmed from early Sumerian and Egyptian sources and that the latter eventually had some effect on Hellenistic religion.

Early Indo-Iranian religions. For almost a millennium close relations existed between the Vedic and Iranian religions—from before the time of the Iranian prophet Zoroaster, who reformed the ancient religion in the late 7th and early 6th centuries BC, back to the time of the Vedic religion of the Aryans, who invaded India about 1500 BC. Zoroaster, in his reforms, succeeded in excising the many gods, some of whom were subsumed as qualities of the supreme Ahura Mazdā. The rich pantheon of the Vedic hymns developed into the world of classical Hindu mythology, which was fed by streams other than the Aryan.

Classical and modern Hinduism. Certain gods of no great importance in the Vedic tradition came to dominate classical Hinduism, above all Śiva and Vishnu. The latter was associated with belief in *avatar,* or incarnation. Most male gods in the Hindu pantheon also came to be represented with a female consort, symbolizing the *śakti,* or creative power of the deity. The increasing elaboration of Hindu cults as different groups were absorbed into a systematized social fabric has led to the estimate of as many as 33,000,000 Hindu gods. It has been common practice for devotees to select the form under which the divine is worshipped, and such a deity is called the *iṣṭadevatā.* Most Hindus are inclined to interpret the many gods as being symbols of the one divine reality.

Buddhism. Buddhism's tolerance of popular cults, provided that the main essentials of the faith are maintained, means that in most Buddhist cultures several gods are worshipped. In Mahāyāna Buddhism, increased devotion to the Buddha became elaborated as a belief in many celestial beings, notably Amitābha and Avalokiteśvara (Kwannon), who were, however, in essence all unified in the absolute (*śūnya,* the void). In Tibet, a synthesis between the indigenous religion and Buddhism was established. The most notable feature of this form of Buddhism, known as Vajrayāna (Vehicle of the Thunderbolt), was the use of divine forms to symbolize the various factors of existence, such as the different elements making up human personality.

Far Eastern religions. In ancient China the cult of Heaven and ancestor worship were elements woven into the system of Confucianism. Numerous lesser deities were worshipped in popular Chinese practice, and the dividing lines between Confucianism, religious Taoism, and Buddhism were hard to draw. In Taoism, an elaborate pantheon was evolved, modelled in part on the Imperial bureaucracy, and was presided over by the Jade Emperor (Yü-Huang). Other deities included atmospheric gods, gods of locality, and functional gods (of wealth, literature, agriculture, and so on). The Taoist gods were in part a response to the richness of Mahāyāna myth, with its cults of celestial buddhas and *bodhisattvas.*

The Taoist pantheon

The religions practiced in China influenced Japanese culture, which took over some main elements of Confucianism and Buddhism, that interacted with the indigenous polytheistic religion, Shintō (Way of the Gods). The divinities of Shintō tend to be connected with natural forces and localities; the most important deity is Amaterasu, who is the sun goddess and divine ancestress of the emperor.

Religions of ancient Meso-America. The Aztec culture, successor of earlier civilizations, together with the associated Maya culture, laid great emphasis on astronomical observation and on a complex religious calendar. Important were the high god Ometecuhtli, the morning star Quetzalcóatl, and the various legends woven round Tezcatlipoca, patron of warriors, who in the form of Huitzilopochtli was patron of the Aztec nation. Inca religion

also possessed a high god, Viracocha; a number of the most important deities were associated with celestial bodies, notably the sun, patron of the Incas. Both in Central and South America, the fertility aspects of deities were also emphasized.

Modern ethnic religions in Africa and elsewhere. In some areas, such as much of Africa and Oceania, the indigenous religions are ethnic or tribal; each group has its own particular tradition. These traditions have been affected considerably by the impact of Christian missions and Western technology. Clearly there is no single pattern of belief, though certain patterns do recur in some of the cultures, such as belief in a high god, totemism (characterized by recognition of a relationship between certain human groups and particular classes of animal, plant, or inanimate object in nature), spirit possession, and so on. In various respects there are matches between myth and social organization that are likewise quite varied. Anthropologists, however, are far from a consensus on the role and origin of the gods.

(N.Sm.)

Pantheism and panentheism

Both "pantheism" and "panentheism" are terms of recent origin, coined to describe certain views of the relationship between God and the world that are different from that of traditional Theism (see *Theism* below). As reflected in the prefix "pan-" (Greek *pas,* "all"), both of the terms stress the all-embracing inclusiveness of God, as compared with his separateness as emphasized in many versions of Theism. On the other hand, pantheism and panentheism, since they stress the theme of immanence—*i.e.,* of the indwelling presence of God—are themselves versions of Theism conceived in its broadest meaning. Pantheism stresses the identity between God and the world; panentheism (Greek *en,* "in") holds that the world is included in God but that God is more than the world.

The adjective "pantheist" was introduced by the Irish Deist, John Toland, in a book, *Socinianism Truly Stated* (1705). The noun "pantheism" was first used in 1709 by one of Toland's opponents. The term "panentheism" appeared much later, in 1828, when it was used to characterize the view that the world is a finite creation within the infinite being of God.

Although the terms are recent, they have been applied retrospectively to alternative views of the divine being as found in the entire philosophical traditions of both East and West.

NATURE AND SIGNIFICANCE

Pantheism and panentheism can be explored by means of a three-way comparison with traditional or Classical Theism viewed from eight different standpoints—*i.e.,* from those of immanence or transcendence; of monism, dualism, or pluralism; of time or eternity; of the world as sentient or insentient; of God as absolute or relative; of the world as real or illusory; of freedom or determinism; and of sacramentalism or secularism.

Immanence or transcendence. The poetic sense of the divine within and around mankind, which is widely expressed in religious life, is frequently treated in literature. It is present in the Platonic Romanticism of Wordsworth and Coleridge, as well as in Tennyson, Emerson, and Goethe. Expressions of the divine as intimate rather than as alien, as indwelling and near dwelling rather than remote, characterize pantheism and panentheism as contrasted with Classical Theism. Such immanence encourages man's sense of individual participation in the divine life without the necessity of mediation by any institution. On the other hand, it may also encourage a formless "enthusiasm," without the moderating influence of institutional forms. In addition, some theorists have seen an unseemliness about a point of view that allows the divine to be easily confronted and appropriated. Classical Theism has, in consequence, held to the transcendence of God, his existence over and beyond the universe. Recognizing, however, that if the separation between God and the world becomes too extreme, man risks the loss of communication with the divine, panentheism—unlike pantheism, which holds to the divine immanence—maintains that the divine can be both transcendent and immanent at the same time.

Distinctions viewed from the metaphysical standpoint

Monism, dualism, or pluralism. Philosophies are monistic if they show a strong sense of the unity of the world, dualistic if they stress its twoness, and pluralistic if they stress its manyness. Pantheism is typically monistic, finding in the world's unity a sense of the divine, sometimes related to the mystical intuition of personal union with God; Classical Theism is dualistic in conceiving God as separated from the world and mind from body; and panentheism is typically monistic in holding to the unity of God and the world, dualistic in urging the separateness of God's essence from the world, and pluralistic in taking seriously the multiplicity of the kinds of beings and events making up the world. One form of pantheism, present in the early stages of Greek philosophy, held that the divine is one of the elements in the world whose function is to animate the other elements that constitute the world. This point of view, called Hylozoistic (Greek *hylē,* "matter," and *zōē,* "life") pantheism, is not monistic, as are most other forms of pantheism, but pluralistic.

Time or eternity. Most, but not all, forms of pantheism understand the eternal God to be in intimate juxtaposition with the world, thus minimizing time or making it illusory. Classical Theism holds that eternity is in God and time is in the world but believes that, since God's eternity includes all of time, the temporal process now going on in the world has already been completed in God. Panentheism, on the other hand, espouses a temporal–eternal God who stands in juxtaposition with a temporal world; thus, in panentheism, the temporality of the world is not cancelled out, and time retains its reality.

The world as sentient or insentient. Every philosophy must take a stand somewhere on a spectrum running from a concept of things as unfeeling matter to one of things as psychic or sentient. Materialism holds to the former extreme, and Panpsychism to the latter. Panpsychism offers a vision of reality in which to exist is to be in some measure sentient and to sustain social relations with other entities. Dualism, holding that reality consists of two fundamentally different kinds of entity, stands again between two extremes. A few of the simpler forms of pantheism support Materialism. Panentheism and most forms of pantheism, on the other hand, tend toward Panpsychism. But there are differences of degree, and though Classical Theism tends toward dualism, even there the insentient often has a tinge of Panpsychism.

God as absolute or relative. God is absolute insofar as he is eternal, cause, activity, creator; he is relative insofar as he is temporal, effect, passive (having potentiality in his nature), and affected by the world. For pantheism and Classical Theism, God is absolute; and for many forms of pantheism, the world, since it is identical with God, is likewise absolute. For Classical Theism, since it envisages a separation between God and the world, God is absolute and the world relative. For panentheism, however, God is absolute and relative, cause and effect, actual and potential, active and passive. The panentheist holds that, inasmuch as they refer to different levels of the divine nature, both sets of claims can be attributed to God without inconsistency, that just as a man can have an absolute, unchanging purpose, which gains now one embodiment and now another, so God's absoluteness can be an abstract unchanging feature of a changing totality.

Distinctions viewed from the human standpoint

The world as real or illusory. Panentheism, Classical Theism, and many forms of pantheism hold the world to be part of the ultimate reality. But for Classical Theism the world has a lesser degree of reality than God; and for some forms of pantheism, for which Hegel coined the term Acosmism, the world is unreal, an illusion, and God alone is real.

Freedom or determinism. In those forms of pantheism that envisage the eternal God literally encompassing the world, man is an utterly fated part of a world that is necessarily just as it is, and freedom is thus illusion. To be sure, Classical Theism holds to the freedom of man but insists that this freedom is compatible with a divine omniscience

that includes his knowledge of the total future. Thus the question arises whether or not such freedom is illusory. Panentheism, by insisting that future reality is indeterminate or open and that man and God, together, are in the process of determining what the future shall be, probably supports the doctrine of man's freedom more completely than does any alternative point of view.

Sacramentalism or secularism. Insofar as God is the indwelling principle of the world and of man, as in pantheism, so far do these take on a sacramental character; and insofar as God is separated from the world as in 18th-century Deism (see also *Deism* below), so far does it become secular, neutral, or even fallen. In contrast, Classical Theism, though basically sacramental, places this quality in an enclave, the church.

DIVERSE VIEWS OF THE RELATION OF GOD TO THE WORLD

On the basis of the preceding characteristics, seven forms of pantheism can be distinguished in addition to Classical Theism and panentheism:

Hylozoistic pantheism. The divine is immanent in, and is typically regarded as the basic element of, the world, providing the motivating force for movement and change. The world remains a plurality of separate elements.

Immanentistic pantheism. God is a part of the world and immanent in it. Though only a part, however, his power extends throughout its totality.

Absolutistic monistic pantheism. God is absolute and identical with the world. The world, although real, is therefore changeless.

Relativistic monistic pantheism. The world is real and changing and is within God (*e.g.*, as the body of God). But God remains nonetheless absolute and is not affected by the world.

Acosmic pantheism. The absolute God makes up the total reality. The world is an appearance and ultimately unreal.

Identity of opposites pantheism. The opposites of ordinary discourse are identified in the supreme instance. God and his relation to the world are described in terms that are formally contradictory; thus reality is not subject to rational description. Whether being is stressed or the void, whether immanence is or transcendence, the result is the same: one must go beyond rational description to an intuitive grasp of the ultimate.

God as related to time and change

Classical Theism. God is absolute, eternal, first cause, pure actuality, an omniscient, omnipotent, and perfect being. Though related to the world as its cause, he is not affected by the world. He is essentially transcendent over the world; and the world exists relative to him as a temporal effect of his action—containing potentiality as well as actuality and characterized by change and finitude. Since all of time is part of God's eternal "Now," and since God's knowledge now includes the total future as though laid out before him like a landscape, it is not clear that, in this system, man can have freedom in any significant sense; for although foreknowledge does not of itself determine anything, it vouches for the existence of such determination. Nonetheless, human freedom is in fact asserted by Classical Theists.

Neoplatonic or emanationistic pantheism. God is absolute in all respects, remote from the world and transcendent over it. This view is like Classical Theism except that, rather than saying that God is the cause of the world, it holds that the world is an emanation of God, occurring by means of intermediaries. God's absoluteness is thus preserved while a bridge to the world is provided as well. In Plotinus (3rd century AD), the foremost Neoplatonist, the Nous (Greek, "mind"), a realm of ideas or Platonic forms, serves as the intermediary between God and the world, and the theme of immanence is sustained by positing the existence of a World-Soul that both contains and animates the world.

Panentheism. In this alternative, both sets of categories, those of absoluteness and of relativity, of transcendence and of immanence, are held to apply equally to God, who is thus dipolar. He is the cause of the world and its effect; his essence is eternal, but he is involved in time. God's knowledge includes all that there is to know; since the future is genuinely open, however, and is not in any sense real as yet, he knows it only as a set of possibilities or probabilities. In this alternative man is held to have significant freedom, participating as a co-creator with God in the continuing creation of the world.

With only slight attention being accorded to Classical Theism (which is covered in another article), the incidence of the preceding eight forms of pantheism and panentheism in cultural history remains to be explored.

PANTHEISM AND PANENTHEISM IN NON-WESTERN CULTURES

The Vedas and *Upaniṣads*

Hindu doctrines. The gods of the Vedas, the ancient scriptures of India (*c.* 1200 BC), represented for the most part natural forces. Exceptions were the gods Prajāpati (Lord of Creatures) and Puruṣa (Supreme Being or Soul of the Universe), whose competition for influence provided, in its outcome, a possible explanation of how the Indian tradition came to be one of pantheism rather than of Classical Theism. By the 10th book of the Ṛigveda, Prajāpati had become a lordly, monotheistic figure, a creator deity transcending the world; and in the later period of the sacred writings of the *Brāhmaṇas* (*c.* 7th century BC), prose commentaries on the Vedas, he was moving into a central position. The rising influence of this Theism was later eclipsed by Puruṣa, who was also represented in Ṛigveda X. In a creation myth Puruṣa was sacrificed by the gods in order to supply (from his body) the pieces from which all the things of the world arise. From this standpoint the ground of all things lies in a Cosmic Self, and all of life participates in that of Puruṣa. The Vedic hymn to Puruṣa may be regarded as the starting point of Indian pantheism.

In the *Upaniṣads* (*c.* 1000–500 BC), the most important of the ancient scriptures of India, the later writings contain philosophic speculations concerning the relation between the individual and the divine. In the earlier *Upaniṣads*, the absolute, impersonal, eternal properties of the divine had been stressed; in the later *Upaniṣads*, on the other hand, and in the *Bhagavadgītā*, the personal, loving, immanentistic properties became dominant. In both cases the divine was held to be identical with the inner self of each man. At times these opposites were implicitly held to be in fact identical—the view earlier called identity of opposites pantheism. At other times the two sets of qualities were related, one to the unmanifest absolute Brahman, or supreme reality (sustaining the universe), and the other to the manifest Brahman bearing qualities (and containing the universe). Thus Brahman can be regarded as exclusive of the world and inclusive, unchanging and yet the origin of all change. Sometimes the manifest Brahman was regarded as an emanation from the unmanifest Brahman; and then emanationistic pantheism—the Neoplatonic pantheism of the foregoing typology—was the result.

The views of Śaṅkara, Rāmānuja, and Radhakrishnan

Śaṅkara, an outstanding nondualistic Vedāntist and advocate of a spiritual view of life, began with the Neoplatonic alternative but added a qualification that turned his view into what was later called acosmic pantheism. Distinguishing first between Brahman as being the eternal Absolute and Brahman as a lower principle and declaring the lower Brahman to be a manifestation of the higher, he then made the judgment that all save the higher unqualified Brahman is the product of ignorance or nescience and exists (apparently only in men's minds) as the phantoms of a dream. Since for Śaṅkara, the world and individuality thus disappear upon enlightenment into the unmanifest Brahman, and in reality only the Absolute without distinctions exists, Śaṅkara has provided an instance of acosmism.

On the other hand, Rāmānuja, a prominent southern Brahmin who held to a qualified monism, argued strenuously against Śaṅkara's dismissal of the world and of individual selves as being mere products of nescience. In place of this acosmism he substituted the notion of world cycles. In the unmanifest state Brahman has as his body only the very subtle matter of darkness, and he decrees "May I again possess a world-body"; in the manifest state all of the things of the world, including individual selves, are part of his body. The doctrine of Rāmānuja approaches panentheism; he has certainly advanced beyond

emanationistic pantheism. There are two aspects to the single Brahman, one absolutistic and the other relativistic. As in panentheism, the beings of the world have freedom. The only qualification is that, although it is Brahman's will to support the choices of finite beings, he has the power to prohibit any choice that displeases him. This power to prohibit indicates a preference for the absolute in Rāmānuja's thought, which is reflected in many ways; although God is the cause of the world, for example, and includes the world within his being, he is never affected by that world, and his motive in world creation is simply play. In sum, since the absolutistic categories were given the greater emphasis in his thought, Rāmānuja is representative of a relativistic monistic pantheism.

The presence in the Hindu tradition of both absolutistic and relativistic descriptions of the divine suggests that genuine panentheism might well emerge from the tradition; and, in fact, in the former president of India, S. Radhakrishnan, also a religious philosopher, that development did occur. Although Radhakrishnan had been influenced by Western philosophy, including that of A.N. Whitehead, later discussed as a modern panentheist, the sources of his thought lie in Hindu philosophy. He distinguishes between God as the being who contains the world and the Absolute, who is God in only one aspect. He finds that the beings of the world are integral with God, who draws an increase of his being from the constituents of his nature.

Buddhist doctrines. Some 600 years after Buddha, a new and more speculative school of Buddhism arose to challenge the 18 or 20 schools of Buddhism then in existence. One of the early representatives of this new school, which came to be known as Mahāyāna (Sanskrit "Greater Vehicle") Buddhism, was Aśvaghoṣa. Like Śaṅkara (whom he antedated by 700 years), Aśvaghoṣa not only distinguished between the pure Absolute (the Soul as "Suchness"; *i.e.,* in its essence) and the all-producing, all-conserving Mind, which is the manifestation of the Absolute (the Soul as "Birth and Death"; *i.e.,* as happenings), but he also held that the judgment concerning the manifest world of beings is a judgment of nonenlightenment; it is, he said, like the waves stirred by the wind—when the quiet of enlightenment comes the waves cease, and an illusion confronts a man as he begins to understand the world.

Identity of Suchness and the Void

Whereas Aśvaghoṣa treated the world as illusory and essentially void, Nāgārjuna, the great propagator of Mahāyāna Buddhism who studied under one of Aśvaghoṣa's disciples, transferred *Śūnya* ("the Void") into the place of the Absolute. If Suchness, or ultimate reality, and the Void are identical, then the ultimate must lie beyond any possible description. Nāgārjuna approached the matter through dialectical negation: according to the school that he founded, the Ultimate Void is the Middle Path of an eightfold negation; all individual characteristics are negated and sublated, and the individual approaches the Void through a combination of dialectical negation and direct intuition. Beginning with the Middle Doctrine School, the doctrine of the Void spread to all schools of Mahāyāna Buddhism as well as to the Satyasiddhi (Sanskrit: "perfect attainment of truth") group in Hīnayāna Buddhism. Since the Void is also called the highest synthesis of all oppositions, the doctrine of the Void may be viewed as an instance of identity of opposites pantheism.

In the T'ien-t'ai school of Chinese Buddhism founded by Chih-i, as in earlier forms of Mahāyāna Buddhism, the elements of ordinary existence are regarded as having their basis in illusion and imagination. What really exists is the one Pure Mind, called True Thusness, which exists changelessly and without differentiation. Enlightenment consists of realizing one's unity with the Pure Mind. Thus, an additional Buddhist school, T'ien-t'ai, can be identified with acosmic pantheism.

Indeed, although a mingling of types is discernible in the Hindu and Buddhist strands of Oriental culture, acosmic pantheism would seem to be the alternative most deeply rooted and widespread in these traditions.

Baal, Yahweh

Ancient Middle Eastern doctrines. Just as the early gods of the Vedas represented natural forces, so the Canaanite deities known as Baal and the Hebrew God Yahweh both began as storm gods. Baal developed into a Lord of nature, presiding with his consort, Astarte, over the major fertility religion of the Middle East. The immanentism of this nature religion might have sustained the development of pantheistic systems; but, whereas the pantheistic Puruṣa triumphed in India, the Theistic Yahweh triumphed in the Middle East. And Yahweh evolved not into a Lord of nature but into a Lord of history presiding first over his chosen people and then over world history. The requirement that he be a judge of history implied that his natural "place" was outside and above the world; and he thus became a transcendent deity. Through much of the history of Israel, however, the people accepted elements from both of these traditions, producing their own highly syncretistic religion. It was this syncretism that provided the occasion that challenged certain men of prophetic consciousness to embark upon their purifying missions, beginning with Elijah and continuing throughout the Old Testament period. In this development, the absoluteness and remoteness of Yahweh came to be supplemented by qualities of love and concern, as in the prophets Hosea and Amos. In short, the categories of immanence came to supplement the categories of transcendence and, in the New Testament period, became overwhelmingly important. The transcendent Yahweh, on the other hand, had fitted more naturally into the categories of absoluteness. And, in the Christian West, it was the transcendent God who appeared in the doctrines of Classical Theism, while pantheism stood as a heterodox departure from the Christian scheme.

PANTHEISM AND PANENTHEISM IN ANCIENT AND MEDIEVAL PHILOSOPHY

Early Greek religion contained among its many deities some whose natures might have supported pantheism; and certainly the mystery religions of later times stressed types of mystical union that are typical of pantheistic systems. But in fact the pantheism of ancient Greece was related almost exclusively to philosophical speculation. For this reason it is more rationalistic, possessing a style quite different from the Pantheisms thus far examined.

Pre-Socratic views

Greco-Roman doctrines. The first philosophers of Greece, all of whom were 6th-century-BC Ionians, were hylozoistic, finding matter and life inseparable. The basic substances that they identified as the elements of reality—the water proposed by Thales, the boundless infinite suggested by Anaximander, and the air of Anaximenes—were presumed to have the motive force of living things and thus to be a kind of life, a position here called hylozoistic pantheism.

Impressed by the absolute unity of all things, the adherents of another philosophic position, that of Eleaticism (see PHILOSOPHICAL SCHOOLS AND DOCTRINES: *Eleaticism*), so-named from its centre in Elea, a Greek colony in southern Italy, found it impossible to believe in multiplicity and change. The first step in this direction was taken by Xenophanes, a religious thinker and rhapsodist, who, on rational grounds, moved from the gods and goddesses of Homer and Hesiod to a unitary principle of the divine. He believed that God is the supreme power of the universe, ruling all things by the power of his mind. Unmoved, unmoving, and unitary, God perceives, governs, and apparently contains, or at least he "embraces," all things. So interpreted, Xenophanes provides an instance of monistic pantheism, inasmuch as, in this view, the Absolute God is united with a changing world, while the reality of neither is attenuated. This paradox may have encouraged Parmenides, possibly one of Xenophanes' disciples (according to Aristotle), to accept the changeless Absolute, eliminating change and motion from the world. Reality thus became for him a unitary, indivisible, everlasting, motionless whole. This position is basically that of absolutistic monistic pantheism in that it views the world as real but changeless. Insofar as the change and variety of the world are only apparent, Parmenides also approaches acosmic pantheism.

A third fundamental position is that of the Ephesian critic Heracleitus, among whose cryptic sayings were many that stressed the role of change as the basic reality. Heracleitus continued the hylozoistic tendencies of the Ionian philoso-

phers. Fire, his basic element, is also the universal *logos*, or reason, controlling all things; and since fire not only has a life of its own but exercises control to the boundaries of the universe as well, the system is more complex than hylozoistic pantheism. In view of the circumstance that everything is either on the way from, or to, fire, this basic element is actually or incipiently everywhere. Since the divine works here from within the universe, indeed from within a single, but basic, aspect of it, the system is an instance of immanentistic pantheism.

The philosopher Anaxagoras, one of the great dignitaries at Athens in the golden age of Pericles, approached the problem somewhat in the manner of Heracleitus. *Nous* (or Mind) he held to be the principle of order for all things as well as the principle of their movement. It is the finest and purest of things and is diffused throughout the universe. This, like the preceding system, is an instance of immanentistic pantheism.

Plato's dualism

From the standpoint of the typology here employed, Plato may be regarded as the first Western philosopher to treat the problem of the absoluteness and the relativity in God with any degree of adequacy. In the *Timaeus* an absolute and eternal God was recognized, existing in changeless perfection in relation to the world of forms, along with a World-Soul, which contained and animated the world and was as divine as a changing thing could be. Although the material can be variously interpreted, panentheists hold that Plato has adopted a dual principle of the divine, uniting both being and becoming, absoluteness and relativity, permanence and change in a single context. To be sure, he envisioned the categories of absoluteness as situated in one deity, and those of relativity in another; but the separation seems not to have pleased him, and in the tenth book of the *Laws*, by invoking the analogy of a circular motion, which combines change with the retention of a fixed centre, he explained how deity could exemplify both absoluteness and change. Plato thus may be viewed as a quasi-panentheist.

Aristotle, on the other hand, with his exclusivistic, transcendent God, exemplifying only the categories of absoluteness, anticipated the absolute God of Classical Theism, existing above and beyond the world.

Stoic pantheism, Neo-platonism

Stoicism, one of the foremost of the post-Aristotelian schools of thought, represents an immanentistic pantheism of the Heracleitean variety. First of all, the Stoics accepted the decision of Heracleitus that an indwelling fire is the principal element entering into all transformations and is also the principle of reason, the *logos*, ordering as well as animating all things, but that, second, there is a World-Soul, which is diffused throughout the world and penetrates it in every part. Rather than approximating Plato's spiritual World-Soul, the Stoic World-Soul is more like the *Nous* of Anaxagoras. The Stoics were Materialists, and their diffuse World-Soul is, thus, an extended form of subtle matter. That everything is determined by the universal reason is an unvarying theme in Stoicism; and this fact suggests that Stoic pantheism, despite its immanentism, stresses the categories of absoluteness rather than those of relativity in the relations holding between God and the world.

The life of reason brings man into harmony with God and with nature and helps him to understand his fate, which is his place in the universal system. Although the view is an amalgam of several types of pantheism, this particular mixture has retained its identity. It is therefore useful to call this position, or any similar combination of themes, by the name Stoic pantheism.

Plotinus, the creator of one of the most thoroughgoing philosophical systems of ancient times, may be taken to represent Neoplatonism, an influential modification of Plato's attempt to deal with absoluteness and relativity in the divine. Plotinus' system consists of the One—the absolute God who is the supreme power of the system—the intermediate *Nous*, and the World-Soul (with the world as its internal content). His World-Soul follows the Platonic model. The system really blends pantheism with Classical Theism, since the categories of absoluteness apply to the One, and the relativistic categories apply to the World-Soul. The doctrine of emanation, whereby the power of the One comes into the world, is a clear attempt to bridge the gap between absoluteness and relativity. For Plotinus, as for Classical Theism, there is immanent in man an image of the divine, which serves as well to relate man to God as does the divine spark in Stoic pantheism. Even Classical Theism may thus contain a touch of immanentistic pantheism. This view, or any similar combination of themes, is an instance of emanationistic or Neoplatonic pantheism.

Medieval doctrines. Though Scholasticism, with its doctrine of a separate and absolute God, was the crowning achievement of medieval thought, the period was, nonetheless, not without its pantheistic witness. Largely through Jewish and Christian mysticism, an essentially Neoplatonic Pantheism ran throughout the age.

The only important Latin philosopher for six centuries after St. Augustine was John Scotus Erigena. Inasmuch as, in his system, Christ's redemptive sacrific helps to effect a Neoplatonic return of all beings to God, Erigena can be said to have turned Neoplatonism into a Christian drama of fall into sin and redemption from its power. When Erigena said that, even in the stage of separation from God, God in his superessentiality is identical with all things, he advanced beyond a strictly Neoplatonic pantheism to some stronger form of immanentistic or monistic pantheism.

In the two principal writings of the esoteric Jewish movement called the Kabbala, known for its theosophical interpretations of the Scriptures, a mystically oriented system of 10 emanations is presented. A Spaniard, Avicebrón, a Jewish poet and philosopher, similarly presented a Neoplatonic scheme of emanations. And in Spain, Averroës, the most prominent Arabic philosopher of the period, represented an Aristotelian tradition that is heavily overladen with Neoplatonism. For Averroës, the active intellect in man is really an impersonal divine reason, which alone lives on when man dies.

The German Meister Eckehart, probably the most significant of philosophical mystics, developed a markedly original theology. From his Stoic pantheism there arose his most controversial thesis—that there resides in every man a divine, uncreated spark of the Godhead, making possible both a union with God and a genuine knowledge of his nature. But Eckehart also distinguished between the unmanifest and barren Godhead and the three Persons who constitute a manifest and personal God. Thus, the system has similarities to both Stoic and Neoplatonic pantheism.

Cardinal Nicholas of Cusa, whose broad scholarship and scientific approach anticipated the coming Renaissance, continued the tradition into the 15th century. The "learned ignorance," in which a man separates himself from every affirmation, can have positive results, in Nicholas' view, because man is a microcosm within the macrocosm (or universe), and the God of the macrocosm is thus mirrored in all of his creatures. He also held that, in reference to God, contradictions are compatible—his "coincidence of opposites" doctrine, in which God is at once all extremes. Clearly, Nicholas wished to ascribe to God both the categories of transcendence and those of immanence without distinction. But in fact he displayed some preference for the categories of the absolute, insisting, for example, that the creatures of the world can add nothing to God since they are merely his partial appearances. Despite this bias toward absolutism, and even to acosmism, Nicholas can be appropriately viewed as espousing an identity of opposites pantheism.

PANTHEISM AND PANENTHEISM IN MODERN PHILOSOPHY

Renaissance and post-Renaissance doctrines. The humanism of the Renaissance included an enlarged interest in Platonism and in its historical carrier, Neoplatonism, as well as influences from Aristotle and from Kabbalistic sources. The view of man as a microcosm of the universe was widespread. Marsilio Ficino, one of the first leaders of the Florentine Academy, found the image and reflection of God in all men and anticipated the divinization of man and the entire cosmos. The humanist and syncretistic philosopher Pico della Mirandola, also a leading figure in the Academy, substituted for creation a Neoplatonic emanation from the divine.

The views of Ficino, Bruno, and Böhme

The most famous scholar of the Italian Renaissance was Giordano Bruno. Combining Copernican astronomy with Neoplatonism, Bruno thought of the universe as an infinite organism with monads as its ultimate constituents and world-systems as its parts. The universe, he held, is in a continual process of development and is infused with the divine life. Accepting Nicholas of Cusa's doctrine of the identity of opposites, he taught that contradictory ascriptions apply equally to God in particular and that claims concerning his immanence and transcendence are equally valid. More open to the categories of relativity than Nicholas, Bruno, however, exemplified a neatly balanced instance of identity of opposites pantheism.

The next great innovator of mystical religious thought was Jakob Böhme, who, in developing the concept of the divine life, took a decisive step beyond mere absoluteness. God goes through stages of self-development, he taught, and the world is merely the reflection of this process. Böhme anticipated Hegel in claiming that the divine self-development occurs by means of a continuing dialectic, or tension of opposites, and that it is the negative qualities of the dialectic that men experience as the evil of the world. Even though Böhme, for the most part, stressed absoluteness and relativity equally, his view that the world is a mere reflection of the divine—apparently denying self-development on the part of creatures—tends toward acosmic pantheism.

Spinoza's rational pantheism

In the 17th century the foremost pantheist was a Jewish rationalist, Benedict Spinoza, whose training in the history of philosophy included both medieval Jewish philosophy and the Kabbala. He championed a rational rather than a mystical pantheism, so much so that all that remained of mysticism, in fact, was his concept of the intellectual love of God. The rationality of the system is suggested by Spinoza's argument that, since God is the infinite being, he must be identical with the world; for otherwise, God-and-world would be a greater totality than God alone. Also, since God is a necessary being and is identical with the world, the world must also be necessary in all its parts. It follows from this that human freedom is an impossible idea; and the sense that man has of such freedom is based on his ignorance of the causes that have determined him. Spinoza distinguished between God and the world in three ways: first, by stressing God's activity in the active sense of *natura naturans* ("the nature that [creates] nature"; *i.e.*, God) compared to the passive sense of *natura naturata* ("the nature that [is created as] nature"; *i.e.*, the world); second, he related God to eternity and the world to time; and third, he distinguished God as self-existing substance, the whole, from the world, which he conceived as the attributes and modes of that substance. In terms of the present classification, Spinoza represents a monistic pantheism tending toward absolutism.

Goethe, the incomparable German litterateur, claimed that he was a follower of Spinoza. In fact, however, his beliefs were rather different inasmuch as Goethe championed man's individuality; opposed mechanical necessity; and held a hylozoistic, or vitalistic, position in which nature was organic, a living unity. His personalistic pantheism mixes hylozoistic and Stoic types with a touch of relativism added to the mixture.

Nineteenth-century doctrines. During the 19th century, pantheism and panentheism were sustained by various kinds of Idealism that developed during the period. In these systems the categories of relativity gained in prominence; God was conceived as entering history and as being more intimately related to processes of change and development.

German Idealism. Although the philosophy of the German patriot J.G. Fichte, an immediate follower of Kant, began in the inner subjective experience of the individual, with the "I" positing the "not-I"—*i.e.*, feeling compelled to construct a perceived world over against itself—it turns out eventually that, at a more fundamental level, God, as the universal "I," posits the world at large. The world, or nature, is described in organic terms; God is considered not alone as the Universal Ego but also as the Moral World Order, or Ground of ethical principles; and since every man has a destiny as a part of this order, man is in this sense somehow one with God. In the moral world order, then, man has a partial identity with God; and in the physical order he has membership in the organic whole of nature. It is not clear, however, whether in Fichte's view God as Universal Ego includes all human egos, and the organic whole of nature. Should he do so, then Fichte would be a representative of dipolar Panentheism, since in his final doctrine the Universal Ego imitates an Absolute deity who is simply the divine end of all activity, serving equally as model and as goal. In this interpretation God is conceived both as absolute mobility and absolute fixity. It is not entirely clear whether the doctrine is to be understood as referring to two aspects of a single God, the panentheistic alternative, or to two separate gods, the alternative imbedded in Plato's quasipanentheism. In either case, Fichte has enunciated most of the themes of panentheism and deserves consideration either as a representative or precursor of that school.

A second early follower of Kant was F.W.J. von Schelling, who, in contrast to Fichte, stressed the self-existence of the objective world. Schelling's thought developed through several stages. Of particular interest to the problem of God are the final three stages in which his philosophy passed through monistic and Neoplatonic pantheism followed by a final stage that was panentheistic.

In the first of these stages, he posits the Absolute as an absolute identity, which nonetheless includes, as in Spinoza, both nature and mind, reality and ideality. The natural series culminates in the living organism; and the spiritual series culminates in the work of art. The universe is, thus, both the most perfect organism and the most perfect work of art.

In his second, Neoplatonic, stage he conceived the Absolute as separated from the world, with a realm of Platonic ideas interposed between them. In this arrangement, the world was clearly an emanation or effect of the divine.

In the final stage of his thought, Schelling presented a theophany, or manifestation of deity, involving the separation of the world from God, and its return. In appearance this was quite like the views of Erigena or like the unmanifest and manifest Brahman of Indian thought. But, since the power of God continues to infuse the world and there can be no real separation, the entire theophany is clearly the development of the divine life. The Absolute is retained as the pure Godhead, a unity presiding over the world; and the world—having in measure its own spontaneity—is both his antithesis and part of his being, the contradiction accounting for progress. The positing within God of eternity and temporality, of being-in-itself and of self-giving, of yes and no, of participation in joy and in suffering, is the very duality of Panentheism.

It was a disciple of Schelling, Karl Christian Krause, who coined the term panentheism to refer to the particular kind of relation between God and the world that is organic in character.

Hegel's notion of Absolute Spirit

The third, and most illustrious, early post-Kantian Idealist was G.W.F. Hegel, who held that the Absolute Spirit fulfills itself, or realizes itself, in the history of the world. And in Hegel's deduction of the categories it is clear that man realizes himself through the attainment of unity with the Absolute in philosophy, art, and religion. It would appear, then, that God is in the world, or the world is in God, and that, since man is a part of history and thus a part of the divine realization in the world, he shares in the divine life; it would seem, too, that God is to be characterized by contingency as well as necessity, by potentiality as well as actuality, by change as well as permanence. In short, it would seem at first that the panentheistic dipolarity of terms would apply to the Hegelian Absolute. But this is not quite so; for Hegel's emphasis was on the deduction of the categories of logic, nature, and spirit, a deduction that provided the lineaments of Spirit-in-Itself (the categories of the intrinsic logic that the world, as Spirit, follows in its development), Spirit-for-Itself (nature as existing oblivious of its own context), and Spirit-in-and-for-Itself (conscious spiritual life, natural, and yet aware of its role in the developing world). This deduction, moving from the most abstract categories to the most concrete, is partly logical and partly temporal; it cannot be read either as a sheerly

logical sequence or as a sheerly temporal sequence. As a logical sequence, it has the appearance of a Neoplatonic scheme turned on its head, since the Absolute Spirit that emerges from the deduction includes all of the steps of the preceding rich and multifarious deduction. As a temporal sequence, the system would seem to be a species of Stoic (*i.e.,* Heracleitean) pantheism, qualified by a clear Parmenidean motif (see above), which appears in its stress on an absoluteness that, from the eternal standpoint, cancels out time. This Parmenidean quality is to be found not only in Hegel but in most of the Idealists who were influenced by him. Time is real, on this view, and yet not quite real, having already eternally happened. And when Hegel spoke of the Absolute Spirit, this phrase held the internal tension of a near contradiction, for spirit, however absolute, must surely be relative to what is around it, sensitive to and dependent on other spirits. The fact that Hegel wished to give something like equal emphasis, however, both to absoluteness and to relativity in the divine being or process suggests that his goal is identical with that of the panentheists, even though he is perhaps more fairly regarded as a Pantheist of an ambiguous type.

Monism and panpsychism. It is impossible for one to leave the 19th century without mention of the pioneering experimental psychologist Gustav Fechner (1801–87), founder of psychophysics, who developed an interest in philosophy. Fechner pursued the themes of panentheism beyond the positions of his predecessors. A panpsychist with an organic view of the world, he held that every entity is to some extent sentient and acts as a component in the life of some more inclusive entity in a hierarchy that reaches to the divine Being, whose constituents include all of reality. God is the soul of the world, which is, in turn, his body. Fechner contends that every man's volitions provide impulses within the divine experience, and that God gains and suffers from the experiences of men. Precisely because God is the supreme being, he is in process of development. He can never be surpassed by any other, but he surpasses himself continually through time. He, thus, argues that God can be viewed in two ways: either as the Absolute ruling over the world, or as the totality of the world; but both are aspects of the same Being. Fechner's affirmations comprise a complete statement of panentheism, including the dipolar deity with respect to whom the categories of absoluteness and relativity can be affirmed without contradiction.

Fechner's panentheism

Twentieth-century doctrines. The 20th century marks a decisive break with absolutism. In the first half of the century, panentheism gained in authority. The position of the Russian ex-Marxist Nikolay Berdyayev, a religious metaphysician, with his emphasis on divine and human freedom, is a manifesto of panentheism. Even more impressive was the work of the eminent British-American philosopher, Alfred North Whitehead. As in the case of Fechner, Whitehead came to philosophy from science and held an organismic view of the structure of the world. In Whitehead's view God has two natures: his primordial nature is abstract; his consequent nature is concrete and includes within itself the total history of the world. Whitehead was also a panpsychist and believed that feeling is present in some degree at every level of the world process. Whether or not he was, then, also a panentheist is in dispute. He held that the possible future and the total past are in God—in his primordial and consequent natures; but for Whitehead the present moment is relative, and contemporaries exclude each other. In the present moment of any entity, since it is the present of *that* entity, it is appropriate to say that God is in that entity, part of the data on which it acts; thus the Stoic spark of divinity has here a modern application. From the standpoint of God, on the other hand, all entities are part of God; they come from him and return to him in the passage of time, but they are not in God in the sense that their independence in the present moment is prejudiced.

The views of Whitehead and Hartshorne

It was left to Charles Hartshorne, one of Whitehead's followers, to provide the definitive analysis of panentheism. It is Hartshorne's suggestion that the organismic analogy, present in Whitehead as well as in many earlier thinkers, be taken seriously. For Hartshorne, God includes the world even as an organism includes its cells, thus including the present moment of each event. The total organism gains from its constituents, even though the cells function with an appropriate degree of autonomy within the larger organism.

CRITICISM AND EVALUATION OF PANTHEISM AND PANENTHEISM

Panentheism is then a middle way between the denial of individual freedom and creativity characterizing many of the varieties of pantheism and the remoteness of the divine characterizing Classical Theism. Its support for the ideal of human freedom provides grounds for a positive appreciation of temporal process, while removing some of the ethical paradoxes confronting deterministic views. It supports the sacramental value of reverence for life. At the same time the theme of participation with the divine leads naturally to self-fulfillment as the goal of life.

Many pantheistic and Theistic alternatives claim the same advantages, but their natural tendency toward absoluteness may make justification of these claims in some cases difficult and, in others, some argue, quite impossible. It is for this reason that a significant number of contemporary philosophers of religion have turned to panentheism as a corrective to the partiality of the other competing views.

(W.L.Re.)

Religious dualism

Dualism is the doctrine that the world (or reality) consists of two basic, opposed, and irreducible principles that account for all that exists. It has played an important role in the history of thought and of religion.

NATURE AND SIGNIFICANCE

In religion, dualism means the belief in two supreme opposed powers or gods, or sets of divine or demonic beings, that caused the world to exist. It may conveniently be contrasted with monism, which sees the world as consisting of one principle such as mind (spirit) or matter; with monotheism; or with various pluralisms and polytheisms, which see a multiplicity of principles or powers at work. As is indicated below, however, the situation is not always clear and simple, a matter of one or two or many, for there are monotheistic, monistic, or polytheistic religions with dualistic aspects.

Types of dualism

Various distinctions may be discerned in the types of dualism in general. In the first place, dualism may be either absolute or relative. In a radical or absolute dualism, the two principles are held to exist from eternity; for example, in the Iranian dualisms, Zoroastrianism and Manichaeism, both the bright and beneficent and the sinister and destructive principles are from eternity.

In a mitigated or relative dualism, one of the two principles may be derived from, or presuppose, the other as a basis; for example, the Bogomils, a medieval heretical Christian group, held that the devil is a fallen angel who came from God and was the creator of the human body, into which he managed by trickery to have God infuse a soul. Here the devil is a subordinate being and not coeternal with God, the absolute eternal being. This, then, is clearly a qualified, not a radical, dualism. Both radical and mitigated types of dualism are found among different groups of the late medieval Cathars, a Christian heretical movement closely related to the Bogomils.

Another and perhaps more important distinction is that between dialectical and eschatological dualism. Dialectical dualism involves an eternal dialectic, or tension, of two opposed principles, such as, in Western culture, the One and the many, or Idea and matter (or space, called by Plato "the receptacle"), and, in Indian culture, *māyā* (the illusory world of sense experience and multiplicity) and *ātman-brahman* (the essential identity of mind and ultimate reality). Dialectical dualism ordinarily implies a cyclical, or eternally repetitive, view of history. Eschatological dualism—*i.e.,* a dualism concerned with the ultimate destiny of man and the world, how things will be in the "last" times—on the other hand, conceives of a final resolution of the present dualistic state of things, in

which evil will be eliminated at the end of a "linear" history constituted of a series of unrepeatable events, instead of a "cyclical," repetitive one. The ancient Iranian religions, Zoroastrianism and Manichaeism, and Gnosticism—a religiophilosophical movement influential in the Hellenistic world—provide examples of eschatological dualism. A type of thought, such as Platonism, that insists on a profound harmony in the cosmos, is thus more radically dualistic, because of its irreducibly dialectical character (see below) than Zoroastrianism and Manichaeism, with their emphasis on the cosmic struggle between two antithetical principles (good and evil). Midway between these extremes is Gnostic dualism, which has an ontology (or theory of being) of an Orphic-Platonic type (for Orphism, see below *Among ancient civilizations and peoples*) but which also affirms the final disappearance and annihilation of evil with the eventual destruction of the material world—and thus comprises both dialectical and eschatological dualism.

In philosophy, dualism is often identified with the doctrine of transcendence—that there is a separate realm or being "above" and "beyond" the world, as opposed to monism, which holds that the ultimate principle is inside the world (immanent). In the disciplines concerned with the study of religions, however, religious dualism refers not to the distinction or separation of God and the world but to the doctrine of two basic principles; a doctrine that, moreover, may easily be compatible with a form of monism (*e.g.,* Orphism or Vedānta) that makes the opposition between the One and the many absolute and sees in multiplicity merely a fragmentation (or illusory obliteration) of the One.

HISTORICAL VARIETIES OF RELIGIOUS DUALISM

Among ancient civilizations and peoples. Dualism is a phenomenon of major importance in the religions of the ancient world. Those of the Middle East will be considered here.

Egypt and Mesopotamia. While there was generally no explicit dualism in ancient Egyptian religion, there was an implicit dualism in the contrast between the god Seth and the god Osiris. Seth, a violent, aggressive, "foreign," sterile god, connected with disorder, the desert, and loneliness, was opposed to Osiris, the god of fertility and life, active in the waters of the Nile. Seth also possessed some typically dualistic marks of a mythological character; his action, as well as his personality itself, was ambivalent; and, as a typical trickster, he was also capable, at times, of constructive action in the cosmos. The myths of Osiris and Seth may be compared in various ways with those recently discovered among the Dogon people of the western Sudan, which contrast Nommo, a fertile and happily mated primordial being pictured in fish form, with Yurugu ("Pale Fox"), an unhappy, sterile character who lives in the wilderness without a mate. Yurugu is considered to be the element that makes the universe complete (the same role assigned to Seth in the Egyptian myth).

Osiris and Seth

Dualism, broadly speaking, was also present in ancient Mesopotamian religion. In myths pertaining to the origin of the gods and of the cosmos, the opposition between the primordial deities (Apsu, the Abyss; and Tiamat, the Sea) and the new ones (particularly Marduk, the demiurge, or creator) displayed some dualistic aspects. Though the earlier deities had established the basic reality of the universe—its ontological core—because of their chaotic and selfish nature they resisted their own offspring, who were later to create the now existing, definite order of the cosmos. A dualism of the ontological—basic reality or being—versus the cosmological—the form or order of the material universe—is thus implicitly affirmed.

Greece and the Hellenistic world. Analogous dualistic concepts may be found in the early Greek *Theogony* of Hesiod (fl. *c.* 700 BC) in his myths of the gods Uranus, Cronus, and Zeus, and the conflict between primordial and later gods. It was in the later, classical Greek world, however, that dualism was most evident. Many of the pre-Socratic philosophers (6th and 5th centuries BC) were dualistic in various ways. In the teachings of Parmenides, for example, noted for reducing the world to a static One—a classical instance of monism—there is still a radical opposition between the realms of Being and Opinion—between ultimate reality and the world of human sense experience. On the other hand, in the doctrines of Heracleitus, noted for reducing the world to fiery Change, the conflict of opposites (hot–cold, day–night, beginning–end, the-way-up–the-way-down), called by Heracleitus *polemos* ("war"), was exalted to become a metaphysical principle. Though these opposites are piecemeal dyads ("pairs"), their effect, taken together, is, as a whole, dualistic. The dualism of Empedocles, simultaneously a religious teacher and a natural philosopher, is especially striking, for he viewed the primordial sphere of the universe as undergoing cycles alternately under the dominance of the antithetical principles of Love and Discord, which periodically break and then reconstruct it. In this context there exist *daimones* ("souls"), divine beings that have fallen from a superior world into this world and exist clothed in the "foreign robe of the flesh." These souls are therefore subject to transmigration through a series of vegetable, animal, and human bodies, owing to a primitive accident (for which credit was given "to the furious Discord").

The same antithetical principles are to be found in Orphism, a Greek mystical school, which constituted an independent development within Greek religion and philosophy; beginning in the 6th century BC, it was part of a "mysteriosophic" trend that sought to attain the wisdom of secret mystic and cultic doctrines. Orphism is characterized by its *sôma-sêma,* or body-tomb concept, which saw the body as a prison or tomb in which the soul—a divine element, akin to the gods—is incarcerated. In addition to this psychophysical dualism of soul and body, the Orphic idea that "everything comes from the One and returns to the One" demonstrates a typical dialectical dualism, in which an implicit monism is involved. Developing on an analogous level, Pythagorean numerical and mystical speculation, arising from the 6th-century-BC Greek philosopher and religious teacher Pythagoras, also stressed the dualistic opposition of Monad–Dyad (One–Two) and of other dialectical pairs of opposites.

Platonic dualism

Many of these dualistic ideas, especially the Orphic and Pythagorean ones, are also found in writings of the Greek philosopher Plato (428–348/347 BC), such as the *Timaeus, Phaedo, Gorgias,* and *Cratylus.* In these writings a divine part of the human soul that is directly infused by the divinity and a mortal part (passionate and vegetative) are defined and considered. The mortal part is assigned to man by inferior divinities, charged to do so by the supreme divinity; and the appetitive passions involved, if followed, are held to be responsible for the punishments that the soul will suffer during various periods of habitation in the other world and reincarnations in this one. Thus God remains free of blame for the destiny of man. The mortal or spoiled part of man is further attributed, in Plato's *Laws,* to the "titanic nature" within his makeup—an element of violence and impiety inherited from the primordial rebellious Titans, sons of the Earth.

Plato's notions of man were rooted in both ontology and cosmology; *i.e.,* in views on being and on the orderly structure of the universe. In the *Timaeus* he considers the cosmos as a single harmony, which for the sake of completeness requires the existence of inferior levels that are bound not only to matter but also to Necessity (the realm of things that could not have been otherwise, and that are hence not amenable to divine activity). A different view is found in his *Laws,* which describes two "Souls" of the World, one of which causes good and one evil. The *Politicus* is concerned with two eternally recurring, alternating cycles in the cosmos, with successive epochs guided either by the gods or by men.

Plato's central inspiration, which unifies his metaphysics, his cosmology, his theory of man, and his doctrine of the soul, was basically dualistic (in the sense of dialectical dualism) with two irreducible principles: the Idea and the *chora* (or material "receptacle") in which the Idea impresses itself. All of this world is conditioned by materiality and necessity; and because of this, the descent of souls into bodies is said to be rendered necessary as well.

Neoplatonism, a 3rd-century-AD development from Pla-

to's thought, conceived the cosmos as a harmony with a succession of levels emanating from an ultimate unit. There was in the system, nevertheless, a rupture of the harmony of the cosmos called *tolma* ("the audacity"), which served as an explanation for the descent of Soul into the material world—and thus constituted a dualistic element.

Gnosticism

In Gnosticism, a Hellenistic religious movement that entered original Christianity from earlier pagan sources, and which viewed matter as evil and spirit as good, dualism manifested itself in a more dramatic way. Gnostic dualism cannot be understood without reference to both Judaism and Christianity, and perhaps even to Zoroastrianism, since Gnostic eschatological characteristics were derived from them. Gnosticism was also connected with certain principles of Orphism and Platonism; reflecting the Orphic body-tomb doctrine, for example, Gnosticism adopted a firmly antisomatic stance (against the body), and similarly adopted the concept of the divine soul—the pneumatic, or spiritual, soul, as the Gnostic would say, of the same substance as the divinity—that is destined to free itself from the tyranny of a material, cosmic demiurge (or subordinate deity). Certain Gnostics, moreover, developed a radical anticosmism, in which they registered their animosity against the material universe by cursing the stars—which brought them bitter reproach from Plotinus (*c.* AD 205–269/270), the founder of Neoplatonism. As viewed by the Gnostic Ophite sect, which venerated the *ophis* (or "snake") as a symbol of knowledge, the cosmos comprises three parts: the superior world, the inferior world (material and chaotic), and the intermediate world, or *logos* ("word" or "reason")—the *logos* being depicted as a snake that impresses spiritual forms into the chaotic matter. These forms—life, soul, vital masculine substance—are later freed again, a liberation that completely empties the material world. Such Gnostic views are of two types: Iranian and Syrian-Egyptian. Iranian Gnosticism is characterized by an absolute, radical dualism: light and darkness, *pneuma* ("spirit") and chaotic formless matter, oppose each other from eternity. Syrian-Egyptian Gnosticism is characterized by a dualism that is mitigated (as earlier defined) but also drastic: the inferior world, the chaotic darkness, begins to exist only at a special moment owing to an accident in the divine world; and this accident is usually also identified with an "audacity," a defect in one of the "aeons," or divine entities.

Iran. In the Indo-Iranian period (2nd millennium BC) there were already tendencies toward dualistic thought, especially in myths relating to monstrous and demonic beings who still the movement of the waters and thus make cosmic life impossible; in later-archaic Indian speculation there was also a tendency to oppose *devas* ("gods") to *asuras* ("demons"). Iranian dualism, however, expressed itself most characteristically in Zoroastrianism. In the Zoroastrian religious texts, the *Gāthās*, there is an opposition between two spirits, the Beneficent Spirit (Spenta Mainyu) and the Destructive Spirit (Angra Mainyu, or Ahriman). These two spirits are different, irreducible principles; at the beginning they have chosen life and nonlife, respectively. Though the Beneficent Spirit is almost an hypostasis (the substance) of the divinity (Ahura Mazdā), nothing is said in the *Gāthās* about the origin of the Destructive Spirit. In any case, the very fact that the Destructive Spirit is said to be "twin brother" of the Beneficent One does not imply that he is a son of Ahura Mazdā but only that the two spirits are "symmetrical"; *i.e.*, equal and contrary as to their respective efficacy and orientation.

Ahura Mazdā versus Ahriman

Medieval Zoroastrian treatises present radical and eschatological dualisms in their extreme forms. According to the *Bundahishn* ("Primordial Creation") text, Ormazd (Ahura Mazdā) and Ahriman have always existed. Ormazd is represented as lofty, in the light, full of omniscience and goodness, while Ahriman is represented as debased, in darkness, full of aggressiveness and ignorance. Ormazd's omniscience allows him to conceive and to actualize the Creation and Time, because only these can offer him an arena in which to accost Ahriman and eliminate him.

The medieval Zoroastrian treatises also describe another "dual" formulation, the two realms of creation and of reality: the *mēnōk* ("potential, embryonic, initial, heavenly, and invisible") and the *gētīk* ("realized, final, worldly, concrete, and visible"). But this opposition does not imply a devaluation of the *gētīk*, of this world.

Zurvanism, a Zoroastrian heretical movement (*c.* 3rd/4th century BC–7th century AD), was also dualistic. The very names of Zurvān (Time-Destiny) and the partially synonymous *zamān* ("time") already appear in the later Avesta and in medieval treatises, in which Time is the milieu in which Ormazd and Ahriman fight. Also, a myth attributed to Zoroastrian priests by later, non-Iranian sources speaks of Zurvān as the father of Ormazd and Ahriman. At times "Zurvanite" mythology tends toward formulations of a Gnostic and Manichaean type (women paid allegiance, for example, to Ahriman, who has partial authority in the world). Zurvanism also developed theosophic characteristics (involving mystical insights), such as that which discerned the ambivalence of Zurvān—viz., that although an evil element (an evil thought or spiritual corruption) has always existed within him, he nonetheless, so it seems, eliminates the evil by expressing it and is thus worthy to be identified with the supreme divinity (Yazdān).

Among religions of the East. Dualisms have also appeared in various forms in the religions of India and China.

India. Indian dualism has involved the opposition of the One and the many: of reality and appearance. In an ancient Hindu hymn (Rigveda, 10.90), Puruṣa, "the Immortal that is in heaven," is opposed to this world; the three quarters of the Puruṣa that comprise the transcendent world are opposed to the other quarter of him (his limbs) that is this world; *i.e.*, the divine foundation, the divine substance of this world, is made out of his limbs. Early speculation on the identity of the *ātman* ("Self") and Brahman (the very core of reality), as opposed to the material and visible world that is subject to *māyā* (or "mundane illusion"), has been mentioned above.

The Sāṃkhya school of Indian philosophy presents another, probably later, formulation of dualism based on two eternal and opposed cosmic principles: *prakṛti* ("original matter") and *puruṣa* ("spirit"), the name of the ancient primordial Man, substance of the universe. Matter is differentiated into three different *guṇas* (or "qualities") that articulate the three levels of the being and essential nature of man in hierarchical connection with each other. Spirit, in itself free, eternal, and infinite, becomes involved in matter by the development of the latter. Salvation coincides with the knowledge of the state of things: "I (spirit) am one thing and It (matter) is another."

China. The first words of the Taoist text, the *Tao-te Ching*, express a doctrine that is typical of a pervasive Chinese dualism; *i.e.*, that of the two opposed and complementary principles, the Yin and the Yang (respectively, feminine and masculine, lunar and solar, terrestrial and celestial, passive and active, dark and bright; in short, the entire series of opposites). The dialectics of Yin and Yang are the double manifestation of the one and only eternal, undividable, and transcendent principle: Tao ("the Way").

Yin and Yang

Among religions of the West. Dualisms have appeared in Western religions chiefly under the impact of Gnostic influences.

Judaism. No real dualism is found in Judaism, except in the Gnostic and theosophic forms of Jewish mysticism known as Kabbala. The presence of a vigorous and universal monotheism implies not only faith in a single creative god but also faith in a god who is the uncontested master of history; and neither Satan nor Belial detract from this absolute monotheism. Within these limitations, however, a tendency towards dualistic thought could be seen in such late noncanonical texts as the *First Book of Enoch* (*c.* 1st century BC), in which certain angels are said to have fallen as a consequence of their wedding with the daughters of men. These angels, it is held, taught mankind the malevolent arts of magic, seduction, and violence, together with such elements of culture as the use of metals and writing. Though there is no dualism in the proper sense in the *Manual of Discipline*, one of the Qumrān texts of the Dead Sea Scrolls, a certain polarity is nonetheless displayed in a passage that asserts of God that

> he created man to have dominion over the world and made for him two spirits, so that he may walk by them until the

time of his visitation: they are the spirits of truth and error. In the dwelling of light are the origins of the truth, and from a spring of darkness are the origins of error. In the hand of the Prince of Lights is dominion over all the children of righteousness, in the ways of light they walk. And in the hand of the angel of darkness is all dominion over the children of error; and in the ways of darkness they walk.

The context of this passage, however, is completely monotheistic. It expresses a doctrine also found in the *Didachē*, a Jewish-Christian work of the early 2nd century AD (better known as the *Teachings of the Twelve Apostles*), that of the two roads on which a man may walk, the good road and the bad, the road of life and that of death, with God leaving the choice of the road to man's free will; and also the later rabbinic doctrine of the struggle between the good and evil inclinations (*yetzer*) within man. There is also no hint of dualism in the two "sources" mentioned in the Qumrān texts, the bright source and the dark. These are hardly dualistic principles (in the ontological sense of the term) but are simply radical (*i.e.*, original) polarities in spiritual orientation. (Not even the "Angel of Darkness," mentioned in the same context is a principle, though he is a person and a power.)

Non-dualistic polarities in late Judaism and Jewish Christianity

There is thus no true parallelism with the two principles that appear in Iranian Zurvanism. Elements of dualistic thought (in a Platonic sense) are also found in the works of the Jewish Hellenistic philosopher Philo of Alexandria (1st century AD), whose philosophy was dualistic in its doctrines about the universe and man, but without shaking his basic adherence to biblical monotheism.

Christianity. In Christianity dualistic concepts appeared principally in its Gnostic developments. But even in the 2nd-century Judaizing sect of the Encratites, which was not really Gnostic, there were dualistic aspects that had modified some tendencies in later Judaism. These teachings were also particularly prominent in the writings of the supporters of Docetism (the doctrine that Christ, being divine, did not suffer and die; 2nd century), who held that matter is essentially evil and that the soul is a preexistent substance. According to the Encratites, the preexistent soul, once it "gets effeminated by concupiscence," drops into the carnal world. Since generation perpetuates the soul's state of decay in this bodily world, they condemned all sexual relations. The dualism of Marcion (a 2nd-century semi-Gnostic Christian heretic) was really a ditheism (a system positing two gods), though the common Gnostic presuppositions—such as antisomatism and anticosmism, the condemnation of the body and the material universe—were also present in his thought. For Marcion, the God of the Old Testament is an inferior and harsh creator demiurge, author of the world and man, who is nonetheless completely distinct from the supreme divinity, who manifested himself in Jesus and is a stranger to this world. For Saturninus (or Satornil) of Antioch, the founder of a 2nd-century Syrian Gnostic group that was commonly connected with the tradition of Simon Magus (reputed leader of an earlier Gnostic sect), the God of the Old Testament is only one of the angels, the martial angel of the Judaic nation, although (as with Marcion) he is distinct from the devil, who is in fact his opponent. According to Saturninus a primordial accident caused a wave of *pneuma* ("spirit") to land in the inferior darkness, where it is said to have remained prisoner and now continues its existence in those who, characterized by the presence in them of this superior element, will later be conducted back to their heavenly origin by Jesus, a messenger coming from above. Conceptions of a similar type are also found in the "Psalm (or Hymn) of the Naassenes" (Naassene is the Hebrew term for Ophite, mentioned above) and in the "Song of the Pearl" in the Gnostic *Acts of Thomas;* here also occurs the concept of a "saviour to be saved," who has been sent from above and was made a prisoner by darkness. This basic concept was developed fully only in Manichaeism. The Gnostic-dualist view survived in late antiquity and into the Middle Ages, both in the East, among the Mandaeans, Yazīdis, and some extreme sects within the Shīʿah branch of Islām and in the West among the Bogomils and Cathars. It is still present today in modern theosophy.

Encratism

Among religions of modern nonliterate peoples. Religious dualism also manifests itself among nonliterate peoples, especially in the concept of a "second" figure, an ambivalent demiurge-trickster who can be both a collaborator and rival of the supreme being and independent of the latter in origin. Such tricksters include the Coyote (in North American Indian mythology), the Raven (among Paleosiberians), or the Crow (among the Southeast Australian tribes). To these animal figures are attributed the origin of such negative aspects of life as death and illness. But they are also credited as benefactors; *e.g.*, in creating utilities in the cosmos and in the invention of fire. The demiurge-trickster is typically ambivalent, tremendously frightful and efficacious, but also frequently limited in power. For example, such tricksters are often incapable of animating the beings that they have molded and must therefore request the help of the supreme being in bringing them to life. They are said to be selfish, lonely, and unhappy, and because of these qualities, they are moved, despite their arrogance, to attempt to relate themselves to or unite with the supreme being.

Role of the trickster

A typically dual composition (involving the coexistence and cooperation of two elements), or even a dualistic opposition (as two opposed elements that function as principles in respect to the actual creation), is found in the Dogon (western Sudanese) notions about Nommo and Yurugu, already mentioned. A series of "words" refers to both principles; *i.e.*, a series of realities and categories can be named that constitute the world in its functional variety, which transcend the simple good–evil opposition, and according to which both Nommo and Yurugu are dualistic "principles" essential to the actual dynamics of the world.

Other dualistic concepts among primitive peoples posit opposite the supreme being a violent and death-bearing "second" figure of a demiurgical type. The character of Erlik in the mythologies of the Central Asiatic Turks (*e.g.*, among the Altaics) is typical.

Erlik is a king of the dead and master of death who assumes the role of a fraudulent and unfortunate collaborator with the supreme being. In stories about the origin of the universe, he appears as an aquatic bird in charge (under the supreme being) of fishing a little earth from the bottom of the primordial sea—a theme also well-known in East European folklore. In other myths, a similar being spits on human beings at the time they are created by God or breathes his bad spirit into man or woman. Elsewhere there is depicted an opposition of two twin brothers, of whom one is the demiurge-creator of good things and the other of death; both, however, are the sons of a mother goddess of heavenly origin. This pattern is exemplified in the Iroquoian myth of Yoskeha and Tawiskaron—a myth curiously reminiscent of certain aspects of the Iranian Zurvanite mythology.

Other ethnological polarities, or pairs of opposites (eastern–western, celestial–terrestrial, solar–lunar divinities, right–left, full moon–dark moon, etc.) are dualistic in the sense of contrasting principles or creating agencies.

THEMES OF RELIGIOUS DUALISM

The sacred and the profane. Among the various themes of religious dualism the opposition between sacred and profane is also important. This distinction, appearing in some sense in nearly every religion, must be particularly acute, however, to qualify a religion as dualistic. Such an intensification of the sacred–profane opposition to the point at which it becomes a dualism is evident in the mid-20th century historian of religions Mircea Eliade's conception of religion. This contrasts time (the *illud tempus*, "those times," of the intact, sacred, primordial creation that are periodically restored by ritual) and the historical time (marked by decay, profaneness, and loss of plentitude and significance).

Good and evil. More pertinent (even if not always dualistic) is the opposition between good and evil, in the various meanings of these words. Whenever the problem of the origin of evil is solved by conceiving the real existence of another principle separate from the prime principle of the world, or by affirming an inner ambivalence, limited sovereignty, or inadequacy of the prime principle,

or of divine beings, a dualism then emerges; and through this good–evil opposition, the problems of theodicy (*i.e.*, of the doctrine of the justification of divine action in a world in which evil is present) are posed. If evil either is, or comes from, a self-existent principle antithetical to the principle of good, then this provides the divinity with a "justification." Such views are completely different from the justification of God in nondualistic religions, especially the monotheistic ones. In monotheistic religions evil does not originate within the divinity nor in general within a divine world (*plērōma*) as it does in Gnosticism; it arises instead from the improper use of freedom by created beings. In monistic religions—all of which are based on the opposition between the One and the many, seen either as an illusion or as the decay or fragmentation of the One—along with a strong ascetic emphasis, there is a notion of evil as being for man a painful and fatal essence that issues from a metaphysical cause or an ontologically negative principle. For the same reason, it is necessary to distinguish between the nondualistic concept of "original sin" in Christian theology and the concept of "previous sin"—in monistic religions with a dualist aspect; whereas "original sin" arises and spreads within the human sphere, "previous sin" is consummated in some sort of a "prologue in heaven" and generates the very existence of the world and of humanity itself.

The problem of evil

Creation and destruction: life and death. Another important dualistic theme is that which opposes life to death based on two opposing metaphysical principles. A typical example of this dualistic opposition is found in Zoroastrianism. Zoroastrian doctrine is strongly vitalistic: Ahriman's chief acolytes are Aēshma (the fury), the Druj Nasu (the deadly agent of putrefaction), Jēh (the infertile whore), and Apaoša (the demon of sterility)—death-bearing forces. There is also a strong vitalistic formulation of these principles in Gnostic doctrines, especially in the Ophite and Barbelo-Gnostic (worshipping Barbelo as the Great Mother of life) varieties, which identify the *pneuma* and the light with the vital substance. At other times the opposition of life and death is formulated in a dialectical manner as a recurring alternation of the two principles. The complex Egyptian opposition between Osiris, the "dead god," who is nonetheless the principle of fecundity and life, and his counterpart Seth has already been mentioned (see above *Egypt and Mesopotamia*). The same dialectic is typical of the "fecundity cults," in which a god-genius of vegetation, a "dying god," is featured, who undergoes a seasonal disappearance and return (not to be interpreted as a "resurrection"). To such vegetation gods, death- or decay-producing figures are sometimes opposed—as Mot (the Death) opposed to Baal, and an infernal and lethal wild boar opposed to Adonis, and (in German religion and mythology) Loki opposed to Baldr. These figures, the agents for disastrous occurrences, were already implicit in the figure of the dying god himself and in his relation to the seasonal cycle of vegetation. To be sure, the growing season is limited; and the new arrival of vegetation each spring (and the wedding of the fertility god) is terminated in the fall by the god's departure to the netherworld (with appropriate lamentation). But the rise of vegetation, though ephemeral, is nonetheless basically benevolent. This complexity is also manifest in those agricultural religions that present themselves as mystery cults (*e.g.*, the Eleusinian mysteries), bestowing upon the initiate a hope for life after death.

The dying god and vegetation cults

But the dualistic theme is far more evident in "mysteriosophy"; *i.e.*, in the "sophic," or "wise," reinterpretation of mysteries (*e.g.*, Orphism). In this context, the divine soul replaces the dying god in the soul's descent from a superior world into the corporeal world—a concept that was later bequeathed to Gnosticism and is especially apparent in its transposed basic vitalism.

A dialectical formulation of the opposition of life and death is also found in the basic theology of Hinduism: with Viṣṇu (Vishnu) cast as the principle of creation (called Nārāyana) and the sustenance of life and Śiva (Shiva) as the principle of destruction and death. The ambivalence of life–death is also found in a series of Hindu divinities (*e.g.*, Śiva, Kālī) and cults whose death-inflicting characteristics are justified in a paradoxical celebration of the recurring triumph of life.

Polytheistic themes. Among the instances of dualistic structure in polytheistic religions are those that oppose celestial and terrestrial, male and female, actual and mythical primordial-chaotic, "diurnal" and "nocturnal," especially when they do so within the context of mythologies and cosmogonies belonging to the ancient world's polytheistic "high cultures" (see above *Egypt and Mesopotamia: Greece and the Hellenistic World*). Such pairs of opposites often provide a framework for polytheistic pantheons that would otherwise appear anarchic or less than comprehensive (see also *Polytheism* above).

FUNCTIONS OF RELIGIOUS DUALISM

Cosmological and cosmogonic functions. The essential function of any religious dualism is obviously ontological—to account for a duality of opposed principles in being—even when the two principles are not regarded as coeternal; and this underlies the cosmological–cosmogonic, anthropological, and sociological functions and expressions of dualism. Both dialectical dualism (*e.g.*, in the fertility cults, Orphic mysteriosophy, and Platonism) and eschatological dualism (*e.g.*, in the Zoroastrian and Manichaean notion of the "mixture" between the two creations good and bad) have a basically cosmological function—the explanation of the structure of the universe. Whenever the concept of a distinct creator, transcendent with respect to his work, is missing (as, for example, in monistic formulations of the Indian type or in polytheistic milieus), dualism has a cosmogonic function—the explanation of the origin of the universe.

On a cosmogonic level, dualistic opposition may also be manifest in the celestial world; *e.g.*, in the late Zoroastrian opposition between the beneficent fixed stars and the planets (which are negative, because they are alleged to proceed in the reverse sense); or else between the world of the Heptad (again the seven planets, under the dominion of the tyrannic archons, or rulers, that cause human passions) and the superior heaven of the Ogdoad (the group of eight divine beings or aeons), as in Gnosticism so also in Mithraism, where the monstrous figure of Leontocephalos (a human figure with a lion's head, belted by a snake with astral signs) represents the power of astral Destiny-Time to be transcended by the soul—a power that is a basic presupposition of astrology and magic. On the other hand, the heaven–earth opposition cannot be regarded as dualistic if the two elements are represented merely as cosmic progenitors.

Anthropological functions. The anthropological functions of dualism (dealing with the nature and destiny of man) are present in all those doctrines that consider man as a duality, or, rather, as an irreconcilable duality of opposed elements. Of particular importance is the opposition between masculine and feminine, in which their opposition involves a remarkable difference in level of being. In mythologies (whether dualistic or not) with a "second" figure, a demiurge, there is frequently a connection between the demiurge and the origin of women (*e.g.*, the myths of Prometheus-Epimetheus in ancient Greece, and of Paliyan in southeast Australia) or between the demiurge and the origin of sexuality (*e.g.*, the myths of the trickster Coyote and of the Gnostic demiurge). In the Platonic theory of man the first incarnation of the soul occurs in a masculine body, and only a subsequent incarnation, marking a later descent of the soul into the world of bodies, is feminine. In Gnosticism (Ophite sects) the vital substance that animates the universe is masculine (active), while the quality of the material world is feminine (passive); and in the last *logion* ("saying") of the Gnostic *Gospel of Thomas*, it is said that Mary will be saved by being made a male; *i.e.*, she will become a "living spirit" (*pneuma*). Gnostic and Manichaean antifeminism, as well as Encratite (and perhaps Orphic) antifeminism, are motivated by their hatred for procreation, which they believe implies the fall of the soul into the material world and its permanent abode there. At other times procreation is explained in terms of a division of a complete, originally androgynous (both male and female) being (as in Plato's *Symposium* and in

Masculine and feminine

the Gnostic *Gospel of Philip*). There are other nondualistic doctrines in which woman is considered to be connected in some way with the origins of evil but not as the embodiment of the evil principle (*e.g.*, in Genesis and the apocryphal late-Judaic *Book of Adam*).

Sociological functions. The sociological functions of religious dualism are less relevant. Among some Australian peoples the "totems" of the two classes of a tribe that intermarry are the Falcon-Eagle (Bundjil), the supreme being, and the Crow (Waang), a demiurge-trickster. According to the Menominee Indians, the highest region of the universe is inhabited by benevolent gods (among whom the supreme being is Mate Hāwātūk) and the inferior region by bad ones; and these two groups are constantly fighting. The Menominee believe that they come from an alliance of families that once belonged to these two groups, whose respective descendants have particular places in the assembly and clearly differentiated functions.

Sociological and economic class oppositions, however, cannot provide a general explication for dualism. All dualities (*e.g.*, in the social structure) are necessarily relevant to religious dualism. On the ethnic level, sociological functions of dualism are found in the Zoroastrian opposition (even if not absolute) between Iran, with its so-called "good religion," and the Turanians, northern plunderers representing the aggressive world of evil. But this can by no means substantiate general hypotheses that explain dualistic oppositions between divinities or groups of divinities as a "projection" of a previously existing opposition between ethnic layers of conquerors and of conquered populations. (U.B.)

Monotheism

Monotheism and polytheism

Monotheism and polytheism, the beliefs in one god or many gods, are often thought of in rather simple terms; *e.g.*, as a merely numerical contrast between the one and the many. The history of religions, however, indicates many phenomena and concepts that should warn against oversimplification in this matter. There is no valid reason to assume, for example, that monotheism is a later development in the history of religions than polytheism. There exists no historical material to prove that one system of belief is older than the other, although many scholars hold that monotheism is a higher form of religion and, therefore, must be a later development, assuming that what is higher came later. Moreover, it is not the oneness of god that counts in monotheism but his uniqueness; one god is not affirmed as the logical opposite to many gods but as an expression of divine might and power.

The choice of either monotheism, or polytheism, however, leads to problems, because neither can give a satisfactory answer to all questions that may reasonably be put. The weakness of polytheism is especially revealed in the realm of questions about the ultimate origin of things, whereas monotheism runs into difficulties in trying to answer the question concerning the origin of evil in a universe under the government of one god. There remains always an antithesis between the multiplicity of forms of the divine manifestations and the unity that can be thought or posited behind them. The one and the many form no static contradistinction: there is rather a polarity and a dialectic tension between them. The history of religions shows various efforts to combine unity and multiplicity in the conception of the divine. Because Christianity is a monotheistic religion, the monotheistic conception of the divine has assumed for Western culture the value of a self-evident axiom. This unquestioned assumption becomes clear when it is realized that for Western culture there is no longer an acceptable choice between monotheism and polytheism but only between monotheism and atheism.

THE SPECTRUM OF VIEWS: MONOTHEISMS AND QUASI-MONOTHEISMS

The basic monotheistic view. Monotheism is the belief in the existence of one god or, stated in other terms, that God is one. As such it is distinguished from polytheism, the belief in the existence of a number of gods, and atheism, the denial of the belief in any god or gods at all. The God of monotheism is the one real god that is believed to exist or, in any case, that is acknowledged as such. His essence and character are believed to be unique and fundamentally different from all other beings that can be considered more or less comparable; *e.g.*, the gods of other religions. The religious term monotheism is not identical with the philosophical term monism, referring to the view that the universe has its origin in one basic principle (*e.g.*, mind, matter) and that its structure is one unitary whole in accordance with this principle; that is, that there is only a single kind of reality, whereas, for monotheism, there are two basically different realities: God and the universe.

The God of monotheism

God in monotheism is conceived of as the creator of the world and man; he has not abandoned his creation but continues to lead it through his power and wisdom; hence, viewed in this aspect, history is a manifestation of the divine will. God has not only created the natural world and the order existing therein but also the ethical order to which man ought to conform and, implicit in the ethical order, the social order. Everything is in the hands of God. God is holy—supreme and unique in being and worth, essentially other than man—and can be experienced as a *mysterium tremendum* ("a fearful mystery") but at the same time as a *mysterium fascinans*, ("a fascinating mystery"), as a mystery approached by man with attitudes of both repulsion and attraction, of both fear and love. The God of monotheism, as exemplified by the great monotheistic religions—Judaism, Christianity, and Islām—is a personal god. In this respect the one god of monotheism is contrasted with the conception of the divine in pantheism, which may also affirm one god or a divine unity. The god of pantheism, however, is impersonal, rather a divine fluid that permeates the whole world including man himself, so that Hinduism can say: *tat tvam asi*, literally "that is you," where "that" refers to the single, supreme reality or principle.

In monotheistic religions the belief system, the value system, and the action system are all three determined in a significant way by the conception of God as one unique and personal being. Negatively considered, the monotheistic conviction results in the rejection of all other belief systems as false religions, and this rejection partly explains the exceptionally aggressive or intolerant stance of the monotheistic religions in the history of the world. The conception of all other religions as "idolatry" (*i.e.*, as rendering absolute devotion or trust to what is less than divine) has often served to justify the destructive and fanatical action of the religion that is considered to be the only true one.

The concept of idolatry

The symbolic language of the monotheistic belief system has no proper terms of its own in speaking of God that cannot be found elsewhere also. God as Creator, Lord, King, Father, and other descriptive names are expressions found in many religions to characterize the various divine beings; the names do not exclusively belong to the religious language of monotheism. This common language is understandable because the monotheistic conception of God differs essentially only in one respect from that of other religions: in the belief that God is one and absolutely unique. Then, consequently, God is regarded as the one and only Creator, Lord, King, or Father. The conception of a divine Word is also to be found in a large number of religions, in accordance with the widespread belief that creation takes place through the word, or speech, of a god.

The extreme positions. The above is the basic monotheistic view. There is, however, a wide range of positions between exclusive monotheism at one extreme and unlimited polytheism at the other. A survey of the various positions may serve to provide a more adequate picture of the complex reality involved in the monotheisms and quasi-monotheisms.

Exclusive monotheism. For exclusive monotheism only one god exists; other gods either simply do not exist at all, or, at most, they are false gods or demons; *i.e.*, beings that are acknowledged to exist but that cannot be compared in power or any other way with the one and only true God. This position is in the main that of Judaism, Christianity, and Islām. While in the Old Testament the other gods in most cases were still characterized as false gods, in later

Judaism and in Christianity as it developed theologically and philosophically the conception emerged of God as the one and only, and other gods were considered not to exist at all.

Ethical and intellectual monotheism

There are two types of exclusive monotheism: ethical monotheism and intellectual monotheism. In ethical monotheism man chooses one god, because that is the god whom he needs and whom he can adore, and that god becomes for him the one and only god. In intellectual monotheism the one god is nothing but the logical result of questions concerning the origin of the world. In many African religions the one god postulated behind the many gods that are active in the world and in the life of man is little more than the prime mover of the universe. He is the intellectual apex necessitated by the system. In Christian theology, heavily influenced as it is by Greek philosophy, both conceptions can be found, usually together.

Unlimited polytheism. On the other hand, there is the extreme position of unlimited polytheism as, for instance, in the classical religions of Greece and Rome: each god has his own name and his own shape, and these are unalienably his and cannot be exchanged with those of any other god (not counting, of course, those cases in which gods are practically each other's duplicate and only bear a different name). The number of divinities is large and in principle unlimited. There are differences of status and power among the gods, of function and sphere of influence, but they are all equally divine. There is, in fact, an ordered pantheon. In unlimited polytheism, the number of gods that are actually worshipped seldom exceeds a few hundred within one religion, but in theory, as in India, millions and millions of gods may be thought to exist (see above *Polytheism*).

The middle positions. Between the extremes of exclusive monotheism and unlimited polytheism are the middle positions of inclusive monotheism and henotheism.

Inclusive monotheism. Inclusive monotheism accepts the existence of a great number of gods but holds that all gods are essentially one and the same, so that it makes little or no difference under which name or according to which rite a god or goddess is invoked. Such conceptions characterized the ancient Hellenistic religions. A well-known example is that of the goddess Isis in the Greco-Roman mystery religion that is called after her. In *The Golden Ass* of Apuleius, the goddess herself speaks: "My name, my divinity is adored throughout all the world, in divers manners, in variable customs, and by many names." Then there follows a number of divine names, and this enumeration ends: "And the Egyptians, which are excellent in all kind of ancient doctrine, and by their proper ceremonies accustomed to worship me, do call me by my true name, Queen Isis."

Henotheism, or kathenotheism. Henotheism (from Greek *heis theos,* "one god")—a belief in worship of one god, though the existence of other gods is granted—also called kathenotheism (Greek *kath hena theon,* "one god at a time")—which literally implies worship of various gods one at a time—has gone out of fashion as a term. It was introduced by the eminent 19th-century philologist and scholar in comparative mythology and religion Max Müller (1823–1900). Many later authors prefer the term

Monolatry

monolatry—which is the worship of one god, whether or not the existence of other deities is posited—to the term henotheism. Both terms mean that one god has a central and dominating position in such a way that it is possible to address this god as if he were the one and only god, without, however, abandoning the principle of polytheism by denying or in any other way belittling the real existence of the other gods, as the above-mentioned forms of monotheism do. Henotheism as a religious concept is at home in cultures with a highly centralized monarchical government. It was especially prevalent in some periods in the history of Babylonia and Egypt.

Alternate positions. *Pluriform monotheism.* The complicated relations that exist between monotheism and polytheism become clear when pluriform monotheism is considered, in which the various gods of the pantheon, without losing their independence, are at the same time considered to be manifestations of one and the same divine substance. Pluriform monotheism is one of the efforts to solve the problem of the coexistence of divine unity and divine pluriformity (multiplicity of forms), which was not recognized by an older generation of scholars, although part of the material was already available. It seems, indeed, that in many parts of the world and in many times religious thinkers have struggled with the perplexing problem of the unity and the pluriformity of the divine.

Nuer and Shilluk religion

The Nuer, a Nilotic pastoral people of the eastern Sudan, venerate a being called Kwoth, the Nuer term for "spirit" (also translated as "God"). He is considered to be the spirit in or of the sky. Like all spirits Kwoth is invisible and omnipresent, but he manifests himself in a number of forms. Each of these manifestations bears a name of its own, but though they are addressed and treated as separate entities, they are essentially nothing but manifestations of the one spiritual being Kwoth and are themselves considered spirits and called *kwoth.* A sacrifice offered to one of these manifestations—*e.g.,* a spirit of air, totem, or place—is not at the same time an offering to another; but all sacrifices, to whatever spirit they are offered, are sacrifices to the supreme Kwoth, or God. Nuer religion is certainly no clear monotheism as it is understood in the Bible and in the Qur'ān (the sacred book of Islām), but neither is it polytheism in the popular sense of the word.

The case of the Nuer is not unique. The related Shilluk people have similar conceptions, and here again the idea of a kind of divine substance that manifests itself in various shapes and under different names is encountered. To give one instance, Macardit is God, but this pronouncement cannot be turned the other way round—it is not permissible to say that God is Macardit. The divine being Macardit represents the dire and fatal aspect of the divinity who orders everything; that is to say, who also sends misfortune and death. In Macardit the contradiction between the creative and constructive and the destructive forces of the divinity is resolved. The positive function of this representation of God lies in the fact that, without diminishing either the power or the justice of the total divinity, it enables man to find an answer for the vexing question of theodicy—the problem of affirming divine justice and goodness in the face of physical and moral evil. That this question is a difficult one, indeed, becomes clear when the reactions of the tribes of Patagonia in a case of death are compared. These tribes believe in a high god, a supreme being, who rules everything and is also responsible for misfortune and death. When someone dies they accuse their god of murder.

Many other instances of pluriform monotheism could be mentioned, and many more presumably still await detection. An interesting pluriform system is that of the Oglala Sioux of the United States, who venerate 16 gods divided into four groups of four. Each group of four forms one god. Thus there are four gods, but these four gods again are one god, Wakan Tanka—the Great Spirit or the Great Mystery.

Religious dualism. Some religions are in the main dualistic; they view the universe as comprising two basic and usually opposed principles, such as good and evil or spirit and matter. Insofar as the conception of a god and antigod rather than that of two gods is encountered, this kind of religion can be considered as another variation of monotheism. In some Gnostic systems (ancient heresies based on esoteric knowledge and the dualism of matter and spirit), Christianity came near to this idea: the demiurge who created the world and man is considered as an evil being and contrasted to the good god. The most important instance of a dualistic religion is the Persian religion Zoroastrianism as founded by Zoroaster (7th–6th century BC) in which Ormazd (the good god) and Ahriman (the evil god) are each other's opposite and implacable enemies. Dualism, the existence of two contrary and, as a rule, mutually inimical principles, must not be confused with the notion of polarity, in which both principles are mutually dependent so that the one cannot exist without the other. Within the religion of Zoroaster, this notion is also found. In the Zoroastrian variation known as Zurvanism, as it is called after the god Zurvān Akarana (Limitless Time), good and evil proceed from one and the

Zoroastrian dualism

same source and in the end they come together again (see above *Religious dualism*).

Pantheism and panentheism. Pantheism and panentheism are not necessarily connected with the notion of either monotheism or polytheism. In both cases the conception of the god or gods is impersonal, which tends, of course, to the conception of one god, of one divine substance, like Spinoza's *deus sive natura*, "god or nature." In pantheism god is immanent, in monotheism god is mostly transcendent, but in polytheism the gods may be either. Pantheism, however, is in most cases more a philosophical than a religious category. Sometimes the term panentheism is used to distinguish between the view that all is in God and that god is in all (see also *Pantheism and panentheism* above).

Primitive monotheism. In connection with monotheism it is necessary to mention the so-called high gods—the remote gods, usually sky gods, found in many primitive and archaic cultures—because this type of divine being has given rise to the theory of primitive monotheism (*Urmonotheismus*). After the Scottish scholar Andrew Lang (1844–1912) had drawn attention to these gods, the Austrian scholar Wilhelm Schmidt (1868–1954) based on their existence in primitive culture and beliefs the theory that the oldest religion of mankind had been monotheistic and that polytheism as well as magic were later degenerations in the course of the history of a pure primeval religion. This theory, defended with great skill and an enormous mass of ethnological material by Schmidt and his collaborators, has long since been proved unsound and was abandoned even by his own students. The connection postulated between the high gods and monotheism has in most respects obscured rather than illuminated the situation. It is true that in many cultures the particular high god is considered as the creator, the founder of the order of the world, and also in some cultures as the reigning god according to whose will everything now happens, but such a god is rarely considered to be the one and only god that counts. Exclusive monotheism is not to be found in either primitive or archaic religions, according to present knowledge. The high god, however, can become a god of exclusive monotheism when circumstances are favourable, at least if he belongs to the active type of high god and not to the intellectual type, which serves mainly as an idea to answer the questions concerning the ultimate origin of things. (See the distinction above between ethical and intellectual monotheism.) This transformation probably occurred in the case of the Islāmic god Allāh. It seems to be more common, however, even for the active type of high god gradually to disappear behind a host of other, often minor, deities who are more concerned with the daily affairs of mankind.

The Deism of the 17th and 18th centuries is often compared to the conception of high gods as *dei otiosi*, "inactive gods," who have created the world and put it into order but after their work was done retreated from the world and left it to run in accordance with the order installed at the creation (see below *Deism*). Not all high gods, however, are inactive.

MONOTHEISM IN THE WORLD RELIGIONS

Classical monotheism. *Religion of Israel and Judaism.* There may be some reason to speak of the Old Testament conception of God as monolatry rather than as monotheism, because the existence of other gods is seldom explicitly denied and many times even acknowledged. The passionate importance given to the proclamation of Yahweh as the one god who counts for Israel and the equally passionate rejection of other gods, however, make it truer to speak of the monotheism of Israel; as in what became the Judaic affirmation of faith, "Hear, O Israel, the Lord is our God, one Lord" (from New English Bible) (Deut. 6:4). The eminent Dutch Old Testament scholar Theodorus C. Vriezen writes: "It is striking how the whole life of the people is seen as dominated by Yahweh and by Yahweh alone. Even if one cannot speak of a strictly maintained monotheistic way of thinking, it is yet clear that faith in Yahweh is the foundation of life for the Israelite." Monotheism is not a matter of mathematics—of opting for the number one as against other numbers—but the conscious choice of a person or group committing himself or themselves to one god rather than to any other ones and putting their faith in that one god; Joshua proclaims: "But as for me and my house, we will serve the Lord" (Josh. 24:15). In Israel the ethical aspect was as important as the exclusiveness of their one god; the prophets stressed the ethical elements of an essentially exclusive God. The God of Israel was a jealous god who forbade his believers to worship other gods. In this respect he differed from other gods in the ancient Near Eastern religions who, as a rule, did not put such exclusive obligation on their adherents.

Yahweh, the God of Israel

In later times—beginning in the 6th century BC and continuing into the early centuries of the Christian Era—Judaic monotheism developed in the same direction as did Christianity and also later Islām under the influence of Greek philosophy and became monotheistic in the strict sense of the word, affirming the one God for all men everywhere.

Christianity. Among the three great monotheistic religions, Christianity has a place apart, because of the trinitarian creed of this religion in its classic forms, in contradistinction to the unitarian creed of Judaism and Islām. The Christian Bible, including the New Testament, has no trinitarian statements or speculations concerning a trinitary deity, only triadic liturgical formulas invoking God the Father, Son, and Holy Spirit. It is true that Christianity also has had its Unitarians, such as the 16th-century Italian theologian Faustus Socinus, but this religion in its three classic forms of Roman Catholicism, Eastern Orthodoxy, and Protestantism acknowledges one God in three Persons: God the Father, God the Son, and God the Holy Spirit. According to Christian theology, this acknowledgment is not a recognition of three gods but that these three persons are essentially one, or as the dogmatic formulation, coined by the early Church Father Tertullian (*c.* 160–after 220), has it: three Persons and one substance. This conception was not accepted without contradiction as is proved by theological disputes of the 3rd and 4th century. It is evident that trinitarian speculation greatly resembles the way of thinking of pluriform monotheism. It is, of course, unlikely that there are any historical connections between these phenomena; both, however, try to solve what is more or less the same problem in more or less the same manner. The main distinction is that Christianity as a monotheistic religion restricts itself to three Persons, though primitive religions have no reason to restrict the number of possible forms of the one divine substance. Like other religions that cover a large territory and have a long history, Christianity appears in a multitude of variations: there is Christian pantheism, Deism, and even, paradoxically, Christian atheism, as exemplified in the mid-20th-century Death of God theologies.

Trinitarianism

Islām. No religion has interpreted monotheism in a more consequential and literal way than Islām. According to Islāmic doctrine the Christian dogma of a trinitarian god is a form of tritheism—of a three-god belief. There is no issue upon which this religion is so intransigent as the one of monotheism. The profession of faith, the first of the so-called Five Pillars of Islām (the basic requirements for the faithful Muslim), states clearly and unambiguously that "there is no God but Allāh," and in accordance with this principle the religion knows no greater sin than *shirk* ("partnership"), the attribution of partners to Allāh; that is to say, polytheism, or anything that may look like it—*e.g.*, the notion of a divine trinity. The Qur'ān declares: "Say: He, Allāh, is one. Allāh, the eternal. Neither has he begotten, nor is he begotten. And no one is his equal" (112). This profession of faith in Allāh as the one god is encountered in a more popular form, for example, in the stories of *The Thousand and One Nights*: "There is no god except Allāh alone, he has no companions, to him belongs the power and he is to be praised, he gives life and death and he is mighty over all things." In only one respect has the uncompromising monotheism of Islām shown itself to be vulnerable; *i.e.*, in the doctrine of the Qur'ān as uncreated and coeval with Allāh himself.

Monotheistic elements in ancient Middle Eastern and Mediterranean religions. *Egyptian religion.* Egyptian religion is of special interest with regard to the various topics treated in this article, for in it are found polytheism,

henotheism, pluriform monotheism, trinitary speculations, and even a kind of monotheism. Especially in the time of the New Kingdom (16th–11th century BC) and later, there arose theological speculations about many gods and the one god, involving concepts that belong to the realm of pluriform monotheism. These ideas are especially interesting when related to trinitarian conceptions, as they sometimes are. In a New Kingdom hymn to Amon are the words: "Three are all gods: Amon, Re and Ptah . . . he who hides himself for them [mankind] as Amon, he is Re to be seen, his body is Ptah." As Amon he is the "hidden god" (*deus absconditus*); in Re, the god of the sun, he becomes visible; as Ptah, one of the gods of the earth, he is immanent in this world.

Monotheistic reform of Akhenaton

Much attention has been given to the reform of Egyptian religion as effected by the pharaoh Akhenaton (Amenophis IV) in the 14th century BC. This reform has been judged in many ways, favourably and unfavourably; it is, however, clear that Akhenaton's theology, if not fully monotheistic, in any case strongly tends toward monotheism. It is even possible to follow the gradual development of his ideas in this direction. At first he only singled out Aton, one of the forms of the sun god, for particular worship, but gradually this kind of henotheism developed in the direction of exclusive monotheism and even took on the intolerance peculiar to this religious concept. The names of the other gods were to be deleted. This un-Egyptian intolerance was probably the main reason for the speedy decline of this creed.

Babylonian religion. As far as is known, monotheism was largely absent from Babylonian religion. There henotheism seems to have been very important, since a person could choose one god for particular worship as if he were the only god.

Greco-Roman religions. The classic religions of Greece and Rome were in the main purely polytheistic, but in later times tendencies arose, partly stimulated by philosophy and later also by Judaism and Christianity, toward inclusive monotheism. The hymn to Zeus by the Stoic philosopher Cleanthes (*c.* 330–*c.* 230 BC) is the best known document of this process. It praises Zeus as the essence of divinity in all gods, creator and ruler of the cosmos, omnipotent, the giver of every gift, and the father of mankind. In the mystery religions of the Greco-Roman world and in the religious philosophies of later antiquity, such as Neoplatonism, Neopythagoreanism, and others, inclusive monotheism was more or less the rule.

Monotheistic elements in Indian and Chinese religions. The religions of India and China show an astonishing multiplicity of form, but exclusive monotheism, unless imported or stimulated by foreign influences, seems to be absent. All other phenomena treated in this survey of monotheism, however, are to be found in their religions. Inclusive monotheism and pantheism fit very well with the Indian notions of religion, particularly in Hinduism, as is witnessed by the reflections on Brahman, the self of the world, and *Ātman*, the self of individual man. As the *Upaniṣads* say: "Truly, in the beginning existed this Brahman, that only knew itself, saying: I am Brahman." Although in many cases one god, such as Śiva (Shiva) or Vishnu, receives nearly all the attention of the faithful, this emphasis never leads to a negation of other gods as such. Jainism does not differ in this respect. Only the religion of the Sikhs, heavily influenced by Islām, can be said to teach a kind of exclusive monotheism.

Buddhism teaches in essence that there are no gods in the full sense of the world. Gods are higher beings, but they belong to the cosmos and are as much in need of salvation as man is.

The ancient Chinese religion of the heaven (Shang Ti T'ien) is comparable to the religions that proclaim a high god as creator and ruler of the world, but monotheism has not resulted from this conception. Neither has Taoism led to monotheism. (T.P.v.B.)

Theism

Theism is the view that all limited or finite things are dependent in some way on one supreme or ultimate reality of which one may also speak in personal terms.

Theism's view of God can be clarified by contrasting it with that of deism, of pantheism, and of mysticism. Deism closely resembles theism; but for the deist, God is not involved in the world in the same personal way. He has made it, so to speak, or set the laws of it—and to that extent he sustains it in being. But subject to this final and somewhat remote control, God, as the deist sees him, allows the world to continue in its own way. This view simplifies some problems, especially those that arise from the scientific account of the world: one does not have to allow for any factor that cannot be handled and understood in the ordinary way. God is in the shadows or beyond; and, though men may still in some way centre their lives upon him, this calls for no radical adjustment at the human or finite level. The deist proceeds, for most purposes at least, as if there were no God—or only an absent one; and this approach is especially true of man's understanding of the world. This is why deism appealed so much to thinkers in the time of the first triumphs of modern science. They could indeed allow for God, but they had "no need of that hypothesis" in science or in their normal account of things. Religion, being wholly superadded, was significant only in a manner that involved little else in the world or in the way man lives. The theist, on the other hand, questions this view and seeks in various ways (as noted below) to bring man's relation to God into closer involvement with the way he understands himself and the world around him.

Theism, pantheism, and mysticism

Theism also sharply contrasts with pantheism, which identifies God with all that there is; and with various forms of monism, which regards all finite things as parts, modes, limitations, or appearances of some one ultimate Being, which is all that there is. Some types of absolute Idealism, a philosophy of all-pervading Mind, while regarding every finite thing as comprising some limitation of the one whole of Being, seek also to retain the theistic element in their view of the world; and they do this normally—as in the works of A.E. Taylor, Andrew Pringle-Pattison, or G.F. Stout—by stressing the role of unifying finite centres, such as self-conscious human beings, in the way the universe as a whole functions. But there is no recognition here of the finality of what is technically known as "the distinctness of persons." The theist, by contrast, considers the world to be quite distinct from its Author or Creator, human life being thus in no sense strictly the life of God, while also making room for a peculiarly intimate involvement of God in the world and in human life.

Mysticism in practice comes close to theism; but mystical thought, and much of its practice, has often involved a repudiation of the proper reality of finite things and sometimes (as in a work by W.T. Stace, *Mysticism and Philosophy*) tends to dismiss all of the finite manifold or multiplicity of things as some wholly unreal phantasm that has no place in the one undiversified Being, which alone is real. Theism is very far removed from ideas of this kind.

THEISM IN WESTERN THOUGHT

God encountered as person. The idea that the world, as man understands it in a finite way, is dependent on some reality altogether beyond his comprehension, perfect and self-sustained but also peculiarly involved in the world and its events, is presented with exceptional sharpness and discernment in the Old Testament, whence it became a formative influence in Hebrew history and subsequently in Christianity and Islām. Behind the creation stories; behind the patriarchal narratives, like that of Jacob at Bethel (Gen. 28) or wrestling with his strange visitor at Penuel (Gen. 32); and behind the high moments of prophecy, like Isaiah's famous vision in the Temple (Isa. 6), and of moving religious experience in the Psalms, in the Book of Job, and (with remarkable explicitness) in some well-known passages, like the story of Moses at the burning bush (Ex. 3)—behind all of these there lies a sense of some mysterious, all-encompassing reality by which man is also in some way addressed and which he may also venture to address in turn. Moses wished to see God, to have some explicit sign that could convince the people and establish his own authority; but he was shown, instead, that this is just what he could not have: all that he could be assured

Old Testament theophanies

of was that God is real and is bound to be—"I am who I am," he was told. On the other hand, in the throes of this humbling and staggering experience, Moses began to learn also what was expected of him and how his people should live and be led. The God who was so strange and elusive was somehow found to be a God who "talked" to him and with whom people could "walk." The same seemingly bewildering claim of remoteness, almost to the point of unreality, linked with a compelling explicitness and closeness, is also found in other cultures, as illustrated below. This claim presents the reflective thinker with the twofold problem of theism, viz., how, in the first place, a reality as remote and mysterious as the God of theism—the "wholly other," in the famous words of the German theologian Rudolf Otto—can be known at all; and, second, how, if it can be known, it can be spoken of in precise and intimate ways and encountered as a person.

The existence of God. There have been many attempts to establish the existence of one supreme and ultimate Being—whom in religion one speaks of as God—and some of these have been given very precise forms in the course of time.

The influence of Plato and Aristotle. The pattern for many of these was laid down in ancient Greece by Plato. He taught about God mostly in mythical terms, stressing the goodness of God (as in *The Republic* and *Timaeus*) and his care for man (as in the *Phaedo*); but in the *Phaedrus,* and much more explicitly in the *Laws,* he presented a more rigorous argument, based on the fact that things change and are in motion. Not all change comes from outside; some of it is spontaneous and must be due to "soul" and ultimately to a supreme or perfect soul. Whether God so conceived quite gives the traditional theist all that he wants, however, is not certain. For God, in Plato, fashions the world on the pattern of immutable Forms and, above all, on "the Good," which is "beyond being and knowledge"; *i.e.,* it is transcendent and beyond the grasp of thought. But Plato's combination of the notion of the transcendent, which is also supremely good, and the argument from change, provided the model for much of the course that subsequent philosophical arguments were to take. Aristotle made the argument from motion more precise, but he coupled it with a doubtful astronomical view and a less theistic notion of God, who, as the unmoved mover, is the ultimate source of all other movement, not by expressly communicating it but by being a supreme object of aspiration, all appetite and activity being in fact directed to some good. Aristotle thus set the pattern for the more deistic view of God, whereas the theist, taken in the strict sense, turns more for his start and inspiration to Plato.

The five ways of Aquinas

The causal argument. The argument for the existence of God inferred from motion was given a more familiar form in the first of the five ways of St. Thomas Aquinas, five major proofs of God that also owed much to the emphasis on the complete transcendence of God in the teaching of Plotinus, the leading Neoplatonist of the 3rd century AD, and his followers. (The word that Plotinus used for the ultimate but mysterious dependence of all things on God is emanation; but this characterization was not understood by him, as it has been by some later thinkers, as questioning the genuine independent existence of finite things.) In the first way, Aquinas put forward the view that all movement implies, in the last analysis, an unmoved mover; and though this argument, as he understood it, presupposes certain views about movement and physical change that may not be accepted today, it does make the main point that finite processes call for some ground or condition other than themselves.

This becomes more explicit in the second way, which proceeds from the principle that everything must have an "efficient cause"—*i.e.,* a cause that actively produces and accounts for it—to the notion of a first cause required to avoid an infinite regress, or tracing of causes endlessly backward. As normally found, the idea of efficient causality, in respect to change and process, has many difficulties; and some would prefer to speak instead of regular or necessary sequence. But a more serious objection stresses the apparent inconsistency of thinkers who invoke a general principle of causality and then exempt the alleged first cause. As the child is apt to put it, "Who then made God?" To this a defender of St. Thomas, or at least of the present approach to the idea of God, would reply that the first cause is not supposed to be itself a member of any ordinary causal sequence but altogether beyond it, an infinite reality not itself a part of the natural or temporal order at all. This point, in fact, is what the third way, starting from the contingency of the world, brings out more explicitly. Nothing explains itself, and all other explanations fall short of showing in any exhaustive way why anything is as it is, or why there is anything at all. But it is also hard to suppose that things just happen to be. Nothing could come out of just nothing, and so the course of events as men find and explain them points to some reality that is not itself to be understood or explained in the normal way at all: it is Explanation with a capital *E,* as it were, that is seen to be necessitated by all that there is—of whose nature, however, nothing may be directly discerned beyond the inevitability of its being as the ultimate or unconditioned ground of all else and in this way transcendent or utterly mysterious in itself.

This way of thinking of the being and necessity of God has been impressively presented in the mid-20th century by notable thinkers like Austin Farrer, E.L. Mascall, and H.P. Owen and also by the present writer (see below *Bibliography*). Generally known as the cosmological approach to the idea of God, it has much in common with the insistence on the transcendence of God in recent theology.

The arguments of Anselm and Descartes

The ontological argument. Scholars have often converged upon the same theme in what appears to be a very different line of argument, namely the ontological one, with which are associated especially the names of St. Anselm, first of the Scholastic philosophers (in the 11th century), and René Descartes, first major modern philosopher (in the mid-17th century). Proponents of this argument try to show that the very idea of God implies his existence. God is the greatest or most perfect being. If the attribute of existence, however, is not included in man's concept of God, he can then think of something more perfect, viz., that which has existence as well. Critics, such as Gaunilo—a monk of Marmoutier—in Anselm's day and Immanuel Kant—one of the major architects of modern philosophy—many centuries later, have fastened on the weakness that existence is not a predicate or attribute in the same way, at least, as colour or shape; but there have been highly ingenious attempts by influential religious thinkers of today to restate the argument in an acceptable form. (See especially the writings of Charles Hartshorne and Norman Malcolm.) Others find in the argument an oblique and needlessly elaborate way of eliciting the feeling that there must be some reality that exists by the very necessity of its own nature and to which everything else directs man's thought.

The reference to value and design. Attempts to arrive at the idea of God in somewhat more comprehensible terms are reflected in the references to value and design in the fourth and fifth ways of St. Thomas; this approach, however, has been given a more explicit presentation and critical discussion in the works of David Hume, a mid-18th-century Scottish Skeptic, and in Kant. The main idea of the teleological argument, as it is called, is that of the worth and purpose, or apparent design, to be found in the world. This purposiveness is taken to imply a supreme Designer. It has been questioned, however (by Kant, for example), whether this argument can really get started without presupposing some feature of the causal argument. The presence of seemingly purposeless features of the world and of much that is positively bad, like wickedness and suffering, while always embarrassing for a theistic view, presents peculiar difficulties here. For the arguer is now throwing hostages to fortune in the shape of a special assessment of the way things actually happen, which goes far beyond the mere requirement of some ultimate ground, whatever the world appears to be like. The arguments from worth and design have, however, one considerable advantage, viz., that they provide a fairly straightforward way of learning about the nature of God and of ascribing a certain aim and character to him from

one's understanding of the phenomena that he is required to explain. The supreme Designer or Architect is known from his works, especially perhaps as reflected in the lives of men; and this approach opens up one way of speaking of God, not just as mysterious power behind the world but as some reality whom man may come to know in a personal way from the way the world goes and from his understanding of what it means.

The views of Ward, Tennant, and Teilhard

Many thinkers in the late 19th and early 20th centuries sought to establish man's knowledge of God in the way suggested through his understanding of himself and the world; and of these the most notable and valuable still today are the British theists James Ward, a psychologist, and F.R. Tennant, a philosophical theologian. But the work of thinkers like Pierre Teilhard de Chardin, a Jesuit paleoanthropologist, and the spate of discussion that he has provoked are also relevant here; and such work, in turn, owes much—directly or otherwise—to the work of evolutionary thinkers like Samuel Alexander and Henri Bergson and of modern scientists like Julian Huxley.

The problem of particular knowledge of God. If the central theme of traditional theism, viz., that the finite world depends in some way on one transcendent and infinite Being, can be sustained, then a crucial problem presents itself at once: the question of how a being whose essence can never be known to man—who, as infinite, is bound to be beyond the grasp of reason and to remain wholly mysterious—how such a being can be said to be known at all, much less known and experienced in the close and intimate personal ways that the theist makes equally central to his claim. Part of the answer is that the theist does not claim to fathom the ultimate mystery of God or to know him as he is in himself. All that is claimed on this score is that man sees the inevitability of there being God in the contingent and limited character of everything else; and though this line of thought could not be adopted for any finite existence—since one could not normally affirm in any sensible way the existence of anything without specifying in some measure, however slight, what it is like—one can, nonetheless, regard the case of God as unique and not subject to the conditions of finite intelligibility. In these ways, an insight or intuition into the being of God may be claimed without a commitment to anything about his nature beyond the sort of completeness or perfection required to account for there being limited finite things. This insight is much in line with the "deliverances of religious consciousness" in which it is claimed that God is "hidden," is "past finding out," that his ways are not man's ways, that he is eternal, uncreated, and so on. But the theist still has a major problem on his hands, for he also makes a central issue of the claim that God can be known—"met" and "encountered" in some way—indeed, that some very bold affirmations about God and his dealings with men may be made.

Theism and natural theology. Theists have tried to deal with this problem in various ways. One of these ways is their use of the doctrine of analogy, which owes a great deal to the teaching of St. Thomas Aquinas. Various types of analogy are distinguished in the traditional doctrine; but the central claim is that certain predicates, such as "love," "faithfulness," or "justice," may be affirmed of God in whatever way may reflect his involvement as the author of the limited realities, such as man, of which such predicates may be affirmed in the normal, straightforward way. The difficulty with this procedure is that, whatever it yields, the content of faith is still very thin and remote, far from the warm fellowship of personal relations. Most of the traditional sponsors of the doctrine admit this and contend, therefore, that the findings of their "natural theology," as it is called, must be supplemented by that of revelation or of divine disclosure. Theism, in fact, is hardly conceivable without some doctrine of revelation. But even if the theologian says that God takes the initiative in communicating himself to man, the epistemological problem remains of how men's essentially finite minds can apprehend anything pertaining to infinite or eternal Being.

Revelation and authority

Theism and religious experience. At this point, recourse is sometimes had to authority, the authority of a sacred book, an institution, or a system of doctrines, or one of divinely implanted images. But there must at least be some initial justification of an authority, to say nothing of an evaluation of rival claims. A more attractive solution, then, especially for those who stress the personal involvement of God in men's lives, is one posed in terms of religious experience. Such experience is usually given prominence in theistic contexts. It is sometimes understood in terms of paranormal phenomena, like hearing voices or seeing visions, which have no natural origin, or like being in some peculiar psychical state. Some of the faithful believe that God literally speaks to them (or spoke in times past to prophets) in this way. A more subtle view holds that men have reason to regard certain experiences as their clue to what they should say of God in his relation to them. The question then arises of how these experiences should be recognized; and various answers are given, such as that which stresses the formative influence (within such experiences) of the initial insight into the being of God and the patterning of the experiences, in themselves and in wider ramifications, as a result. Much use is made in this context of the analogy with men's knowledge of one another. Men do not know one another's minds, it is alleged, as they know their own but only as mediated through bodily states and behaviour. So a man may come to know God, who in his essence is impenetrable to him, from the impact that he makes within experiences and events that one would otherwise understand and handle just as one does other finite occurrences. In the molding and perpetuating of such experiences, prominence is given to imagination and to the place of figurative terms and symbolism. These forms have therefore a place of special importance in theistic types of religion, the personal encounter being extended and deepened through art and literature, song, dance, myth, and ritual. This fact, in turn, presents problems for thought and practice, since the art forms and ritual must not be allowed to take wing on their own and thereby be loosed from the discipline and direction of the proper dynamic of religious life.

Theism and religious language. Preoccupation with the forms in which religious life expresses itself has led some theistic writers to lean heavily on the contribution made to religious understanding today by studies of religious language. In some cases this concern has carried with it, as generally in much linguistic philosophy of today, a skeptical or agnostic view of the transcendent factor in religion. It is hard to see, however, how attenuations of this kind could be strictly regarded as forms of theism; though clearly, within their more restricted scope, they can retain many of the other characteristics of theism, such as the stress on personal involvement and response. This tendency is very marked in some recent studies of religion, in which the inspiration and form of theism are retained without the substance—though how long and how properly are moot points. There are others who, while retaining the transcendent reference of theism, look for the solution of the central problem less in the substance of religious awareness and in varieties of experience than in the modes of articulation and religious language. Controversy centres, to a great degree, on which of these approaches is the most fruitful.

The question of transcendence in theism

In the work of some theists today, the preoccupation with language is also combined with the existentialist stress on personal involvement and commitment. A good example of this approach is found in the work of I.T. Ramsey, the bishop of Durham, who, in spite of his insistence on disclosure situations, in which something peculiarly significant becomes alive to man, seemed to concede more than a theist should to the skeptical strain in recent studies of religious language.

The nature of God in modern thought. Modern thought has thrown new light on issues, both old and new, regarding the nature of God.

Theism and incarnation. The core of human personality has often been thought to be man's moral existence, and, accordingly, theists have often taken this fact to be the main clue to the way they are to think of divine perfection and to the recognition of a peculiar divine involvement in the world. Prominence is thus accorded to the high ethical teaching and character of saints and prophets, who have

a special role to play in transmitting the divine message. In some religions this tendency culminates in doctrines of incarnation, of God manifesting himself expressly in refined or perfected human form. This trend is peculiarly marked in the Christian religion, in which the claim is usually made that a unique and "once for all" incarnation of God has occurred in Christ. Incarnational claims seem certainly to take their place easily in some main forms of theism. The vindication of such claims, however—especially today—relies much on consideration of the personal factor in religion generally.

For these and related reasons, the theist today may find himself calling to his aid certain other disciplines that centre upon men as persons, such as psychology and anthropology. Not all of the forms and findings of these studies favour the theist, and he should take special note of their challenge when they seem hostile, for they may touch him at his tenderest spot. He may, on the other hand, find in such studies, and certain general literature that borders on kindred themes, substantial help in reconstructing his case in the full context of contemporary thought and culture.

Humanism and transcendence. It is indeed from certain modern studies of man and his environment that some of the most disturbing challenges to the theist have come. For it has been argued that the very idea of God, as well as the more specific forms that it takes, emanate from man's emotional needs for succour and comfort. It is in fact man himself, it is said, who has created God in his own image, and the attempt is made to substantiate this view from accounts of the proclivity of men, especially in early times, to personify natural objects—rivers, trees, mountains, and so forth—and, in due course, to confer peculiar properties upon them, leading in time to the notion of some superbeing in whom these powers and properties are concentrated. The classical statement of this position appeared before the development of anthropology and the modern systematic study of religions, viz., in David Hume's essay "The Natural History of Religion" (1757). This short but splendidly lucid and challenging work set the pattern for the more scientific and empirical studies of religion that began to take shape in the 19th century in pioneer work by E.B. Tylor, a British ethnologist and anthropologist, in his *Primitive Culture* (1871), and by Sir James Frazer, an ethnographer and historian of religion, in his *Golden Bough* (1890–1915). But a corrective to this approach was soon provided by other scholars equally renowned, who started from the historical and empirical evidence available to them at the time. Andrew Lang, a Scottish litterateur, drew attention to the phenomenon, among very early peoples, of the High God, a Supreme Being who created himself and the earth and dwelt at one time on earth. John H. King, in *The Supernatural: Its Origin, Nature and Evolution* (1892), stressed the importance of the element of mystery in all religions, and another pioneer of religious anthropology, R.R. Marett, showed how extensively the savage ascribes the mysteries of life and power to a supernatural source. Lucien Lévy-Bruhl, a French sociologist, noted the pervasiveness of prelogical factors in primitive mentality, and Rudolf Otto, the most famous name in this context, found evidence in early forms of religion of a response to "the wholly other," the *mysterium tremendum et fascinans.*

Role of the sciences of religion

The idea of a finite God. Concern with the problem of evil—*i.e.*, with reconciling the existence of evil with that of a good God—becomes acute for thinkers who rest their case mainly on what they find in the world around them; and this has led many to retreat to the notion of a finite God, according to which the world may be under the direction of a superior being who is nonetheless limited in power, though not in goodness. This is a serious alternative to the idea of a supreme and unlimited source of all reality as found in the usual forms of theism. Indeed, it is a moot point whether the idea of a finite God should be classified as a form of theism. It does come close to traditional theism, however, in its insistence on the unity and absolute benevolence of God. There are clearly advantages in the notion of God as a limited being, especially where evil is concerned; for though one could still insist that God intends nothing that is not wholly good, he can now account for extensive suffering and other ills on the basis of the limits to God's power. He is doing his utmost, the finitist holds, but there are things—refractory materials or explicitly evil powers—that he has not yet subdued, though hopefully he will eventually do so. There is also induced in this way a sense of urgency in man's own obligation to cooperate with God—to be a "fellow worker." God will clearly need his help though he himself is in the vanguard of the battle against evil. Thus, those who incline to the idea of a finite God usually have been activist in thought and practice.

There are also grave difficulties to be met. For if a thinker has recourse to the idea of God simply to account for what is otherwise bewildering in the finite course of things, he may find no warrant for the inference involved and indeed may find himself desperately clinging to what is sometimes called "the God of the gaps" (*i.e.*, of the gaps in man's explanations). If, on the other hand, he starts from the inherently incomplete character of finite explanation as such, or from the contingency of finite things, nothing short of an infinite or absolute God will meet the case. In addition, the usual attitude of religious people, or of what is sometimes known as "the religious consciousness," is that of a profound assurance and serenity that presupposes that God is "all in all" and beyond any possibility of being thwarted. It is also questionable whether the attitude of worship is appropriate for a limited being, however superior he may be to man.

The views of James, Perry, and the process philosophers

Among the outstanding advocates of the idea of a finite God were, at the turn of the 20th century, the U.S. Pragmatist William James and some of his disciples, notably Ralph Barton Perry. Thus, it is not surprising that a closely similar notion arising in the mid-20th century finds its main inspiration and support in the United States, viz., in the work of process philosophers, such as Charles Hartshorne and Schubert Ogden, who have developed some of the leading ideas of A.N. Whitehead, an eminent metaphysician. In their view, God is himself in process of fulfillment in some kind of identification with the world, which at the same time leaves him distinct in some sense from the universe, which he permeates and unifies. There are grave and admitted paradoxes in this view; and, in spite of the remarkable ingenuity of its advocates and their logical nimbleness, it is not clear that the paradoxes can be sustained nor that the difficulties that are shared with the simpler notion of a finite God can be overcome. Much in recent religious thought centres on this issue.

Theism in Islām. The Muslim faith owes much to the Semitic outlook from which the Old Testament and Christianity arose. It centres on a transcendent personal deity; but, in its regard for the holiness and majesty of God, it rejects incarnational doctrines as a form of blasphemy. There is, however, a paradoxical side to one form of Islām: while insisting that God is all in all, it sometimes tends to represent all of man's own actions as the action of God within him and thus has some tendency to identify man with God. This tendency, most marked in the mysticism of the Ṣūfīs, seems, as respects its monism, to veer away from theism but seems, as respects the sense of devotion and personal excitation that it inspires, to be in line with the more explicit forms of theism. In its main form, Islām, with its quite exceptional sense of the transcendence of God, is one of the most distinctively theistic religions, though at odds with the incarnational factor in Christian theism.

THEISM IN EASTERN THOUGHT

The trend toward the testing of theistic thought in the crucible of the special disciplines was continued not only in further anthropological studies (see *The Worship of the Sky-God*, by E.O. James) but also in extensive scholarly studies and translations of the sacred books of the great religions of the East.

Hindu theism. It was noted, for example, that the Vedic hymns that appear in the earliest Hindu scriptures contain significant intimations of a sense of "the wonder of existence," "the outpourings," as Savepalli Radhakrishnan, the former philosopher-president of India, has expressed it, "of poetic minds who were struck by the immensity of

the universe and the inexhaustible mystery of life." Note was taken also of early manifestations of henotheism, a view that exalts several deities to the first place. The theme of some one supreme reality, the first principle, or the supreme self becomes more explicit in the *Upaniṣads*, ancient Hindu scriptures, while retaining a sense of its ineffableness. One hears of "the way of silence" and of the ultimate absorption of all into the one supreme reality, the "one who breathes breathless." This one is variously conceived in its relation to finite things; and although the transcendent reference is rarely absent, there is not the same recognition of the distinctness of finite beings that there is in Western theism or of the eternal self being involved in the world in a personal way. The *Upaniṣads* have, in fact, a variety of themes and emphases, tending generally toward a monistic and mystical philosophy; but on occasion the theistic element is very marked, as in the *Kaṭha* and the *Śvetāśvatara* books of the *Upaniṣads*. The absolutist and the theistic views are not always felt to be exclusive. This climate of thought has set the course for much of subsequent Hinduism, in which, along with the persistence of the monistic strain, the theistic note is sounded much more distinctly, especially in the doctrine and practice of *bhakti*—devotion to a personal God who bestows grace. In the famous *Bhagavadgītā* (probably 3rd or 4th century BC), a classic of religious literature, and in the teaching of the Brahmin Rāmānuja (11th century), considered the founder of the Viśiṣṭadvaita (qualified nondualism) school, the flowering of the more theistic side of Hinduism is found. In the Śaiva-siddhānta theology of South Indian Śaivism (a major cult of Hinduism), there is a further insistence that the soul, in being united with God, is not annihilated or negated but only fused into the likeness of God, who, in turn, is always in loving pursuit of the soul. This doctrine makes the system "perhaps the highest form of theism that India was ever to develop" (R.C. Zaehner). In the closing words of the *Bhagavadgītā* is an insistence on a love of God for man and of man for God that represents a decisive turning point in the history of Hinduism:

> Think on me, worship me, sacrifice to me, pay me homage, so shalt thou come to me, I promise thee truly, for I love thee well. Give up all things of *dharma*, turn to me only as thy refuge. I will deliver thee from all evil. Have no care.

Theism in the *Upaniṣads*

This theology has been well reflected in the 20th century in the devotionalism of Gandhi and in the writings of Sri Aurobindo, a philosopher and Yoga devotee, which reflect an indwelling of the divine within the world and a summons to high moral endeavour on the part of man that comes close to theism without explicitly accepting it.

Buddhism and theism. The same diversity of strains is found in Buddhism. Though Buddhism was at one time regarded as an atheistic religion leading to total elimination of self in a state of Nirvāṇa, a close examination of the evidence—in the Pāli *Tipiṭaka*, for example, the canon of the Theravāda school of Buddhism—leads to a revision in favour of the view that the seeming negativism of early Buddhist scriptures and the rejection of metaphysics reflect chiefly the caution arising from a profound recognition of the characterless elusiveness of the transcendent. And although the Buddhist doctrine of compassion and its rigorous intellectual and moral discipline may lack something of the warmth of a close personal commitment, the Buddhist adoration of the Buddha and of the *bodhisattvas* (those on their way to Enlightenment) afforded much scope to the religious responses that find their full expression in overt theism. This trend became more marked in the more popular forms of Buddhism and in the mythologies that centre upon the idea of the *bodhisattvas*.

Theism in other religions. In the same way, the seeming agnosticism of Confucian religion is qualified by its teaching about a power from beyond the world working for justice within it, a "Heaven-ordained relationship" that provides the basis of ethics and induces a deep consciousness of individuality. This trend became intensified in the conflations that resulted from the extension of Buddhism into China.

In the doctrines of Sikhism, a religion of the eastern Punjab that combines certain Muslim and Hindu elements, stress is laid upon personal awareness of God as a central and unifying factor in religion. In doctrine though not always in practice, however, the Sikhs reject every notion of an avatar, or incarnation. The religion of the Jains is nontheistic in theory, but the great figures of its tradition come to function as gods in popular religion. For a period in ancient Persia, there was established in the teaching of Zoroaster (Zarathushtra) a form of ethical monotheism in which the god Ahura Mazdā is the creator of the physical and moral world—though limited, for a time at least, by an opposing principle of evil (Ahriman).

Theism in Confucianism, Sikhism, and Jainism

The clue to the theistic element in the religions of primitive peoples may well be found in an observation by H.H. Farmer, a British philosophical theologian:

> We may surmise that at moments of living prayer and worship there is in primitive man a turning to *a* god as if he were in fact the one and only God, though without any expressly formulated denial of the existence of others; for the time being, the god worshipped fills the whole sphere of the divine.

(H.D.L.)

Deism

Deism, as the word is customarily employed, describes an unorthodox religious attitude that found expression among a group of English writers beginning with Edward Herbert (later 1st Baron Herbert of Cherbury) in the first half of the 17th century and ending with Henry St. John, 1st Viscount Bolingbroke, in the middle of the 18th century. In general, however, it refers to what can be called natural religion, the acceptance of a certain body of religious knowledge that is inborn in every person or that can be acquired by the use of reason, as opposed to knowledge acquired through either revelation or the teaching of any church.

The concept of natural religion

NATURE AND SCOPE

Though an initial use of the term occurred in 16th-century France, the later appearance of the doctrine on the Continent was stimulated by the translation and adaptation of the English models. The high point of Deist thought occurred in England from about 1689 through 1742, during a period when, despite widespread counterattacks from the established Church of England, there was relative freedom of religious expression following upon the Glorious Revolution that ended the rule of James II and brought William and Mary to the throne. Deism took deep root in 18th-century Germany after it had ceased to be a vital subject of controversy in England.

At times in the 19th and early 20th centuries, the word Deism was used theologically in contradistinction to theism, the belief in an immanent God who actively intervenes in the affairs of men. In this sense Deism was represented as the view of those who reduced the role of God to a mere act of creation in accordance with rational laws discoverable by man and held that, after the original act, God virtually withdrew and refrained from interfering in the processes of nature and the ways of man. So stark an interpretation of the relations of God and man, however, was accepted by very few Deists during the flowering of the doctrine, though their religious antagonists often attempted to force them into this difficult position. Historically, a distinction between theism and Deism has never had wide currency in European thought. As an example, when encyclopaedist Denis Diderot, in France, translated into French the works of Anthony Ashley Cooper, 3rd earl of Shaftesbury, one of the important English Deists, he often rendered "Deism" as *théisme*. The term is not in current usage as a metaphysical concept, and its significance is really limited to the 17th and 18th centuries.

THE HISTORICAL DEISTS

The English Deists. In 1754–56, when the Deist controversy had passed its peak, John Leland, an opponent, wrote a historical and critical compendium of Deist thought, *A View of the Principal Deistical Writers that Have Appeared in England in the Last and Present Century; with Observa-*

tions upon Them, and Some Account of the Answers that Have Been Published Against Them. This work, which began with Lord Herbert of Cherbury and moved through the political philosopher Thomas Hobbes, Charles Blount, the Earl of Shaftesbury, Anthony Collins, Thomas Woolston, Matthew Tindal, Thomas Morgan, Thomas Chubb, and Viscount Bolingbroke, fixed the canon of who should be included among the Deist writers. In subsequent works Hobbes usually has been dropped from the list and John Toland included, though he was closer to pantheism than most of the other Deists were. Herbert was not known as a Deist in his day, but Blount and the rest who figured in Leland's book would have accepted the term Deist as an appropriate designation for their religious position. Simultaneously, it became an adjective of opprobrium in the vocabulary of their opponents. Bishop Edward Stillingfleet's *Letter to a Deist* (1677) is an early example of the orthodox use of the epithet.

Herbert's five fundamental beliefs

In Lord Herbert's treatises five religious ideas were recognized as God-given and innate in the mind of man from the beginning of time: the belief in a supreme being, in the need for his worship, in the pursuit of a pious and virtuous life as the most desirable form of worship, in the need of repentance for sins, and in rewards and punishments in the next world. These fundamental religious beliefs, Herbert held, had been the possession of the first man, and they were basic to all the worthy positive institutionalized religions of later times. Thus, differences among sects and cults all over the world were usually benign, mere modifications of universally accepted truths; they were corruptions only when they led to barbarous practices such as the immolation of human victims and the slaughter of religious rivals.

In England at the turn of the 17th century this general religious attitude assumed a more militant form, particularly in the works of Toland, Shaftesbury, Tindal, Woolston, and Collins. Though the Deists differed among themselves and there is no single work that can be designated as the quintessential expression of Deism, they joined in attacking both the existing orthodox church establishment and the wild manifestations of the dissenters. The tone of these writers was often earthy and pungent, but their Deist ideal was sober natural religion without the trappings of Catholicism and the High Church in England and free from the passionate excesses of Protestant fanatics. In Toland there is great emphasis on the rational element in natural religion; in Shaftesbury more worth is ascribed to the emotive quality of religious experience when it is directed into salutary channels. All are agreed in denouncing every kind of religious intolerance because the core of the various religions is identical. In general, there is a negative evaluation of religious institutions and the priestly corps who direct them. Simple primitive monotheism was practiced by early men without temples, churches, and synagogues, and modern men could readily dispense with religious pomp and ceremony. The more elaborate and exclusive the religious establishment, the more it came under attack. A substantial portion of Deist literature was devoted to the description of the noxious practices of all religions in all times, and the similarities of pagan and Roman Catholic rites were emphasized.

The Deists who presented purely rationalist proofs for the existence of God, usually variations on the argument from the design or order of the universe, were able to derive support from the vision of the lawful physical world that Sir Isaac Newton had delineated. Indeed, in the 18th century, there was a tendency to convert Newton into a matter-of-fact Deist—a transmutation that was contrary to the spirit of both his philosophical and his theological writings.

When Deists were faced with the problem of how man had lapsed from the pure principles of his first forebears into the multiplicity of religious superstitions and crimes committed in the name of God, they ventured a number of conjectures. They surmised that men had fallen into error because of the inherent weakness of human nature; or they subscribed to the idea that a conspiracy of priests had intentionally deceived men with a "rout of ceremonials" in order to maintain power over them.

The role of Christianity in the universal history of religion became problematic. For many religious Deists the teachings of Christ were not essentially novel but were, in reality, as old as creation, a republication of primitive monotheism. Religious leaders had arisen among many peoples—Socrates, Buddha, Muḥammad—and their mission had been to effect a restoration of the simple religious faith of early men. Some writers, while admitting the similarity of Christ's message to that of other religious teachers, tended to preserve the unique position of Christianity as a divine revelation. It was possible to believe even in prophetic revelation and still remain a Deist, for revelation could be considered as a natural historical occurrence consonant with the definition of the goodness of God. The more extreme Deists, of course, could not countenance this degree of divine intervention in the affairs of men.

Natural religion was sufficient and certain; the tenets of all positive religions contained extraneous, even impure elements. Deists accepted the moral teachings of the Bible without any commitment to the historical reality of the reports of miracles. Most Deist argumentation attacking the literal interpretation of Scripture as divine revelation leaned upon the findings of 17th-century biblical criticism. Woolston, who resorted to an allegorical interpretation of the whole of the New Testament, was an extremist even among the more audacious Deists. Tindal was perhaps the most moderate of the group. Toland was violent; his denial of all mystery in religion was supported by analogies among Christian, Judaic, and pagan esoteric religious practices, equally condemned as the machinations of priests.

The Deists were particularly vehement against any manifestation of religious fanaticism and enthusiasm. In this respect Shaftesbury's *Letter Concerning Enthusiasm* (1708) was probably the crucial document in propagating their ideas. Revolted by the Puritan fanatics of the previous century and by the wild hysteria of a group of French exiles prophesying in London in 1707, Shaftesbury denounced all forms of religious extravagance as perversions of true religion. These false prophets were directing religious emotions, benign in themselves, into the wrong channels. Any description of God that depicted his impending vengeance, vindictiveness, jealousy, and destructive cruelty was blasphemous. Because sound religion could find expression only among healthy men, the argument was common in Deist literature that the preaching of extreme asceticism, the practice of self-torture, and the violence of religious persecutions were all evidence of psychological illness and had nothing to do with authentic religious sentiment and conduct. The Deist God, ever gentle, loving, and benevolent, intended men to behave toward one another in the same kindly and tolerant fashion.

French Deists

Deists in other countries. Ideas of this general character were voiced on the Continent at about the same period by such men as Pierre Bayle, a French philosopher famous for his encyclopaedic dictionary, even though he would have rejected the Deist identification. During the heyday of the French Philosophes in the 18th century, the more daring thinkers—Voltaire among them—gloried in the name Deist and declared the kinship of their ideas with those of Rationalist English ecclesiastics, such as Samuel Clarke, who would have repudiated the relationship. The dividing line between Deism and atheism among the Philosophes was often rather blurred, as is evidenced by *Le Rêve de d'Alembert* (written 1769; "The Dream of d'Alembert"), which describes a discussion between the two "fathers" of the *Encyclopédie*: the Deist Jean Le Rond d'Alembert and the atheist Diderot. Diderot had drawn his inspiration from Shaftesbury, and thus in his early career he was committed to a more emotional Deism. Later in life, however, he shifted to the atheist materialist circle of the Baron d'Holbach. When Holbach paraphrased or translated the English Deists, his purpose was frankly atheist; he emphasized those portions of their works that attacked existing religious practices and institutions, neglecting their devotion to natural religion and their adoration of Christ. The Catholic Church in 18th-century France did not recognize fine distinctions among heretics, and Deist and atheist works were burned in the same bonfires.

German Deists

English Deism was transmitted to Germany primarily through translations of Shaftesbury, whose influence upon

thought was paramount. In a commentary on Shaftesbury published in 1720, Gottfried Wilhelm Leibniz, a Rationalist philosopher and mathematician, accepted the Deist conception of God as an intelligent Creator but refused the contention that a god who metes out punishments is evil. A sampling of other Deist writers was available particularly through the German rendering of Leland's work in 1755 and 1756. H.S. Reimarus, author of many philosophical works, maintained in his *Apologie oder Schutzschrift für die vernünftigen Verehrer Gottes* ("Defense for the Rational Adorers of God") that the human mind by itself without revelation was capable of reaching a perfect religion. Reimarus did not dare to publish the book during his lifetime, but it was published in 1774–78 by Gotthold Ephraim Lessing, one of the great seminal minds in German literature. According to Lessing, common man, uninstructed and unreflecting, will not reach a perfect knowledge of natural religion; he will forget or ignore it. Thus, the several positive religions can help men achieve more complete awareness of the perfect religion than could ever be attained by any individual mind. Lessing's *Nathan der Weise* (1779; "Nathan the Sage") was noteworthy for the introduction of the Deist spirit of religion into the drama; in the famous parable of the three rings, the major monotheistic religions were presented as equally true in the eyes of God. Although Lessing's rational Deism was the object of violent attack on the part of Pietist writers and the more mystical thinkers, it influenced such men as Moses Mendelssohn, a German Jewish philosopher who applied Deism to the Jewish faith. Immanuel Kant, the most important figure in 18th-century German philosophy, stressed the moral element in natural religion; moral principles are not the result of any revelation but originate from the very structure of man's reason. English Deists, however, continued to influence German Deism. Witnesses attest that virtually the whole officer corps of Frederick the Great was infected with Deism and that Collins and Tindal were favourite reading in the army.

American Deists

By the end of the 18th century, Deism had become a dominant religious attitude among intellectual and upper class Americans. Benjamin Franklin, the great sage of the Colonies and then of the new republic, summarized in a letter to Ezra Stiles, president of Yale College, a personal creed that almost literally reproduced Herbert's five fundamental beliefs. The first three presidents of the United States also held Deistic convictions, as is amply evidenced in their correspondence. "The ten commandments and the sermon on the mount contain my religion," John Adams wrote to Thomas Jefferson in 1816. (F.E.M.)

Agnosticism

The word agnosticism was first publicly coined in 1869 at a meeting of the Metaphysical Society in London by T.H. Huxley, a British biologist and champion of the Darwinian theory of evolution. He coined it as a suitable label for his own position. "It came into my head as suggestively antithetical to the 'Gnostic' of Church history who professed to know so much about the very things of which I was ignorant."

NATURE AND KINDS OF AGNOSTICISM

Huxley's statement brings out both the fact that agnosticism has something to do with not knowing, and that this not knowing refers particularly to the sphere of religious doctrine. Etymology, however, and now common usage, do permit less limited uses of the term. The Soviet leader Lenin, for instance, in his *Materialism and Empirio-Criticism* (1908), distinguished the extremes of true Materialism on the one hand and the bold Idealism of George Berkeley, an 18th-century Idealist, on the other. He recognized as attempted halfway houses between them the "agnosticisms" of the Scottish Skeptic David Hume and the great German critical philosopher Immanuel Kant—agnosticisms that here consisted in their contentions about the unknowability of the nature, or even the existence, of "things-in-themselves" (realities beyond appearances).

Huxley's nonreligious agnosticism. The essence of Huxley's agnosticism—and his statement, as the inventor of the term, must be peculiarly authoritative—was not a profession of total ignorance, nor even of total ignorance within one special but very large sphere; rather, he insisted, it was "not a creed but a method, the essence of which lies in the rigorous application of a single principle," viz., to follow reason "as far as it can take you"; but then, when you have established as much as you can, frankly and honestly to recognize the limits of your knowledge. It is the same principle as that later proclaimed in an essay on "The Ethics of Belief" (1876) by the British mathematician and philosopher of science W.K. Clifford: "It is wrong always, everywhere and for everyone to believe anything upon insufficient evidence." Applied by Huxley to fundamental Christian claims, this principle yields characteristically skeptical conclusions: speaking, for example, of the Apocrypha (ancient scriptural writings excluded from the biblical canon), he wrote: "One may suspect that a little more critical discrimination would have enlarged the Apocrypha not inconsiderably." In the same spirit, Sir Leslie Stephen, 19th-century literary critic and historian of thought, in *An Agnostic's Apology, and Other Essays* (1893), reproached those who pretended to delineate "the nature of God Almighty with an accuracy from which modest naturalists would shrink in describing the genesis of a black beetle."

Contrasts with atheism, Skepticism, Positivism

Agnosticism in its primary reference is commonly contrasted with atheism thus: "The Atheist asserts that there is no God, whereas the Agnostic maintains only that he does not know." This distinction, however, is in two respects misleading: first, Huxley himself certainly rejected as outright false—rather than as not known to be true or false—many widely popular views about God, his providence, and man's posthumous destiny; and second, if this were the crucial distinction, agnosticism would for almost all practical purposes be the same as atheism. It was indeed on this misunderstanding that Huxley and his associates were attacked both by enthusiastic Christian polemicists and by Friedrich Engels, the co-worker of Karl Marx, as "shame-faced atheists," a description that is perfectly applicable to many of those who nowadays adopt the more comfortable label.

Agnosticism, moreover, is not the same as Skepticism, which, in the comprehensive and classical form epitomized by the ancient Greek Skeptic Sextus Empiricus (2nd and 3rd centuries AD), confidently challenges not merely religious or metaphysical knowledge but all knowledge claims that venture beyond immediate experience. Agnosticism is, as Skepticism surely could not be, compatible with the approach of Positivism, which emphasizes the achievements and possibilities of natural and social science—though most agnostics, including Huxley, have nonetheless harboured reserves about the more authoritarian and eccentric features of the system of Auguste Comte, the 19th-century founder of Positivism.

Religious agnosticism. It is also possible to speak of a religious agnosticism. But if this expression is not to be contradictory, it has to be taken to refer to an acceptance of the agnostic principle, combined either with a conviction that at least some minimum of affirmative doctrine can be established on adequate grounds, or else with the sort of religion or religiousness that makes no very substantial or disputatious doctrinal demands. If these two varieties of agnosticism be admitted, then Huxley's original agnosticism may be marked off from the latter as (not religious but) secular and from the former as (not religious but) atheist—construing "atheist" here as a word as wholly negative and neutral as "atypical" or "asymmetrical." These, without pejorative insinuations, mean merely "not typical" or "not symmetrical" (the atheist is thus one who is simply without a belief in God).

Huxley himself allowed for the possibility of an agnosticism that was in these senses religious—even Christian—as opposed to atheist. Thus, in another 1889 essay "Agnosticism and Christianity," he contrasted "scientific theology," with which "agnosticism has no quarrel," with "Ecclesiasticism, or, as our neighbours across the Channel call it, Clericalism"; and his complaint against the latter's proponents was not that they reach substantive conclusions different from his own but that they maintain "that

it is morally wrong not to believe certain propositions, whatever the results of strict scientific investigation of the evidence of these propositions." The second possibility, that of an agnosticism that is religious as opposed to secular, was realized perhaps most strikingly in the Buddha (Gautama). Typically and traditionally, the ecclesiastical Christian has insisted that absolute certainty about some minimum approved list of propositions concerning God and the general divine scheme of things was wholly necessary to salvation. Equally typically, according to the tradition, the Buddha sidestepped all such speculative questions. At best they could only distract attention from the urgent business of salvation—salvation, of course, in his own very different interpretation.

The agnosticism of the Buddha

HISTORICAL ANTECEDENTS OF MODERN AGNOSTICISM

It is convenient to distinguish the antecedents of secular agnosticism from those of religious agnosticism.

Antecedents of secular agnosticism. The ancestry of modern secular and atheist agnosticism may be traced back to the Sophists and to Socrates in the 5th century BC; not, of course, the "Socrates" of Plato's *Republic*—the would-be founding father of an ideal totalitarian state—but the shadowy historical Socrates supposedly hailed by the oracle of Apollo's Delphi as the wisest of men—who knew what, and how much, he did not know. But the most important and immediate source of such agnostic ideas was surely Hume, while Hume's successor Kant may well be seen as the prime philosophical inspirer of religious reactions against them.

Huxley, as noted above, demanded that a thinker recognize and accept the limits of his knowledge. In taking it that these limits do not include either the findings of a general positive natural theology or the contents of a particular special divine revelation, Huxley was accepting a Humean critique. (It is significant that Huxley's study of Hume was the most sympathetic appraisal to be published in the 19th century.) Hume's critique is found in his *Enquiry Concerning Human Understanding* (first published in 1748 under another title), which attempts, in the manner of Locke and later Kant, to determine the limits of man's possible knowledge, and in his posthumous *Dialogues Concerning Natural Religion* (1779).

The work of David Hume

Two sections of the *Enquiry* refer directly to these limits: "Of a Particular Providence and of a Future State" and "Of Miracles." In the first, Hume starts from his basic Empiricist claims: that, generally, "matters of fact and real existence" cannot be known a priori (prior to and apart from experience); and that, particularly, one cannot know a priori that any thing or kind of thing either must be or cannot be the cause of any other thing or kind of thing. These considerations dispose of all the classical arguments for the existence of God other than the argument to design—that the structure and order of the universe and its constituents implies a design and a designer. But here, Hume urges, argument from experience can find no purchase because both the supposed effect, the universe as a whole, and the putative cause, God, are essentially unique and incomparable. Later, in his *Dialogues*, he develops the suggestion—which he acknowledges as stemming from the 3rd-century-BC philosopher Strato of Lampsacus, next but one after Aristotle as head of his Lyceum—that whatever order man discerns should be attributed to the universe itself and not to any postulated outside cause.

In the section "Of Miracles," Hume takes his stand on the agnostic principle: "A wise man . . . proportions his belief to the evidence." He then argues that no attempt to appeal to the alleged occurrence of miracles—conceived as authoritative endorsements by a power beyond and greater than nature—can succeed in establishing the truth of a claim to constitute special divine revelation. Hume's distinctive contribution here is methodological: the contention that the principles and presuppositions upon which the critical historian must rely, in first interpreting the remains of the past as historical evidence and in then building up from this evidence his account of what actually happened, are such as to make it impossible for him "to prove a miracle and make it a just foundation for any such system of religion."

In this two-phase attack, Hume challenged what was in his day, and long remained, the standard framework for systematic Christian apologetics. Indeed, the contrary contentions—of the possibilities, both of developing a positive natural theology and of establishing the authenticity of a supposed revelation by discovering endorsing miracles—were defined as essential and constitutive dogmas of Roman Catholicism by decrees of the first Vatican Council of 1869–70.

In view of the future history of Western thought, it must be emphasized that Hume's position, like Kant's, was (officially) that knowledge in this area is practically impossible. This thesis is stronger than that of those who simply confess that they just do not know:

> The God-men say when die go sky
> Through pearly gates where river flow,
> The God-men say when die we fly
> Just like eagle, hawk and crow—
> Might be, might be; I don't know.
> (Aboriginal song from the Northern Territory, Australia.)

Yet Hume's thesis was, on the other hand, weaker than that of his 20th-century neo-Humean successors, the logical positivists of the Vienna Circle, who held that any talk about a transcendent God must be "without literal significance." This view was presented brilliantly, and in an uncompromisingly drastic form, by A.J. Ayer in his *Language, Truth and Logic* (2nd ed., 1946). Similar conclusions were reached less high-handedly by several contributors to *New Essays in Philosophical Theology* (ed. by A. Flew and A. MacIntyre, 1955).

Antecedents of religious agnosticism. Looking backward, it is possible now to see what Hume himself did not know—that his attack on the possibility of a positive natural theology had to a considerable extent been anticipated by 14th-century Christian Scholastics: generally, by William of Ockham; and, with particular reference to the lack of a priori knowledge of causal relations, by Nicholas of Autrecourt.

The claims of Hume and Kant—and, indeed, those of the logical positivists and their successors—about the practical, or theoretical, impossibility of such knowledge should also be compared with the long traditions of "negative theology." Such a theology maintains that the nature of God passes so far beyond the comprehension of any creature that God must be characterized largely or entirely by indirection—as Infinite, as Incomparable, and so on. Thus Thomas Aquinas, the foremost Scholastic of the 13th century—who contrived on other occasions to tell his readers as much as his most practical church could wish about the deeds, plans, and demands of the Ineffable—nevertheless had his agnostic moments as well. But he did elaborate a doctrine of so-called analogical predication designed to show how it is possible for finite creatures to say and to understand something positive about God by means of comparisons with known entities or qualities. By contrast, the 12th-century philosopher Moses Maimonides, often dubbed anachronistically "the Jewish Aquinas," had been much more drastic than his successor, "the Christian Maimonides," in his insistence that everything that can be truly said about the Creator—not excluding the proposition that he exists—has to be construed as purely negative.

Negative theology

Although it is clearly possible to speak of a religious agnosticism without self-contradiction, the foregoing considerations suggest the difficulty of intermingling religious and agnostic concerns. The easiest case is that in which the religion is altogether without metaphysical content: thus, one of Huxley's biographers reports that the 19th-century Scottish sage Thomas Carlyle "taught him that a deep sense of religion was compatible with an entire absence of theology." The next simplest case is that in which worship is combined with a total noncommitment about the attributes of the object of worship:

Tension between the religious and agnostic

> He is not a male: He is not a female: He is not a neuter.
> He is not to be seen: He neither is nor is not.
> When He is sought He will take the form in which
> He is sought.
> It is indeed difficult to describe the name of the Lord.
> (Poem from the Telugu, inscribed on a cult object in the Royal Ontario Museum.)

In its original setting this expression of a Hindu piety has power and charm. Yet its intellectual inadequacy becomes manifest when the doctrine of the Unknowable in the broad synthetic system of Herbert Spencer, a late-19th-century evolutionary philosopher, is recalled. For to affirm, as Spencer did, the existence of a being about whom absolutely nothing else can be said is a rather comical hypostatization (taking of an abstraction as real), which is surely indiscernible from affirming no being at all. Nor, perhaps, is it any great improvement to aver that much else can indeed be said about him, but only in words that here must bear an extraordinary meaning—unless, of course, those meanings can be specified. It was the suggestion that the goodness of God might thus be goodness in a quite unusual sense—what would elsewhere be called badness—that provoked the ire of John Stuart Mill, a mid-19th-century Empiricist, against certain developments from Sir William Hamilton's "Philosophy of the Unconditioned." Mill wrote: "I will call no being good, who is not what I mean when I apply that epithet to my fellow creatures."

The third, and surely the most promising, way in which the reconciliation may be attempted is by essaying some distinction between the essence or the internal nature of God and his external relations with the creation. It may then be suggested that, whereas man's knowledge of the former must be at least exiguous and at worst simply lacking, he can nevertheless know as much as he needs to know about the latter. As to the rest, he should be reverently agnostic.

INCOMPATIBILITY WITH FIDEISM

What cannot, however, by any means be squared with agnosticism in Huxley's sense are attempts to transmute the very limitations of human knowledge into grounds for accepting some wholly unevidenced faith. Such transmutations have been made in the interests of many mutually irreconcilable systems, and they apparently remain perennially attractive to thinkers with a different understanding of the ethics of belief.

St. Augustine of Hippo, near the end of the 5th century, felt the challenge of classical Skepticism in Cicero's *Academica* and *De natura deorum* ("On the Nature of the Gods") and gave his response in *Contra academicos* ("Against the Academics"). Skepticism, he thought, can be overcome only by revelation. The orthodox Muslim philosopher and mystic al-Ghazālī (late 11th century) deployed Skeptical arguments similarly, as a propaedeutic, or study preparatory to the acceptance of his rival revelation. With the rediscovery in the 16th century of the works of Sextus Empiricus, a course of Skepticism became commonly a preliminary to fideist commitment. Fideism is the thesis that truth in religion is accessible only to faith. The course persuaded the inquirer that reason cannot attain truth; yet certainty in true religious belief was still thought absolutely necessary for salvation. Martin Luther was speaking for his times (first half of the 16th century) when he thundered against the extremely cautious and restricted agnosticism of Desiderius Erasmus, foremost figure of the northern Renaissance: "Spiritus sanctus non est Skepticus" ("The Holy Spirit is not a Skeptic").

The only resort was, it seemed, faith: whether the easy-going Roman Catholic faith of the 16th-century Skeptic Michel de Montaigne; the polemical Counter-Reformation fervour of his contemporary Gentian Hervet, veteran of the Council of Trent and Latin translator of the *Adversus mathematicos* (1569; "Against the Pundits") of Sextus Empiricus; or, one century later, the vestigial Huguenot loyalty of Pierre Bayle—stocker of a great arsenal of secular argument, the *Dictionnaire historique et critique* (1695–97).

Locke's definition of faith

The decisive objection to any and every such rationally unfounded flight into faith was posed by John Locke, the 17th-century British Empiricist, who set a tone of coolly unfervent Anglicanism for the following century:

> We may as well doubt of our being, as we can whether any revelation from God be true. So that faith is a settled and sure principle of assent and assurance, and leaves no room for doubt or hesitation. Only we must be sure that it be a divine revelation, and that we understand it right: else we shall expose ourselves to all the extravagancy of enthusiasm, and all the error of wrong principles . . . (*An Essay Concerning Human Understanding*, Book IV, ch. xvi, 14).

Many thinkers have agreed that it is all very well to depreciate the potentialities of unaided natural reason and to insist that if man is to have any knowledge of God this must depend largely or wholly upon whatever special steps God may have taken to reveal himself; and they have also agreed that, if man's commitment of faith is not to be arbitrary and frivolous, then he clearly must have some good reason for believing, first, that there is a God who has so revealed himself, and, second, that his preferred candidate—and not one of its innumerable rivals—truly is that revelation.

These points are crucial—both for the appreciation of the history of ideas and for a reasonable contemporary understanding. Clearly, they were upheld by Aquinas, who in the *Summa contra gentiles*—before proceeding to present his own reasons for accepting Christianity, rather than Islām, as the authentic revelation—applied that same word frivolous to any such unsupportable commitment. Again, Judah ha-Levi, an early 12th-century Jewish poet and philosopher, has been authoritatively described as "concerned to bring men to a mystical and non-rational appreciation of religious truths" by his Skeptical attacks on the established Aristotelian natural theology. Yet ha-Levi's main work, entitled *Kuzari: The Book of Proof and Argument in Defence of the Despised Faith*, does in fact offer rational evidences of the truth of Judaism.

Skeptical propaedeutics to faith are now out of fashion. But the same challenge applies to all of the various responses to Kant's famous invitation: "I have found it necessary to deny *knowledge* in order to make room for *faith*" (Preface to the *Critique of Pure Reason*). Natural theology may, indeed, for Hume's reasons as reinforced by Kant, be impossible. The way of religious discovery may indeed be mystical experience, personal encounter with the divine Thou, or whatever else. But there is, and can be, no substitute for a man's having some sound grounds for identifying his experience not only as really mystical but also as experience of the real God; for holding his faith in some putative revelation not only to be real religious faith but also to be faith in a genuine revelation of the Real; and so on.

REJECTIONS OF THE AGNOSTIC PRINCIPLE

Anyone who insists on the foregoing touchstones may still be agnostic as well as religious. What cannot consist with agnosticism is a calculated commitment to faith seen as altogether without evidential warrant. The classic example of such commitment was provided in the 17th century by the Wager Argument of the French mathematician Blaise Pascal, who assumed, for the sake of the argument, that "reason can decide nothing here" and then urged that the only sane bet is Roman Catholicism; for we have nothing but this one short life to lose, and all eternity to win.

Blaise Pascal and William James

Pascal's Wager Argument is unsound because, on its own stated assumption of total and inescapable ignorance, the gambler is not entitled to limit the betting options to two—and to one particular two, at that. A similarly parochial inattention to the variety of candidacies for belief has characterized most fideists. Thus Søren Kierkegaard, an influential mid-19th-century Danish lay theologian, happily glorified the essential irrationality of religious faith, while taking it always that faith will, of course, be Protestant. Elsewhere, Pascal himself did notice, and tried to meet, some of the competition; his neglect here is the more remarkable because his wager was originally imported into Christendom from Islām (see Miguel Palacios, *Los precedentes de Pari de Pascal*). What makes it a landmark is that it constituted a direct, reasoned rejection of the agnostic principle—a rejection in which the reason proposed for believing was explicitly a motive for self-persuasion rather than some evidence of truth. Thus, when William James, a pre-World War I American psychologist and philosopher, in *The Will to Believe*, developed the best known systematic attack on that principle it was, rightly, Pascal whom he hailed as his first inspiration. James distinguished those hypotheses that, for any individual, represent psychologi-

cally "live options" from those that do not, and he urged that, when evidential grounds are lacking, the choice may properly be determined by one's passional nature. For men often have to act on some unproved hypothesis, and sometimes such firm commitments may help to make the belief come true. Consider, for example, some belief that a man is trustworthy. The objections are that belief in the existence of God is clearly not of this case, and generally that to act decisively on some hypothesis does not require the agent to believe it as a known truth. (A.G.N.F.)

Atheism

The dialectic of the argument between forms of belief and unbelief raises questions concerning the most perspicuous delineation, or characterization, of atheism, agnosticism, and theism. It is necessary not only to probe the warrant for atheism but also carefully to consider what is the most adequate definition of atheism. This section will start with what have been some widely accepted, but still in various ways mistaken or misleading, definitions of atheism and move to more adequate formulations that better capture the full range of atheist thought and more clearly separate unbelief from belief and atheism from agnosticism. In the course of this delineation the section also will consider key arguments for and against atheism.

ATHEISM AS REJECTION OF RELIGIOUS BELIEFS

A central, common core of Judaism, Christianity, and Islām is the affirmation of the reality of one, and only one, God. Adherents of these faiths believe that there is a God who created the universe out of nothing and who has absolute sovereignty over all his creation: this includes, of course, human beings—who are not only utterly dependent on this creative power but also sinful and who, or so the faithful must believe, can only make adequate sense of their lives by accepting, without question, God's ordinances for them. The varieties of atheism are numerous, but all atheists reject such a set of beliefs.

Atheism, however, casts a wider net and rejects all belief in "spiritual beings," and to the extent that belief in spiritual beings is definitive of what it means for a system to be religious, atheism rejects religion. So atheism is not only a rejection of the central conceptions of Judeo-Christianity and Islām, it is, as well, a rejection of the religious beliefs of such African religions as that of the Dinka and the Nuer, of the anthropomorphic gods of classical Greece and Rome, and of the transcendental conceptions of Hinduism and Buddhism. Generally atheism is a denial of God or of the gods, and if religion is defined in terms of belief in spiritual beings, then atheism is the rejection of all religious belief.

It is necessary, however, if a tolerably adequate understanding of atheism is to be achieved, to give a reading to "rejection of religious belief" and to come to realize how the characterization of atheism as the denial of God or the gods is inadequate.

ATHEISM AND THEISM

To say that atheism is the denial of God or the gods and that it is the opposite of theism, a system of belief that affirms the reality of God and seeks to demonstrate his existence, is inadequate in a number of ways. First, not all theologians who regard themselves as defenders of the Christian faith or of Judaism or Islām regard themselves as defenders of theism. The influential 20th-century Protestant theologian Paul Tillich, for example, regards the God of theism as an idol and refuses to construe God as a being, even a supreme being, among beings or as an infinite being above finite beings. God, for him, is "being-itself," the ground of being and meaning. The particulars of Tillich's view are in certain ways idiosyncratic, as well as being obscure and problematic, but they have been influential; and his rejection of theism, while retaining a belief in God, is not eccentric in contemporary theology, though it may very well affront the plain believer.

Second, and more important, it is not the case that all theists seek to demonstrate or even in any way rationally to establish the existence of God. Many theists regard such a demonstration as impossible, and fideistic believers (*e.g.*, Johann Hamann and Søren Kierkegaard) regard such a demonstration, even if it were possible, as undesirable, for in their view it would undermine faith. If it could be proved, or known for certain, that God exists, people would not be in a position to accept him as their sovereign Lord humbly on faith with all the risks that entails. There are theologians who have argued that for genuine faith to be possible God must necessarily be a hidden God, the mysterious ultimate reality, whose existence and authority must be accepted simply on faith. This fideistic view has not, of course, gone without challenge from inside the major faiths, but it is of sufficient importance to make the above characterization of atheism inadequate.

Atheism and fideism

Finally, and most important, not all denials of God are denials of his existence. Believers sometimes deny God while not being at all in a state of doubt that God exists. They either willfully reject what they take to be his authority by not acting in accordance with what they take to be his will, or else they simply live their lives as if God did not exist. In this important way they deny him. Such deniers are not atheists (unless we wish, misleadingly, to call them "practical atheists"). They are not even agnostics. They do not question that God exists; they deny him in other ways. An atheist denies the existence of God. As it is frequently said, atheists believe that it is false that God exists, or that God's existence is a speculative hypothesis of an extremely low order of probability.

Yet it remains the case that such a characterization of atheism is inadequate in other ways. For one it is too narrow. There are atheists who believe that the very concept of God, at least in developed and less anthropomorphic forms of Judeo-Christianity and Islām, is so incoherent that certain central religious claims, such as "God is my creator to whom everything is owed," are not genuine truth-claims; *i.e.*, the claims could not be either true or false. Believers hold that such religious propositions are true, some atheists believe that they are false, and there are agnostics who cannot make up their minds whether to believe that they are true or false. (Agnostics think that the propositions are one or the other but believe that it is not possible to determine which.) But all three are mistaken, some atheists argue, for such putative truth-claims are not sufficiently intelligible to be genuine truth-claims that are either true or false. In reality there is nothing in them to be believed or disbelieved, though there is for the believer the powerful and humanly comforting illusion that there is. Such an atheism, it should be added, rooted for some conceptions of God in considerations about intelligibility and what it makes sense to say, has been strongly resisted by some pragmatists and logical empiricists.

While the above considerations about atheism and intelligibility show the second characterization of atheism to be too narrow, it is also the case that this characterization is in a way too broad. For there are fideistic believers, who quite unequivocally believe that when looked at objectively the proposition that God exists has a very low probability weight. They believe in God not because it is probable that he exists—they think it more probable that he does not—but because belief is thought by them to be necessary to make sense of human life. The second characterization of atheism does not distinguish a fideistic believer (a Blaise Pascal or a Kierkegaard) or an agnostic (a T.H. Huxley or a Leslie Stephen) from an atheist such as Baron d'Holbach or Thomas Paine. All believe that "There is a God" and "God protects humankind," however emotionally important they may be, are speculative hypotheses of an extremely low order of probability. But this, since it does not distinguish believers from nonbelievers and does not distinguish agnostics from atheists, cannot be an adequate characterization of atheism.

It may be retorted that to avoid apriorism and dogmatic atheism the existence of God should be regarded as a hypothesis. There are no ontological (purely a priori) proofs or disproofs of God's existence. It is not reasonable to rule in advance that it makes no sense to say that God exists. What the atheist can reasonably claim is that there is no evidence that there is a God, and against that background he may very well be justified in asserting that there is

no God. It has been argued, however, that it is simply dogmatic for an atheist to assert that no possible evidence could ever give one grounds for believing in God. Instead, atheists should justify their unbelief by showing (if they can) how the assertion is well-taken that there is no evidence that would warrant a belief in God. If atheism is justified, the atheist will have shown that in fact there is no adequate evidence for the belief that God exists, but it should not be part of his task to try to show that there could not be any evidence for the existence of God. If the atheist could somehow survive the death of his present body (assuming that such talk makes sense) and come, much to his surprise, to stand in the presence of God, his answer should be, "Oh! Lord, you didn't give me enough evidence!" He would have been mistaken, and realize that he had been mistaken, in his judgment that God did not exist. Still, he would not have been unjustified, in the light of the evidence available to him during his earthly life, in believing as he did. Not having any such postmortem experiences of the presence of God (assuming that he could have them), what he should say, as things stand and in the face of the evidence he actually has and is likely to be able to get, is that it is false that God exists. (Every time one legitimately asserts that a proposition is false one need not be certain that it is false. "Knowing with certainty" is not a pleonasm.) The claim is that this tentative posture is the reasonable position for the atheist to take.

Burden-of-proof argument

An atheist who argues in this manner may also make a distinctive burden-of-proof argument. Given that God (if there is one) is by definition a very recherché reality—a reality that must be (for there to be such a reality) transcendent to the world—the burden of proof is not on the atheist to give grounds for believing that there is no reality of that order. Rather, the burden of proof is on the believer to give some evidence for God's existence; *i.e.*, that there is such a reality. Given what God must be, if there is a God, the theist needs to present the evidence, for such a very strange reality. He needs to show that there is more in the world than is disclosed by common experience. The empirical method, and the empirical method alone, such an atheist asserts, affords a reliable method for establishing what is in fact the case. To the claim of the theist that there are in addition to varieties of empirical facts "spiritual facts" or "transcendent facts," such as it being the case that there is a supernatural, self-existent, eternal power, the atheist can assert that such "facts" have not been shown.

Fallibilistic atheism

It will, however, be argued by such atheists, against what they take to be dogmatic aprioristic atheists, that the atheist should be a fallibilist and remain open-minded about what the future may bring. There may, after all, be such transcendent facts, such metaphysical realities. It is not that such a fallibilistic atheist is really an agnostic who believes that he is not justified in either asserting that God exists or denying that he exists and that what he must reasonably do is suspend belief. On the contrary, such an atheist believes that he has very good grounds indeed, as things stand, for denying the existence of God. But he will, on the second conceptualization of what it is to be an atheist, not deny that things could be otherwise and that, if they were, he would be justified in believing in God or at least would no longer be justified in asserting that it is false that there is a God. Using reliable empirical techniques, proven methods for establishing matters of fact, the fallibilistic atheist has found nothing in the universe to make a belief that God exists justifiable or even, everything considered, the most rational option of the various options. He therefore draws the atheistical conclusion (also keeping in mind his burden-of-proof argument) that God does not exist. But he does not dogmatically in a priori fashion deny the existence of God. He remains a thorough and consistent fallibilist.

ATHEISM AND METAPHYSICAL BELIEFS

Such a form of atheism (the atheism of those pragmatists who are also naturalistic humanists), though less inadequate than the first formation of atheism, is still inadequate. God in developed forms of Judaism, Christianity, and Islām is not, like Zeus or Wotan, construed in a relatively plain anthropomorphic way. Nothing that could count as "God" in such religions could possibly be observed, literally encountered, or detected in the universe. God, in such a conception, is utterly transcendent to the world; he is conceived of as "pure spirit," an infinite individual who created the universe out of nothing and who is distinct from the universe. Such a reality—a reality that is taken to be an ultimate mystery—could not be identified as objects or processes in the universe can be identified. There can be no pointing at or to God, no ostensive teaching of "God," to show what is meant. The word God can only be taught intralinguistically. "God" is taught to someone who does not understand what the word means by the use of descriptions such as "the maker of the universe," "the eternal, utterly independent being upon whom all other beings depend," "the first cause," "the sole ultimate reality," or "a self-caused being." For someone who does not understand such descriptions, there can be no understanding of the concept of God. But the key terms of such descriptions are themselves no more capable of ostensive definition (of having their referents pointed out) than is "God," where that term is not, like "Zeus," construed anthropomorphically. (That does not mean that anyone has actually pointed to Zeus or observed Zeus but that one knows what it would be like to do so.)

God as "pure spirit"

In coming to understand what is meant by "God" in such discourses, it must be understood that God, whatever else he is, is a being that could not possibly be seen or be in any way else observed. He could not be anything material or empirical, and he is said by believers to be an intractable mystery. A nonmysterious God would not be the God of Judaism, Christianity, and Islām.

This, in effect, makes it a mistake to claim that the existence of God can rightly be treated as a hypothesis and makes it a mistake to claim that, by the use of the experimental method or some other determinate empirical method, the existence of God can be confirmed or disconfirmed as can the existence of an empirical reality. The retort made by some atheists, who also like pragmatists remain thoroughgoing fallibilists, is that such a proposed way of coming to know, or failing to come to know, God makes no sense for anyone who understands what kind of reality God is supposed to be. Anything whose existence could be so verified would not be the God of Judeo-Christianity. God could not be a reality whose presence is even faintly adumbrated in experience, for anything that could even count as the God of Judeo-Christianity must be transcendent to the world. Anything that could actually be encountered or experienced could not be God.

At the very heart of a religion such as Christianity there stands a metaphysical belief in a reality that is alleged to transcend the empirical world. It is the metaphysical belief that there is an eternal, ever-present creative source and sustainer of the universe. The problem is how it is possible to know or reasonably believe that such a reality exists or even to understand what such talk is about.

God and theoretical entities

It is not that God is like a theoretical entity in physics such as a proton or a neutrino. They are, where they are construed as realities rather than as heuristically useful conceptual fictions, thought to be part of the actual furniture of the universe. They are not said to be transcendent to the universe, but rather are invisible entities in the universe logically on a par with specks of dust and grains of sand, only much, much smaller. They are on the same continuum; they are not a different kind of reality. It is only the case that they, as a matter of fact, cannot be seen. Indeed no one has an understanding of what it would be like to see a proton or a neutrino—in that way they are like God—and no provision is made in physical theory for seeing them. Still, there is no logical ban on seeing them as there is on seeing God. They are among the things in the universe, and thus, though they are invisible, they can be postulated as causes of things that are seen. Since this is so it becomes at least logically possible indirectly to verify by empirical methods the existence of such realities. It is also the case that there is no logical ban on establishing what is necessary to establish a causal connection, namely a constant conjunction of two discrete empirical realities. But no such constant conjunction can

be established or even intelligibly asserted between God and the universe, and thus the existence of God is not even indirectly verifiable. God is not a discrete empirical thing or being, and the universe is not a gigantic thing or process over and above the things and processes in the universe of which it makes sense to say that the universe has or had a cause. But then there is no way, directly or indirectly, that even the probability that there is a God could be empirically established.

ATHEISM AND INTUITIVE KNOWLEDGE

The gnostic may reply that there is a nonempirical way of establishing or making it probable that God exists. The claim is that there are truths about the nature of the cosmos neither capable of verification nor standing in need of verification. There is, gnostics claim against empiricists, knowledge of the world that transcends experience and comprehends the sorry scheme of things entire.

Influence of Hume and Kant

Since the thorough probings of such epistemological foundations by David Hume and Immanuel Kant, scepticism about how, and indeed even that, such knowledge is possible is very strong indeed. With respect to knowledge of God in particular, both Hume and Kant provide powerful critiques of the traditional attempts to prove the existence of God (notwithstanding the fact that Kant remained a Christian). While some of the details of their arguments have been rejected and refinements rooted in their argumentative procedure have been developed, there is a considerable consensus among philosophers and theologians that arguments of the general type as those developed by Hume and Kant show that no proof of God's existence is possible. Alternatively, to speak of "intuitive knowledge" (an intuitive grasp of being or of an intuition of the reality of the divine being) is to make an appeal to something that is not sufficiently clear to be of any value in establishing anything.

Revelation and religious authority

Prior to the rise of anthropology and the scientific study of religion, an appeal to revelation and authority as a substitute for knowledge or warranted belief might have been thought to have considerable force. But with a knowledge of other religions and their associated appeals to revealed truth, such arguments are without probative force. Claimed, or alleged, revelations are many, diverse, and not infrequently conflicting; without going in a small and vicious circle, it cannot be claimed, simply by appealing to a given putative revelation, that the revelation is the "true revelation" or the "genuine revelation" and that others are mistaken or, where nonconflicting, mere approximations to the truth. Similar things need to be said for religious authority. Moreover, it is at best problematic whether faith could sanction speaking of testing the genuineness of revelation or of the acceptability of religious authority. Indeed, if something is a "genuine revelation," there is no using reason to assess it. But the predicament is that plainly, as a matter of anthropological fact, there is a diverse and sometimes conflicting field of alleged revelations with no way of deciding or even having a reasonable hunch which, if any, of the candidate revelations is the genuine article. But even if the necessity for tests for the genuineness of revelation is allowed, there still is a claim that clearly will not do, for such a procedure would make an appeal to revelation and authority supererogatory. It is, where such tests are allowed, not revelation or authority that can warrant the most fundamental religious truths on which the rest depend. It is something else—that which establishes the genuineness of the revelation or authority—that guarantees these religious truths (if such there be), including the proposition that God exists. But the question returns, like the repressed, what that fundamental guarantee is or could be. Perhaps such a belief is nothing more than a cultural myth. There is, as has been shown, neither empirical nor a priori knowledge of God, and talk of intuitive knowledge is without logical force.

If these considerations are near to the mark, it is unclear what it means to say, as some agnostics and even atheists have, that they are sceptical God-seekers who simply have not found, after a careful examination, enough evidence to make belief in God a warranted or even a reasonable belief. It is unclear what it would be like to have, or for that matter fail to have, evidence for the existence of God. It is not that the God-seeker has to be able to give the evidence, for if that were so no search would be necessary, but that he, or at least somebody, must be able to conceive what would count as evidence if he had it so that he (and others) have some idea of what to look for. But it appears to be just that which cannot be done.

Perhaps there is room for the retort that it is enough for the God-seeker not to accept any logical ban on the possiblility of there being evidence. He need not understand what it would be like to have evidence in this domain. But, in turn, when one considers what kind of transcendent reality God is said to be, there seems to be an implicit logical ban on there being empirical evidence (a pleonasm) for his existence. It would seen plausible to assert that there is such a ban, though any such assertion should, of course, be made in a tentative way.

Someone trying to give empirical anchorage to talk of God might give the following hypothetical case. (It is, however, important in considering the case to keep in mind that things even remotely like what is described do not happen.) If thousands of people were standing out under the starry skies and all saw—the thing went on before their very eyes—a set of stars rearrange themselves to spell out "God," they would indeed rightly be utterly astonished and think that they had gone mad. Even if they could somehow assure themselves that this was not in some way a form of mass hallucination—how they could do this is not evident—such an experience would not constitute evidence for the existence of God, for they still would be without a clue as to what could be meant by speaking of an infinite individual transcendent transcendent to the world. Such an observation (the stars so rearranging themselves), no matter how well confirmed, would not ostensively fix the reference range of "God." Talk of such an infinite individual is utterly incomprehensible and has every appearance of being incoherent. No one knows what he is talking about in speaking of such a transcendent reality. All they would know is that something very strange indeed had happened. The doubt arises whether believers, or indeed anyone else in terms acceptable to believers, can give an intelligible account of the concept of God or of what belief in God comes to once God is de-anthropomorphized.

COMPREHENSIVE DEFINITION OF ATHEISM

Reflection on this should lead to a more adequate statement of what atheism is and indeed as well to what an agnostic or religious response to atheism should be. Instead of saying that an atheist is someone who believes that it is false or probably false that there is a God, a more adequate characterization of atheism consists in the more complex claim that to be an atheist is to be someone who rejects belief in God for the following reasons (which reason is stressed depends on how God is being conceived):

Reasons for rejecting belief in God

for an anthropomorphic God, the atheist rejects belief in God because it is false or probably false that there is a God; for a nonanthropomorphic God (the God of Luther and Calvin, Aquinas, and Maimonides), he rejects belief in God because the concept of such a God is either meaningless, unintelligible, contradictory, incomprehensible, or incoherent; for the God portrayed by some modern or contemporary theologians or philosophers, he rejects belief in God because the concept of God in question is such that it merely masks an atheistic substance—*e.g.*, "God" is just another name for love, or "God" is simply a symbolic term for moral ideals.

This atheism is a much more complex notion, as are its various reflective rejections. It is clear from what has been said about the concept of God in developed forms of Judeo-Christianity that the more crucial form of atheist rejection is not the assertion that it is false that there is a God but instead the rejection of belief in God because the concept of God is said not to make sense—to be in some important way incoherent or unintelligible.

Such a broader conception of atheism, of course, includes everyone who is an atheist in the narrower sense, but the converse does not obtain. Moreover, this conception of atheism does not have to say that religious claims are

meaningless. The more typical and less paradoxical and tendentious claim is that utterances such as "There is an infinite, eternal creator of the universe" are incoherent and that the conception of God reflected in such a claim is unintelligible, and in that important sense the claim is inconceivable and incredible—incapable of being a rational object of belief for a philosophically and scientifically sophisticated person touched by modernity. It is this that is a central belief of many contemporary atheists. There are good empirical grounds for believing that there are no Zeus-like spiritual beings, and as this last, more ramified form of atheism avers, if there are sound grounds for believing that the nonanthropomorphic or at least radically less anthropomorphic conceptions of God are incoherent or unintelligible, the atheist has the strongest grounds for rejecting belief in God.

Atheism

Atheism is a critique and a denial of the central metaphysical beliefs of systems of salvation involving a belief in God or spiritual beings, but a sophisticated atheist does not simply claim that all such cosmological claims are false but takes it that some are so problematic that, while purporting to be factual, they actually do not succeed in making a coherent factual claim. The claims, in an important sense, do not make sense, and, while believers are under the illusion that there is something intelligible to be believed in, in reality there is not. These seemingly grand cosmological claims are in reality best understood as myths or ideological claims reflecting a confused understanding of their utterers' situation.

It is not a well-taken rejoinder to atheistic critiques to say, as have some contemporary Protestant theologians, that belief in God is the worst form of atheism and idolatry, since the language of Jewish and Christian belief, including such sentences as "God exists" and "God created the world," is not to be taken literally but symbolically and metaphorically. Christianity, as Reinhold Niebuhr, a theologian who defends such views, once put it, is "true myth." The claims of religion are not, on such account, to be understood as metaphysical claims trying to convey extraordinary facts but as metaphorical and analogical claims that are not understandable in any other terms. But if something is a metaphor it must at least in principle be possible to say what it is a metaphor of. Thus metaphors cannot be understandable only in metaphorical terms. There can be no unparaphrasable metaphors or symbolic expressions though, what is something else again, a user of such expressions may not be capable on demand of supplying that paraphrase. Moreover, if the language of religion becomes simply the language of myth and religious beliefs are viewed simply as powerful and often humanly compelling myths, then they are conceptions that in reality have only an atheistic substance. The believer is making no cosmological claim that the atheist is not; it is just that his talk, including his unelucidated talk of "true myths," is language that for many people has a more powerful emotive force.

Agnosticism

Agnosticism has a parallel development to that of atheism. An agnostic, like an atheist, asserts either that he does not know that God exists—or, more typically, that he cannot know or have sound reasons for believing that God exists—but unlike the atheist he does not think that he is justified in saying that God does not exist or, stronger still, that God cannot exist. Similarly, while some contemporary atheists say that the concept of God in developed theism does not make sense and thus that Jewish, Christian, and Islāmic beliefs must be rejected, many contemporary agnostics believe that the concept of God is radically problematic. They maintain that they are not in a position to be able to decide whether, on the one hand, the terms and concepts of such religions are so problematic that such religious beliefs do not make sense or whether, on the other, though the talk is indeed radically paradoxical and in many ways incomprehensible, such talk has sufficient coherence to make reasonable a belief in an ultimate mystery. Such an agnostic recognizes that the puzzles about God cut deeper than perplexities concerning whether it is possible to attain adequate evidence for God's existence. Rather, he sees the need to exhibit an adequate nonanthropomorphic, extralinguistic referent for "God." (This need not commit him to the belief that there are any observations independent of theory.) Believers think that, though God is a mystery, such a referent has been secured, though what it is remains a mystery. Atheists, by contrast, believe that it has not been, and indeed some of them believe that it cannot be, secured. To talk about mystery, they maintain, is just an evasive way of talking about what is not understood. Contemporary agnostics (those agnostics who parallel the atheists characterized above) remain in doubt and are convinced that there is no rational way of resolving the doubt about whether talk in a halting fashion of God just barely secures such reference or whether it, after all, fails and that nothing religiously acceptable is referred to by "God."

Intense religious commitment, as the history of fideism makes evident, has sometimes gone hand in hand with deep scepticism concerning man's capacity to know God. It is agreed by all parties to the dispute between belief and unbelief that religious claims are paradoxical. Furthermore, criteria for what is meaningless and what is not or for what is intelligible and what is not are deeply contested. It is perhaps fair enough to say that there are no generally accepted criteria.

Keeping these diverse considerations in mind in the arguments between belief, agnosticism, and atheism, it is crucial to ask whether there is any good reason at all to believe that there is a personal creative reality that is beyond the bounds of space and time and transcendent to the world. Is there even a sufficient understanding of such talk so that such a reality can be the object of religious commitment? (One cannot have faith in or take on faith what one does not at all understand. People must at least in some way understand what it is that they are to have faith in to be able to have faith in it. If a person is asked to trust Irglig, he cannot do so no matter how strongly he wants to take something simply on trust.)

The nature of faith

It appears to be a brute fact that there just is that indefinitely immense collection of finite and contingent masses or conglomerations of things and processes the phrase "the universe" refers to. People can come to feel wonder, awe, and puzzlement that there is a universe at all. But that fact, or the very fact that there is a world at all, does not license the claim that there is a noncontingent reality on which the world (the sorry collection of things entire) depends. It is not even clear that such a sense of contingency gives an understanding of what such a noncontingent thing could be. Some atheists think that the reference range of "God" is so indeterminate and the concept of God so problematic that it is impossible for someone fully aware of that reasonably to believe in God; believers, by contrast, think that, though the reference range of "God" is indeterminate, it is not so indeterminate and the concept of God so problematic as to make belief irrational or incoherent. It is known, they claim, that talk of God is problematic, but it is not known, and cannot be known, whether it is so problematic as to be without a religiously appropriate sense. Agnostics, in turn, say that there is no reasonable decision procedure. It is not known and cannot be ascertained whether or not "God" secures a religiously adequate referent. What needs to be kept in mind, in reflecting on this issue, is whether a "contingent thing" is a pleonasm and "infinite reality" is without sense and whether, when people go beyond anthropomorphism (or try to go beyond it), it is possible to have a sufficient understanding of what is referred to by "God" to make faith a coherent possibility.

Problems of the term "God"

Finally, it will not do to take a Pascalian or Dostoyevskian turn and claim that, intellectual absurdity or not, religious belief is necessary, since without belief in God morality does not make sense and life is meaningless. That claim is false, for even if there is no purpose to life there are purposes in life—things people care about and want to do—that can remain perfectly intact even in a godless world. God or no God, immortality or no immortality, it is vile to torture people just for the fun of it, and friendship, solidarity, love, and the attainment of self-respect are human goods even in an utterly godless world. There are intellectual puzzles about how people know that these things are good, but that is doubly true for the distinctive

claims of a religious ethic. The point is that these things remain desirable and that life can have a point even in the absence of God. (K.E.N.)

BIBLIOGRAPHY

Nature worship: Apart from some older works about particular forms of nature worship, the most important discussions of this form of religion and the related nature mythology appear in journals. JAMES G. FRAZER, *The Worship of Nature* (1926), is a classic work mainly concerned with the worship of the heavens, the sun, and the earth; and *The Golden Bough*, 3rd ed. rev., 12 vol. (1907–15; abridged ed., 1922 and 1959), contains an important discussion of nature gods, especially vegetation gods, and their cults. Important references to the worship of nature may be found under various headings in *The Mythology of All Races*, ed. by LOUIS H. GRAY, 13 vol. (1916–32); and *The Encyclopaedia of Religion and Ethics*, ed. by JAMES HASTINGS, 13 vol. (1908–26; reprinted 1955).

Animism: The original and classic source is E.B. TYLOR, *Primitive Culture*, 2 vol. (1871 and later editions, reprinted 1958). For the "pre-animism" thesis, see R.R. MARETT, *The Threshold of Religion*, 4th ed. (1929). An extended analysis and review of theories may be found in ADOLF E. JENSEN, *Mythos und Kult bei Naturvölkern*, 2nd ed. (1960; Eng. trans., *Myth and Cult Among Primitive Peoples*, 1963); or for a short review, see E.E. EVANS-PRITCHARD, *Theories of Primitive Religion* (1965). Still to be recommended for an overview is ROBERT H. LOWIE, *Primitive Religion*, new ed. (1970). An ample source for the animism of China is the 5th volume of J.J.M. DE GROOT's monumental study, *The Religious System of China* (1907, reprinted 1969); and for Burma the best introduction is M.E. SPIRO, *Burmese Supernaturalism* (1967). Ethnographic studies of animistic peoples abound; for an annotated bibliography, see WILLIAM A. LESSA and EVON Z. VOGT, *Reader in Comparative Religion*, 2nd ed., pp. 640–645 (1965). A provocative consideration of the relation of animism to modern world views is presented in JACQUES MONOD, *Le Hasard et la nécessité* (1970; Eng. trans., *Chance and Necessity*, 1971).

Totemism: FRANZ BOAS, "The Origin of Totemism," *Am. Anthrop.*, 18:319–326 (1916), contains a variety of ideas and objects defined as totemism, and thus maintains little unity; A.P. ELKIN, "Studies in Australian Totemism: The Nature of Australian Totemism," *Oceania*, 4:113–131 (1933–34); E.E. EVANS-PRITCHARD, "Nuer Totemism," *Annali Lateran.*, 13:225–248 (1949), describes the manifestations of totemism combined with the conception of the spirit in the "Nilot" tribe; J.V. FERREIRA, *Totemism in India* (1965), a critical evaluation of hitherto existing works on totemism in general and that of India in particular; R. FIRTH, "Totemism in Polynesia," *Oceania*, 1:291–321, 377–398 vol. 1 (1930–31), a useful survey article; J.L. FISCHER, "Totemism on Truk and Ponape," *Am. Anthrop.*, 59:250–265 (1957), describes and interprets the highly contrasting forms of totemism found in Micronesia, using not only the unusual sociological–ideological organization but also containing psychological aspects. G. FOSTER, "Nagualism in Mexico and Guatemala," *Acta Am.*, 2:85–103 (1944), deals with important borderline cases of totemism with particular regard to the problem of personal totemism; J.G. FRAZER, *Totemism and Exogamy*, 4 vol. (1910) and *Totemica: A Supplement to Totemism and Exogamy* (1937), comprehensive and informative reference works although the hypotheses are now out of date; A.A. GOLDENWEISER, "Totemism: An Analytic Study," *J. Am. Folklore*, 23:179–293 (1910); "Totemism," *Encyclopaedia of the Social Sciences*, vol. 14 (1934); J. HAEKEL, "Der heutige Stand des Totemismusproblems," *Mitt. Anthrop. Ges. Wien*, 82:33–49 (1952), attempts to give a critical examination with concrete examples of the complex question in its various forms; CLAUDE LÉVI-STRAUSS, *Le Totémisme aujourd'hui* (1962; Eng. trans., *Totemism*, 1963), contains a detailed critical evaluation of existing hypotheses of Anglo-American and French authors; R. PIDDINGTON, *An Introduction to Social Anthropology* 1:200–206 (1950), offers a short but sufficient characterization of the totemic phenomena, the difficulty in defining it, its great variability, and some concrete examples.

Ancestor worship: J.T. ADDISON, *Chinese Ancestor Worship* (1925), descriptive information on traditional practices in China; E. BENDANN, *Death Customs* (1930), reprinted 1969), descriptive material useful in distinguishing between death cults and ancestor worship; R.F. FORTUNE, *Manus Religion* (1935, reprinted 1965), a detailed account, including information on the role of ancestor worship as a sanction for ethical behaviour; J.G. FRAZER, *The Belief in Immortality and Worship of the Dead*, 3 vol. (1913–24, reprinted 1968), a classic work presenting extensive information on its subject among societies of the world; J.R. GOODY, *Death, Property and the Ancestors* (1962), a detailed account of mortuary customs of the Lodagaa of West Africa; B. MALINOWSKI, "Baloma, the Spirits of the Dead in the Trobriand Islands," in *Magic, Science and Religion and Other Essays* (1948).

Polytheism: S.G.F. BRANDON (ed.), *A Dictionary of Comparative Religion* (1970), contains articles on the various gods and also on theories, and these have bibliographical sections. A useful compendium is the *Larousse Encyclopedia of Mythology* (1959). One example of the numerous recent anthropological literature is G.E. SWANSON, *The Birth of the Gods: The Origin of Primitive Beliefs* (1960).

Pantheism and panentheism: CHARLES HARTSHORNE and W.L. REESE, *Philosophers Speak of God* (1953), offers an extensive historical exploration of pantheism, panentheism, and Classical Theism. The fundamental basis of panentheism is discussed not only in the epilogue of the above volume, but in many other works by CHARLES HARTSHORNE, including *The Divine Relativity* (1948) and *The Logic of Perfection* (1962). For the relation of mysticism to pantheism, see W.T. STACE, *Mysticism and Philosophy* (1960). For information concerning any of the philosophers mentioned, reference may be made to their individual entries in *The Encyclopedia of Philosophy*, 8 vol., ed. by PAUL EDWARDS (1967); the *Enciclopedia filosofica*, 6 vol., 2nd ed., ed. by G.C. SANSONI (1967); or the *Diccionario de filosofía*, 2 vol., ed. by JOSE FERRATER MORA (1965).

Religious dualism: UGO BIANCHI, *Il dualismo religioso: saggio storico ed etnologico* (1958), discusses the "dualistic area" extending from ancient Greece to Iran, east European folklore, north and Central Asia, and North America; see also his "Le dualisme en histoire des religions," *Revue de l'histoire des religions*, 159:1–46 (1961); and *Le origini dello gnosticismo* (1967), a collection of papers (in French, German, English, and Italian) presented at the Colloquium of Messina, April 1966, many of which are devoted to dualism in various religions. MIRCEA ELIADE, "Prolegomenon to Religious Dualism: Dyads and Polarities," *The Quest: History and Meaning in Religion*, ch. 8 (1969), is concerned not only with dualism proper but also with the functions of "duality"; his *De Zalmoxis à Gengis-Khan*, ch. 2–3 (1970), is a study of dualism in folklore and ethnology. SIMONE PETREMENT, *Le Dualisme dans l'histoire de la philosophie et des religions* (1946), and *Le Dualisme chez Platon, les Gnostiques et les Manichéens* (1947), are two important general surveys but do not sufficiently distinguish the different meanings of dualism in the philosophical and the religious-historical terminologies. For an analytic exposition of the Gnostic ideology, with modern analogies, see HANS JONAS, *The Gnostic Religion*, 2nd ed. (1963); for a discussion of the varieties of Iranian dualism, R.C. ZAEHNER, *Zurvan: A Zoroastrian Dilemma* (1955), JACQUES DUCHESNE-GUILLEMIN, *The Western Response to Zoroaster* (1958), deals with the history of the problem of Iranian dualism. Other works include: GEO WIDENGREN, "Der iranische Hintergrund der Gnosis," in *Zeitschrift für Religions- und Geistesgeschichte*, 4:97–114 (1952), on dualism in the Indian *Upanisads*; F.K. NUMAZAWA, *Die Weltanfänge in der japanischen Mythologie . . .* (1946), a comparative, ethnological appreciation of the Yin–Yang opposition; MARCEL GRIAULE and GERMAINE DIETERLEN, *Le Renard Pâle*, vol. 1, fasc. 1 (1965), an account of dualism in the ontology and the mythology of the Dogon of West Sudan; HELMER RINGGREN, "Dualism," *The Faith of Qumran*, ch. 2 (1963), a study of Qumränic dualism; and R.M. GRANT, *Gnosticism and Early Christianity*, rev. ed. (1966).

Monotheism: General works include: JOSIAH ROYCE, "Monotheism," in the *Encyclopaedia of Religion and Ethics*, vol. 8 (1928); "Monotheismus und Polytheismus," in *Religion in Geschichte und Gegenwart*, vol. 4 (1913); T.P. VAN BAAREN, *Doolhof der goden* (1960); G. VAN DER LEEUW, *Phänomenologie der Religion* (1933; Eng. trans., *Religion in Essence and Manifestation*, 2 vol., 1963); and R. OTTO, *Das Heilige* (1917; Eng. trans., *The Idea of the Holy*, 2nd ed., 1950). For primitive religions, see T.P. VAN BAAREN, *Menschen wie wir* (1964); and P. RADIN, *Monotheism Among Primitive Peoples* (1954); on the high gods, see H. ZWICKER, *Das Höchste Wesen* (1970); and E.O. JAMES, *The Concept of Deity* (1950), the only general systematic study of recent date treating the problems of monotheism. H.R. NIEBUHR, *Radical Monotheism and Western Culture* (1960), is a modern, incisive Protestant theological presentation of absolute monotheism.

Theism: The classic recent statement of God's transcendence is A.M. FARRER, *Finite and Infinite*, 2nd ed. (1959), a difficult but essential book; C.A. CAMPBELL, *On Selfhood and Godhood* (1957), is an exceptionally lucid presentation that allows for the distinctness of finite beings; see also further statements in WILLIAM TEMPLE, *Nature, Man and God* (1934); H.H. FARMER, *God and Men* (1947); and H.D. LEWIS, *Philosophy of Religion* (1965). A. SETH PRINGLE-PATTISON presents the more traditional Idealist view in *The Idea of God in the Light of Recent Philosophy* (1920). An Idealism stressing the immediate awareness of other minds and of God is found in W.E. HOCKING, *The Meaning of God in Human Experience* (1912); a presentation similarly starting from Empiricism and science that culminates in a "Cosmic Teleology" is that of F.R. TENNANT, *Philosophical*

Theology, 2 vol. (1928–30). E.S. BRIGHTMAN, *The Problem of God* (1930), treats God as a limited being (finitism).

Deism: JOHN LELAND, *A View of the Principal Deistical Writers . . .*, 3rd ed., 3 vol. (1754; also 1837 ed.), the first historical account; FRITZ MAUTHNER, *Der Atheismus und seine Geschichte im Abendlande*, 4 vol. (1921–23), a complete history; ERNST CASSIRER, *Die Philosophie der Aufklärung* (1932; Eng. trans., *The Philosophy of the Enlightenment*, 1951), a description of Deism and its philosophical background; HAROLD G. NICOLSON, *The Age of Reason* (1960), on the nature of 18th-century Rationalism and its connection with Deism; JAMES COLLINS, *God in Modern Philosophy* (1959), a full history of Deism, here called "theism," from Nicolas of Cusa to contemporary theological theories; JOHN ORR, *English Deism: Its Roots and Its Fruits* (1934); GOTTHARD V. LECHLER, *Geschichte des englischen Deismus* (1841), the first full history after the end of Deism; HERBERT OF CHERBURY, *De Veritate* (1624; Eng. trans. by MEYRICK H. CARRE, *On Truth*, 1937), the first English translation of the reputedly "first" classic expression of Deism; MARIO M. ROSSI, *La vita, le opere, i tempi di Edoardo Herbert di Chirbury*, 3 vol. (1947), and *Alle fonti del deismo e del materialismo moderno* (1942), two works that describe Herbert's life and Deistic thought against the background of the history of Deism and the attitude of the church. DAVID HUME, *Dialogues Concerning Natural Religion*, 2nd ed. with suppl. (1947), the beginning of the Deist's self-criticism; THOMAS PAINE, *The Age of Reason*, 3 pt. (1794–1811), the work most influential on the Deism of common people; JOHN S. SPINK, *French Free-Thought from Gassendi to Voltaire* (1960), on French Deism; HENRY E. ALLISON, *Lessing and the Enlightenment* (1966); IMMANUEL KANT, *Die Religion innerhalb der Grenzen der blossen Vernunft* (1793; Eng. trans., *Religion Within the Limits of Reason Alone*, 1947), the classic work of the last stage of German Deism; G.W.F. HEGEL, *Early Theological Writings*, trans. by THOMAS M. KNOX and RICHARD KRONER (1948), early writings to show Hegel's indebtedness to Deistic polemics.

Agnosticism: The fundamental primary sources are T.H. HUXLEY, "Agnosticism" and "Agnosticism and Christianity" in his *Collected Essays*, vol. 5 (1894). Other basic sources are W.K. CLIFFORD, *The Ethics of Belief, and Other Essays* (1876, reprinted 1947); and LESLIE STEPHEN, *An Agnostic's Apology, and Other Essays* (1893), which first appeared as an essay in 1876. An important classical antecedent is DAVID HUME, *An Enquiry Concerning Human Understanding* (1748), of which the best modern edition is that of C.W. HENDEL (1955). For a study of this, which was Hume's first *Enquiry*, see ANTONY FLEW, *Hume's Philosophy of Belief* (1961), which devotes special attention to its implications for religion. Religious agnosticism is treated in HENRY MANSEL, *The Limits of Religious Thought Examined in Eight Lectures* (1858). Notable secondary sources include LESLIE STEPHEN, *History of English Thought in the Eighteenth Century*, 2 vol. (1876), the best analytical study done at the time of Huxley and one that still deserves this high rating; R.F. FLINT, *Agnosticism* (1903); and R.A. ARMSTRONG, *Agnosticism and Theism in the Nineteenth Century* (1905); for a recent sympathetic biography, see CYRIL BIBBY, *T.H. Huxley: Scientist, Humanist, and Educator* (1959). Classic rejections of agnosticism include BLAISE PASCAL, *Pensées* (1960), posthumous fragments given their definitive arrangement by L. LAFUMA; and WILLIAM JAMES, "The Will to Believe," in *The Will to Believe, and Other Essays in Popular Philosophy* (1897).

Atheism: The article "Atheism" in *The Encyclopaedia of Philosophy*, vol. 1, pp. 174–189 (1967, reissued 1972), and the article "Agnoticism" in the *Dictionary of the History of Ideas*, vol. 1, pp. 17–27 (1968), give sophisticated conceptual elucidations of the concept of atheism. See also *Encyclopaedia of Religion and Ethics*, vol. 2, pp. 173–190 (1922); *Reallexikon für Antike und Christentum*, vol. 1, cols. 866–870 (1950); *Die Religion in Geschichte und Gegenwart*, 3rd ed., vol. 1, cols. 670–678 (1957); GEORGE KLAUS and MANFRED BUHR (eds.), *Philosophisches Wörterbuch*, 6th rev. and enl. ed., vol. 1, pp. 125–129 (1969); *Enciclopedia filosofica*, 2nd ed., vol. 1, col. 557–562 (1968); and KAI NIELSEN, *Introduction to the Philosophy of Religion* (1982). A classic extended history of atheism is FRITZ MAUTHNER, *Der Atheismus und seine Geschichte im Abendland*, 4 vol. (1920–23, reissued 1963). See also JACOB PRESSER, *Das Buch "De tribus impostoribus"* (1926); JOHN MACKINNON ROBERTSON, *A Short History of Freethought, Ancient and Modern* (1957, reissued 1972); HENRI BUSSON, *Le Rationalisme dans la littérature française de la Renaissance (1533–1601)*, new ed., rev. and augmented (1971); and CORNELIO FABRO, *God in Exile: Modern Atheism* (1968; orginally published in Italian, 1961). Modern atheism is treated in CHARLES BRADLAUGH, *Champion of Liberty: Charles Bradlaugh* (1933); BARON D'HOLBACH, *The System of Nature* (1756–96, reissued 1970; originally published in French, 1770), and *Common Sense* (1795, reissued 1972; originally published in French, 1772); ARTHUR SCHOPENHAUER, *Complete Essays of Schopenhauer*, trans. by T. BAILEY SAUNDERS (1896); LUDWIG FEUERBACH, *The Essence of Christianity* (1854, reissued 1972, trans. of 2nd German ed.; originally published in German, 1841); FRIEDRICH NIETZSCHE, *Thus Spake Zarathustra* (1896; originally issued in German, 1883–92); KARL MARX and FRIEDRICH ENGELS, *On Religion* (1955); THOMAS H. HUXLEY, *Collected Essays*, vol. 5 (1894); and LESLIE STEPHEN, *An Agnostic's Apology, and Other Essays* (1893), and *History of English Thought in the Eighteenth Century*, 3rd ed., 2 vol., (1902, reissued 1963). Powerful contemporary defenses of atheism, together with some religious responses, are given in NORBERT O. SCHEDLER (ed.), *Philosophy of Religion* (1794).

Contemporary analytical discussions of atheism include MICHAEL SCRIVEN, *Primary Philosophy* (1966); RICHARD ROBINSON, *An Atheist's Values* (1964, reissued 1975); BERTRAND RUSSELL, *Why I Am Not a Christian, and Other Essays on Religion and Related Subjects* (1957); KAI NIELSEN, *Skepticism* (1973), and *Reason and Practice* (1971); SIDNEY HOOK, *The Quest for Being, and Other Studies in Naturalism and Humanism* (1961, reprinted 1971); and RONALD W. HEPBURN, *Christianity and Paradox* (1958, reissued 1968). Two anthologies that give the core debate between belief and unbelief are MALCOLM L. DIAMOND and THOMAS V. LITZENBURG, JR. (eds.), *The Logic of God* (1975), and ANTONY FLEW and ALASDAIR MACINTYRE and PAUL ROCOEUR, *The Religious Significance of Atheism* (1969); and MOSTAFA FAGHFOURY (ed.), *Analytical Philosophy of Religion in Canada* (1982).

Religious Experience

Religious experience is taken here to include such specific experiences as wonder at the infinity of the cosmos, the sense of awe and mystery in the presence of the holy, feelings of dependence on a divine power or an unseen order, the sense of guilt and anxiety accompanying belief in a divine judgment, and the feeling of peace that follows faith in divine forgiveness. Some thinkers also point to a religious aspect to the purpose of life and with the destiny of the individual. In the first sense, religious experience means an encounter with the divine in a way analogous to encounters with other persons and things in the world. In the second case, reference is made not to an encounter with a divine being but rather to the apprehension of a quality of holiness or rightness in reality or to the fact that all experience can be viewed in relation to the ground from which it springs. In short, religious experience means both special experience of the divine or ultimate and the viewing of any experience as pointing to the divine or ultimate.

The first part of this article provides an overview of religious experience from a philosophical/psychological point of view. In the second part, the particular category of religious experience known variously as mysticism, enlightenment, or illumination is examined at some length, from a cross-cultural historical perspective. (See also RELIGIOUS AND SPIRITUAL BELIEF, SYSTEMS OF; RITES AND CEREMONIES, SACRED; SACRED OFFICES AND ORDERS.)

This article is divided into the following sections:

The nature and forms of religious experience 578
- Study and evaluation 578
- Religious experience and other experience 579
  - Views of experience in general
  - Views of religious experience
- The structure of religious experience 580
  - The self and the other
  - Social forms or expressions
  - Objective "intention," or reference
  - Immediacy and mediation
- Situational contexts and forms of expression 581
- Types of religious experience and personality 582

The experience of mysticism 582
- Nature and significance 582
- Relation of mystical experience to other kinds of experience 583
- Definitions of mysticism and mystical experience 583
  - Differences between mysticism and similar phenomena
  - Basic patterns
  - Introvertive mysticism
  - Other definitions and experiences of mysticism
- Universal types of mystical experience 584
  - Intellectual and contemplative forms
  - Devotional forms
  - Ecstatic and erotic forms
- Goal of mystical experience and mysticism 585
  - Experience of the divine or sacred
  - Union with the divine or sacred
  - Experience of the universal
  - Experience of oneness with people
- Mystical relationship between man and the sacred 585
  - Nature of the relationship
  - Means and modes of the relationship
- Semantics and symbolism in mystical experience 587
  - Union of opposites
  - Emptiness and fullness
  - Symbolism of divine messengers
  - Symbolism of love and marriage
  - Symbolism of the journey
- Psychological aspects of mysticism 588
  - Awareness
  - Role of identification
- Systematic exposition of mystical experience 589
  - Attempts of mystics to record the nature of their experiences
  - The value and meaning of mystical experience
  - Problems of communication and understanding
- Mysticism as a social factor 589

Bibliography 590

The nature and forms of religious experience

STUDY AND EVALUATION

"Religious experience" was not widely used as a technical term prior to the publication of *The Varieties of Religious Experience* (1902) by William James, an eminent U.S. psychologist and philosopher, but the interpretation of religious concepts and doctrines in terms of individual experience reaches back at least to 16th-century Spanish mystics and to the age of the Protestant Reformers. A special emphasis on the importance of experience in religion is found in the works of such thinkers as Jonathan Edwards, Friedrich Schleiermacher, and Rudolf Otto. Basic to the experiential approach is the belief that it allows for a firsthand understanding of religion as an actual force in human life, in contrast with religion taken either as church membership or as belief in authoritative doctrines. The attempt to interpret such concepts as God, faith, conversion, sin, salvation, and worship through personal experience and its expressions opened up a wealth of material for the investigation of religion by psychologists, historians, anthropologists, and sociologists as well as by theologians and philosophers. A focus on religious experience is especially important for Phenomenologists (thinkers who seek the basic structures of human consciousness) and Existentialist philosophers.

A number of controversial issues have emerged from these studies, involving not only different conceptions of the nature and structure of religious experience but also different views of the manner in which it is to be evaluated and the sort of evaluation possible from the standpoint of a given discipline. Four such issues are basic: (1) whether religious experience points to special experiences of the divine or whether any experience may be regarded as religious by virtue of becoming related to the divine; (2) the kinds of differentia that can serve to distinguish religion or the religious from both secular life and other forms of spirituality, such as morality and art; (3) whether religious experience can be understood and properly evaluated in terms of its origins and its psychological or sociological conditions or is *sui generis*, calling for interpretation in its own terms; and (4) whether religious experience has cognitive status, involving encounter with a being, beings, or a power transcending human consciousness, or is merely subjective and composed entirely of ideas and feelings that have no reference beyond themselves. The last issue, transposed in accordance with either a Positivist outlook or some types of Empiricism, which restrict assertible reality to the realm of sense experience, would be resolved at once by the claim that the problem cannot be meaningfully discussed, since key terms, such as "God" and "power," are strictly meaningless.

Four basic issues

Proponents of mysticism, such as Rudolf Otto, Rufus Jones, and W.T. Stace, have maintained the validity of immediate experience of the divine; theologians such as Emil Brunner have stressed the self-authenticating char-

acter of man's encounter with God; naturalistically oriented psychologists, such as Freud and J.H. Leuba, have rejected such claims, explaining religion in psychological and genetic terms as a projection of human wishes and desires. Philosophers such as William James, Josiah Royce, William E. Hocking, and Wilbur M. Urban have represented an idealist tradition in interpreting religion, stressing the concepts of purpose, value, and meaning as essential for understanding the nature of God. Naturalist philosophers, of whom John Dewey was typical, have focussed on the "religious" as a quality of experience and an attitude toward life that is more expressive of the human spirit than of any supernatural reality. Theologians Douglas Clyde Macintosh and Henry N. Wieman sought to build an "empirical theology" on the basis of religious experience understood as involving a direct perception of God. Unlike Macintosh, Wieman held that such a perception is sensory in character. Personalist philosophers, such as Edgar S. Brightman and Peter Bertocci, have regarded the person as the basic category for understanding all experience and have interpreted religious experience as the medium through which God is apprehended as the cosmic person. Existential thinkers, such as Søren Kierkegaard, Gabriel Marcel, and Paul Tillich, have seen God manifested in experience in the form of a power that overcomes estrangement and enables man to fulfill himself as an integrated personality. Process philosophers, such as Alfred North Whitehead and Charles Hartshorne, have held that the idea of God emerges in religious experience but that the nature and reality of God are problems calling for logical argument and metaphysical interpretation, in which emphasis falls on the relation between God and the world being realized in a temporal process. Logical Empiricists, of whom A.J. Ayer has been typical, have held that religious and theological expressions are without literal significance, because there is no way in which they can be either justified or falsified (refuted). On this view, religious experience is entirely emotive, lacking all cognitive value. Analytic philosophers following the lead of Ludwig Wittgenstein, an Austrian-British thinker, approach religious experience through the structure of religious language, attempting to discover exactly how this language functions within the community of believers who use it.

RELIGIOUS EXPERIENCE AND OTHER EXPERIENCE

Sense experience and beyond

Views of experience in general. Religious experience must be understood against the background of a general theory of experience as such. Experience as conceived from the standpoint of a British philosophical tradition stemming from John Locke and David Hume is essentially the reports of the world received through the senses. Experience, as a tissue of sensible content, was set in contrast to reason, understood as the domain of logic and mathematics. The mind was envisaged as a wax tablet on which the sensible world imprints itself; and the one who experiences is the passive recipient of what is given. It is possible to distinguish and compare these sensible items by means of understanding, but the data themselves are available only through experience—*i.e.,* the sensation of things and reflection upon thought and mental activities, feelings, and desires. According to this classical empiricist view, all ideas, beliefs, and theories expressed in conceptual form are to be traced back to their origin in sense if they are to be understood and justified.

The above view of experience came under criticism from two sides. Immanuel Kant, an 18th-century German philosopher, who still retained some of the assumptions of the position he criticized, nevertheless declared that experience is not identical with passively received sensible material but must be construed as the joint product of such material and its being grasped by an understanding that thinks in accordance with certain necessary categories not derived from the senses. Kant opened the way for a new understanding of the element of interpretation in all experience, and his successors in the development of German Idealism, Johann Fichte, Friedrich Schelling, and G.W.F. Hegel, came to characterize experience as the many-sided reflection of man's multiple encounters with the world, other men, and himself.

A second attack on the classical conception came from U.S. Pragmatist philosophers, notably Charles Sanders Peirce, William James, and John Dewey, for whom experience was the medium for the disclosure of whatever there is to be encountered; it is far richer and more complex than a passive registry of sensible data. Experience was seen as a human activity related to the purposes and interests of the one who experiences, and it was understood as an interpreted product of multiple transactions between man and the environment. Moreover, stress was placed on the social and funded character of experience in place of the older conception of experience as a private content confined to the mind of an individual. On this view, experience is not confined to its content but includes modes or dimensions that represent frames of meaning—social, moral, aesthetic, political, religious—through which whatever is encountered can be interpreted. James went beyond his associates in developing the broadest theory of experience, known as radical empiricism, according to which the relations and connections between items of experience are given along with these items themselves.

Critics of the classical view of experience, while not concerned exclusively with religious experience, saw, nevertheless, that if experience is confined to the domain of the senses it is then difficult to understand what could be meant by religious experience if the divine is not regarded as one sensible object among others. This consideration prompted attempts to understand experience in broader terms. Cutting across all theories of experience is the basic fact that experience demands expression in language and symbolic forms. To know what has been experienced and how it is to be understood requires the ability to identify things, persons, and events through naming, describing, and interpreting, which involve appropriate concepts and language. No experience can be the subject of analysis while it is being had or undergone; communication and critical inquiry require that experiences be cast into symbolic form that arrests them for further scrutiny. The various uses of language—political, scientific, moral, religious, aesthetic, and others—represent so many purposes through which experience is described and interpreted.

The expression and interpretation of experience

Views of religious experience. Specifically religious experience has been variously identified in the following ways: the awareness of the holy, which evokes awe and reverence; the feeling of absolute dependence that reveals man's status as a creature; the sense of being at one with the divine; the perception of an unseen order or of a quality of permanent rightness in the cosmic scheme; the direct perception of God; the encounter with a reality "wholly other"; the sense of a transforming power as a presence. Sometimes, as in the striking case of the Old Testament prophets, the experience of God has been seen as a critical judgment on man and as the disclosure of his separation from the holy. Those who identify religion as a dimension or aspect of experience point to man's attitude toward an overarching ideal, to a total reaction to life, to an ultimate concern for the meaning of one's being, or to a quest for a power that integrates human personality. In all these cases, it is the fact that the attitudes and concerns in question are directed to an ultimate object beyond man that justifies their being called religious. All interpreters are agreed that religious experience involves what is final in value for man and concerns belief in what is ultimate in reality.

Because of their intimate relation to one another, the religious and the moral have often been confused. The problem has been intensified by many attempts—beginning with Kant's treatise on religion (1793)—to interpret religion as essentially morality or merely as an incentive for doing one's duty. Religion and morality are, however, usually taken to be distinguishable; religion concerns the being of a person, what he is and what he acknowledges as the worshipful reality, while morality concerns what the person does and the principles governing his relation to others. While it is generally acknowledged that religion must affect man's conduct in the world, some have maintained that there is no morality without religion, while others deny this claim on the ground that morality must remain autonomous and free of divine sanctions. Reli-

gious experience may be distinguished from the aesthetic aspect of experience in that the former involves commitment and devotion to the divine, while the latter is focussed on the appreciation and enjoyment of qualities, forms, and patterns in themselves, whether as natural objects or works of art. Anthropological studies have shown that primitive religions gave birth to many forms of art that, in the course of development, won independence as secular forms of expression. The problem of the relation between religion and art is posed in a particularly acute way when reference is made to religious art as a special form of the aesthetic. Since it is concerned with the holy and the purpose of human life as a whole, most scholars would hold that religious experience should be related in an intelligible way to all other experience and forms of experience. The task of tracing out these relationships belongs to theology and the philosophy of religion.

THE STRUCTURE OF RELIGIOUS EXPERIENCE

The three basic elements: subjective, objective, social

The self and the other. All religious experience can be described in terms of three basic elements: first, the personal concerns, attitudes, feelings, and ideas of the individual who has the experience; second, the religious object disclosed in the experience or the reality to which it is said to refer; third, the social forms that arise from the fact that the experience in question can be shared. Although the first two elements can be distinguished for purposes of analysis, they are not separated within the integral experience itself. Religious experience is always found in connection with a personal concern and quest for the real self, oriented toward the power that makes life holy or a ground and a goal of all existence. A wide variety of individual experiences are thus involved, among which are attitudes of seriousness and solemnity in the face of the mystery of human destiny; feelings of awe and of being unclean evoked by the encounter with the holy; the sense of a power or a person who both loves and judges man; the experience of being converted or of having the course of life directed toward the divine; the feeling of relief stemming from the sense of divine forgiveness; the sense that there is an unseen order or power upon which the value of all life depends; the sense of being at one with the divine and of abandoning the egocentric self.

In all these situations, the experience is realized in the life of an individual who at the same time has his attention focussed on an "other," or divine reality, that is present or encountered. The determination of the nature of this other poses a problem of interpretation that requires the use of symbols, analogies, images, and concepts for expressing the reality that evokes religious experience in an understandable way. Four basic conceptions of the divine may be distinguished: the divine as an impersonal, sacred order (Logos, Tao, *ṛta*, Asha) governing the universe and man's destiny; the divine as power that is holy and must be approached with awe, proper preparation, or ritual cleansing; the divine as all-embracing One, the ultimate Unity and harmony of all finite realities and the goal of the mystical quest; and the divine as an individual or self transcending the world and man and yet standing in relation to both at the same time.

The two most important concepts that have been developed by theologians and philosophers for the interpretation of the divine are transcendence and immanence; each is meant to express the relation between the divine and finite realities. Transcendence means going beyond a limit or surpassing a boundary; immanence means remaining within or existing within the confines of a limit. The divine is said to transcend man and the world when it is viewed as distinct from both and not wholly identical with either; the divine is said to be immanent when it is viewed as wholly or partially identical with some reality within the world, such as man or the cosmic order. The conception of the divine as an impersonal, sacred order represents the extreme of immanence since that order is regarded as entirely within the world and not as imposing itself from without. The conception of the divine as an individual or self represents the extreme of transcendence, since God is taken as not wholly identical with either the world or any finite reality within it. Some thinkers have described the divine as wholly transcendent of or "wholly other" than finite reality, some have maintained the total immanence of the divine, and still others claim that both concepts can be applied and therefore that the two characteristics do not exclude each other.

"Solitariness" and community

Social forms or expressions. Most enduring, historical religious traditions find their roots in the religious experience and insight of charismatic individuals who have served as founders; the sharing of their experience among disciples and followers leads to the establishment of a religious community. Thus, the social dimension of religion is a primary fact, but it need not be seen as opposed to religious experience taken as a wholly individual affair. There has been some difference of opinion on the point; Whitehead, for example, put emphasis on the "solitariness" of religious experience precisely in order to deny the claim of those who, like Émile Durkheim, a French sociologist, characterized religion as essentially a social fact. The social expression of religious experience results in the formation of specifically religious groups distinct from such natural groups as the family, the local society, and the state. Religious communities, including brotherhoods, mystery cults, synagogues, churches, sects, and monastic and missionary orders, serve initially to preserve and interpret their traditions or the body of doctrine, practices, and liturgical forms through which religious experience comes to be expressed. Such communities play a significant role in the shaping of religious experience and in determining its meaning for the individual through the structure of worship and liturgy and the establishment of a sacred calendar. Communities differ in the extent to which they stress the importance of individual experience of the divine, as distinct from adherence to a creed expressing the basic beliefs of the community. The tension between social and individual factors becomes apparent at times when the individual experience of the prophet or reformer conflicts with the norm of experience and interpretation established by the community. Therefore, although the religious community aims at maintaining its historic faith as a framework within which to interpret experience of the divine, every such community must find ways of recognizing both novel experience and fresh insight resulting from individual reflection and contemplation.

Objective "intention," or reference. Religious experience is always understood by those who have it as pointing beyond itself to some reality regarded as divine. For the believer, religious experience discloses something other than itself; this referent is sometimes described as the "intentional" object that is meant or aimed at by the experiencing person. Analysis of religious experience, interpretations placed upon it, and the beliefs to which it gives rise may result in the denial that there is any such reality to be encountered or that the assertion of it is justified by the experience in question. This conclusion, however, does not change the fact that all religious experience, whether that of the mystic who strives for unity with God or of the naturalist who points to a religious quality in life, purports to be experience "of" something other than itself. The question of the cognitive import or the objective validity of religious experience is one of the most difficult problems encountered in the philosophy of religion. In confronting the question, it is necessary to distinguish between various ways of describing the phenomena under consideration and the critical appraisal of truth claims concerning the reality of the divine made on the basis of these phenomena. Even if describing and appraising are not utterly distinct and involve one another, it is generally admitted that the question of validity cannot be settled on the basis of historical or descriptive accounts alone. Validity and cognitive import are matters calling for logical, semantic, epistemological, and metaphysical criteria—of the principles of rational order and coherence, meaning, knowledge, and reality—and this means that the appraisal of religious experience is ultimately a philosophical and theological problem. The anthropologist will seek to identify and describe the religious experience of primitive peoples as part of a general history and theory of man; the sociologist will concentrate on the social expression of religious experience and seek to

Cognitive import or objective validity

determine the nature of specifically religious groupings in relation to other groups—associations and organizations that constitute a given society; the psychologist will seek to identify religious experience within the life of the person and attempt to show its relation to the total structure of the self, its behaviour, attitudes, and purposes. In all these cases attention is directed to religious experience as a phenomenon to be described as a factor that performs certain functions in human life and society. As William Warde Fowler, a British historian, showed in his classic *Religious Experience of the Roman People* (1911), the task of elucidating the role of religion in Roman society can be accomplished without settling the question of the validity or cognitive import of the religious feelings, ideas, and beliefs in question. The empirical investigator, as such, has no special access to the critical question of the validity of religious experience.

The most radical form of the denial that religious experience has cognitive import is advanced by the Logical Positivists, who hold that all assertions or forms of expression involving a term such as "God" are meaningless because there is no way in which they can be verified or falsified.

Others who hold that religious utterance based on experience is without cognitive import regard it either as the expression of emotions or an indication that the person using religious language has certain feelings that are associated with religion. Those who follow the lead of Wittgenstein regard religious utterances as noncognitive but attempt to determine the way in which religious language is actually used within a circle of believers. Some psychologists have denied cognitive status to religious experience on the ground that it represents nothing more than man's projection of his own insecurity in the face of problems posed by life in the world and therefore has no referent beyond itself.

Immediacy and mediation. *Revelational and mystical immediacy.* Among defenders of the validity and cognitive import of religious experience, it is necessary to distinguish those who take such experience to be an immediate and self-authenticating encounter with the divine and those who claim that apprehension of the divine is the result of inference from, or interpretation of, religious experience. Two forms of immediacy may be distinguished: the revelational and the mystical (for a detailed treatment of the latter, see below *The experience of mysticism*). Christian theologians, such as Emil Brunner and H.H. Farmer, speak of a "divine-human encounter," and Martin Buber, a Jewish religious philosopher, describes religious experience as an "I Thou" relationship; for all three, religious experience means an immediate encounter between persons. The second form of the immediate is the explicitly mystical sort of experience in which the aim is to pass beyond every form of articulation and to attain unity with the divine.

Mediation through analysis and critical interpretation. A number of thinkers have insisted on the validity of religious experience but have denied that it can be understood as wholly immediate and self-supporting, since it stands in need of analysis and critical interpretation. Some, like Paul Tillich, hold that there are certain "boundary experiences," such as having an ultimate concern or experiencing the unconditional character of moral obligation, that become intelligible only when understood as the presence of the holy in experience. Others, such as H.D. Lewis and Charles Hartshorne, find the divine ingredient in the experience of the transcendent and supremely worshipful reality but demand that this experience be coherently articulated and, in the case of Hartshorne, supplemented by rational argument for the reality of the divine. Dewey envisaged a religious quality in experience pointing to God as an ideal that stands in active and creative tension with the actual course of events. Whitehead identified the presence of the divine with an apprehension of a "permanent rightness" in the scheme of things and based the validity of the experience on the claim that an adequate cosmology requires God as a principle of selection aiming at the realization of the good in the world process. James found the justification of religious experience in its consequences for the life of the individual: valid experience is distinguished by its philosophical reasonableness and moral helpfulness. Finally, some have sought to combine experience and interpretation by taking the traditional proofs of God's existence and pointing to their roots in the experience of perfection, of the contingency of one's own existence, and of the reality of purpose in human life. On this view, the arguments for the reality of God are not wholly formal demonstrations but rather the tracing out of intelligible patterns in experience.

Preparations for experience. Mystics, prophets, and religious thinkers in many traditions, both East and West, have been at one in emphasizing the need for various forms of preparation as a preliminary for gaining religious insight. The basic idea is that ordinary ways of looking at the world, dictated by the demands of everyday life, stand in the way of the understanding of religious truth; man must pass beyond these limitations by the disciplining of his mind and body. Three classic forms of preparation may be distinguished: first, rational dialectic for training the mind to reach insight (this explains why many mystical thinkers from the Pythagoreans to Nicholas of Cusa and Benedict de Spinoza were deeply involved in mathematics); second, moral preparation aiming at purity of heart, which was sometimes conjoined with bodily discipline, as in the Indian Yoga exercises; third, the use of drugs to expand the range of consciousness beyond that required for ordinary life. It is significant that the great mystics invariably regarded such preparation as necessary, but not sufficient, for experience. The self may be prepared, but the vision may not come; being prepared, as it were, establishes no claim on the divine. The experience described by St. John of the Cross, a 16th-century Spanish mystic, as "the dark night of the soul" points precisely to the experience of failure. The soul in this situation is convinced that God has abandoned it, cast it into darkness, perhaps forever. Mystics in the Taoist and Buddhist traditions have often emphasized the spontaneity of insight and the need to seek it through an "effortless striving" that combines the need to search with the awareness that the insight cannot be compelled. Zen Buddhists are fond of pointing to insights that are already possessed but not recognized as such until their holder is shaken loose from ordinary patterns of thought.

Three forms of preparation: rational dialectic, moral purification, drugs

SITUATIONAL CONTEXTS AND FORMS OF EXPRESSION

Cultic and devotional. Religious experience receives its initial, practical expression in the forming of the cult that provides an orderly framework for the worship of the religious object. Worship includes expressions of praise, acknowledgments of the excellency of the divine, communion in the form of prayer, and the use of sacraments or visible objects that signify or represent the invisible sacred beyond them, feelings of joy and of peace expressed often in musical form, and sacrifice or the offering of gifts to the divine or in the name of the divine. Worship is ordered by means of liturgy directing the experience of the worshipper in patterns that combine the written word, the spoken word, and sacred music in a unity aimed at bringing him or her into the presence of the divine.

Life crises and rites of passage. Religious experience has to do with the quality and purpose of life as a whole and with the ultimate destiny of the person. Certain special times and events in the course of life present themselves as occasions that are set apart and celebrated, because they direct man's thought to the divine and the sacred with peculiar forcefulness. These occasions, called life crises, are regarded as dangerous because they are transitional from one stage of life to another and open to view the relation of life as a whole to its sacred ground. Pregnancy and birth, the naming of a child, being initiated into the community—sometimes called "puberty rites"—the choice of a vocation, the celebration of marriage, and the time of death are experienced as special events distinct from the routine happenings of secular life. These events represent "crises"—*i.e.*, turning points—when man's relation to the sacred becomes a matter of special concern. As Gerardus van der Leeuw, a Dutch phenomenologist and historian of religions, points out, these transitional times are occasions for celebration in every culture because they mark the

death of one stage and the birth of another in a universal cycle of life.

Sacred and secular. The marking off of these crisis occasions from the routine events of daily life points to the all-important distinction between the sacred and the secular. As directed toward the sacred, religious experience finds expression in the specifically religious form of the cult and in the cycle of sacred life. There is, however, a secular as well as a sacred life, and, since religious experience concerns the whole of life, the religious meaning must be related to all the dimensions of secular life—political, economic, moral, technological, and other. The relationship is twofold; on the one hand, there is the bearing of the conception of the divine on standards of behaviour, and, on the other, there is the influence that the religious meaning has upon one's general attitude toward life. The sacred, thus, makes its impact on the secular by providing principles that are to govern the relations between persons and by holding before men a vision of the divine that gives purpose to life as a whole. Although the sacred retains its dynamism by becoming related to secular life, there is the constant danger that it will lose itself in the secular, unless specifically religious forms of life are preserved. The existence in every society of secret and mystery cults, of sacred brotherhoods, of groups of disciples devoted to holy men, of monastic orders, and, on the broadest scale, of established churches and denominations, points to the need felt to retain the sacred as a special domain that can neither be merged into nor contained within secular society.

Religious experience and secular life

Verbal, conceptual, and symbolic. In all of the world religions, religious experience receives its most enduring expression in the form of sacred scriptures and the body of commentary through which they are interpreted. Mythological and symbolic forms of expression are older than conceptual forms and systems of doctrine. Myth takes the form of a story and represents the imaginative use of materials drawn from sensible experience in order to express a religious meaning surpassing the sensible world. Myths of creation in many religions give ample evidence of this imaginative function. The task of the theologian using conceptual tools is to elucidate the thought content of the myth and other primary forms of religious expression—legend, parable, confession, lamentation, prophetic vision—and thereby reduce the degree of dependence on the sensible and imaginative elements. It is important to distinguish devotional and liturgical expressions from the theological use of language. Creeds, confessions, psalms and hymns of praise, litanies and scriptures containing a record of the lives and experiences of sacred persons, all give immediate expression to the primary experience upon which a religious tradition is founded. Systems of theology and religious philosophy make their appearance when it becomes necessary to conceptualize and express consistently the body of belief about the divine, the world, and man implied in this primary experience. Tension exists between religious experience and theological expression at two points: first, the pietistic and evangelical spirit in religion, as seen, for instance, in some forms of Protestant Christianity, and the *bhakti* devotional movement in Hinduism, seeks to preserve the primacy of experience at the expense of theology; and, second, those who acknowledge the indispensability of theology will also demand that its formulations remain in accord with the experience it is meant to express and interpret.

Tension between religious experience and theological expression

TYPES OF RELIGIOUS EXPERIENCE AND PERSONALITY

The personal character of religious experience makes it essential to understand its varieties as manifested in different types of personality and the functions they perform. The mystic, a reflective and contemplative type, shuts out the world and all distracting influences in order to reach true selfhood through purification and enlightenment. Although mysticism has social implications, the mystic is primarily an individualist, whereas the prophet, a person of intense but intermittent experience, sees himself called to be a spokesman for the divine to the community or all mankind, and regards his own experience as a message that enables him to interpret the past and the future in the light of the divine will. The priest is a mediator between man and the divine, and his main function is the proper ordering of worship through liturgical forms. By contrast with the prophet, whose insight is spontaneous, the priest attains the authority of his office through education and training; as guardian of the tradition, he must assume administrative responsibilities in addition to his role as spiritual adviser; thus he is both active and contemplative. The reformer is a figure who stands within a religious tradition and seeks to transform or revitalize it in the light of his own experience and insight. The reforms intended may be moral, intellectual, or ecclesiastical, depending on the particular genius of the individual. Common to all reformers is the conviction that some valid and essential feature of traditional faith has been ignored or distorted and that these deficiencies must be overcome if the religion is to be purified. It is characteristic of the reformer to be actively engaged in bringing about the reforms indicated by his renewing experience. The monk or member of a religious order is in search of a special or sacred place set apart from secular life within which a religious life can be lived and moral and religious demands fulfilled to a greater degree than is possible in the world. Different orders stress different aspects of experience: some emphasize ascetic practices and self-discipline; others are devoted to the preservation of learning and the development of theology; still others make missionary zeal uppermost, and the members are impelled by their own experience to seek to convert others. The forerunner of the monk, who lives in a community governed by rule, was the hermit or religious recluse, the type for whom solitary existence, preferably in deserts and barren places, is necessary for communion with the divine and self-purification. The saint is a figure venerated by the religious community as one who embodies perfection in some form. The saint may have been a martyr, exhibiting perfection in faith; a person possessed of intensified capacity for experience and communion with the divine; or one who achieves to a supreme degree the moral and spiritual ideals of the beatific life. The theologian has the task of expressing the historic faith of a community concerning the divine (*theos*) in rational or conceptual form (*logos*). The content of his thought, though handed on to him in its essentials by the tradition, will depend on his own experience and his insight into the special relevance of that tradition for his time. The theologian both interprets and reinterprets. The founder, as might be expected, surpasses all others in importance. The founder's experience forms the basis of his own authority and the substance of the religion he establishes. The intensity of his experience and the effect it has upon his personality are decisive factors determining the response of his initial followers and disciples. There is reason to believe that the founders of the great religions, such as Moses, Buddha, and Jesus, did not intend to fill this role; the founding of the religion in each case was the result of the impact of their personalities and of the profundity of their experience on those who gathered around them. (J.E.Sm.)

The founder's experience: Moses, Buddha, and Jesus

The experience of mysticism

Mysticism, a quest for a hidden truth or wisdom ("the treasure hidden in the centres of our souls"), in the 20th century is undergoing a renewal of interest and understanding and even a mood of expectancy similar to that which had marked its role in previous eras. Such a mood stems in part from the feeling of alienation that many persons experience in the modern world. Put down as a religion of the elite, mysticism (or the mystical faculty of perceiving transcendental reality) is said by many to belong to all men, though few use it. The British author Aldous Huxley has stated that "a totally unmystical world would be a world totally blind and insane," and the Indian poet Rabindranath Tagore has noted that "Man has a feeling that he is truly represented in something which exceeds himself."

NATURE AND SIGNIFICANCE

The goal of mysticism is union with the divine or sacred. The path to that union is usually developed by following

four stages: purgation (of bodily desires), purification (of the will), illumination (of the mind), and unification (of one's will or being with the divine). If "the object of man's existence is to be a Man, that is, to re-establish the harmony which originally belonged between him and the divinized state before the separation took place which disturbed the equilibrium" (*The Life and Doctrine of Paracelsus*), mysticism will always be a part of the way of return to the source of being, a way of counteracting the experience of alienation. Mysticism has always held—and parapsychology also seems to suggest—that the discovery of a nonphysical element in man's personality is of utmost significance in his quest for equilibrium in a world of apparent chaos.

Mysticism's apparent denial, or self-negating, is part of a psychological process or strategy that does not really deny the person. In spite of its lunatic fringe, the maturer forms of mysticism satisfy the claims of rationality, ecstasy, and righteousness.

There is obviously something nonmental, alogical, paradoxical, and unpredictable about the mystical phenomenon, but it is not, therefore, irrational or antirational or "religion without thought." Rather, as Zen (Buddhist intuitive sect) masters say, it is knowledge of the most adequate kind, only it cannot be expressed in words. If there is a mystery about mystical experience, it is something it shares with life and consciousness. Mysticism, a form of living in depth, indicates that man, a meeting ground of various levels of reality, is more than one-dimensional. Despite the interaction and correspondence between levels—"What is below is like what is above; what is above is like what is below" (*Tabula Smaragdina,* "Emerald Tablet," a work on alchemy attributed to Hermes Trismegistus)—they are not to be equated or confused. At once a praxis (technique) and a gnosis (esoteric knowledge), mysticism consists of a way or discipline.

Mysticism and religion

The relationship of the religion of faith to mysticism ("personal religion raised to the highest power") is ambiguous, a mixture of respect and misgivings. Though mysticism may be associated with religion, it need not be. The mystic often represents a type that the religious institution (*e.g.,* church) does not and cannot produce and does not know what to do with if and when one appears. As William Ralph Inge, an English theologian, commented, "institutionalism and mysticism have been uneasy bedfellows." Although mysticism has been the core of Hinduism and Buddhism, it has been little more than a minor strand—and, frequently, a disturbing element—in Judaism, Christianity, and Islām. As the 15th- to 16th-century Italian political philosopher Niccolò Machiavelli had noted of the 13th-century Christian monastic leaders St. Francis and St. Dominic, they had saved religion but destroyed the church.

The founders of religion may have been incipient or advanced mystics, but the inner compulsions of their experience have proved less amenable to dogmas, creeds, and institutional restrictions, which are bound to be outward and majority oriented. There are religions of authority and the religion of the spirit. Thus, there is a paradox: if the mystic minority is distrusted or maltreated, religious life loses its sap; on the other hand, these "peculiar people" do not easily fit into society, with the requirements of a prescriptive community composed of less sensitive seekers of safety and religious routine. Though no deeply religious person can be without a touch of mysticism, and no mystic can be, in the deepest sense, other than religious, the dialogue between mystics and conventional religionists has been far from happy. From both sides there is a constant need for restatement and revaluation, a greater tolerance, a union of free men's worship. Though it validates religion, mysticism also tends to escape the fetters of organized religion.

RELATION OF MYSTICAL EXPERIENCE TO OTHER KINDS OF EXPERIENCE

Mysticism shares a common world with magic, theurgy (power of persuading the supernatural), prayer, worship, religion, metaphysics (transcendent levels of reality), and even science. It may not be always easy to distinguish mysticism from these but its approach and emphasis are different. Though there is an element of magic, psychism, and the occult in much of what passes for mysticism, it is not to be equated with a science of the unseen or with voices and visions. Powers of the occult (or *siddhi*s) are viewed as real, but they can also be dangerous and are not of interest to genuine mystics, who have warned against their likely misuse.

Prayer and worship may form part of mysticism, but they are viewed as means and not as essence; also, they are usually continuations of sensory experience, whereas mysticism is a pure unitary consciousness, or a union with God. As for science, it is analytic and discursive and expresses its findings in precise and abstract formulas; mysticism, however, like poetry, depends more on paradoxes and an unusual use of language. Philosophies may lead to or follow from mysticism, but they are not the same. Nature mysticism is another prominent variant, to which poets and artists are particularly prone. This has often been described or dismissed as pantheism (the divine in all), though it is perhaps other than a simple assertion of identity.

Aesthetic mysticism

Emotionalism and purified emotion are quite different. Emotionalism, a kind of unsuccessful ecstasy, may arise from unpurged elements in the being; it could also be a concession or inability to hold the flow or touch from above. The natural state of man and, even more, that of the true mystic is serene and not agitated, not at the mercy of what the medieval mystical book *The Cloud of Unknowing* calls "monkey tricks of the soul." "Be still, and still, and know." Mysticism, among the many forms of experience, confirms the claims of religion and is viewed as providing a foretaste of the life after death.

DEFINITIONS OF MYSTICISM AND MYSTICAL EXPERIENCE

Differences between mysticism and similar phenomena. To define is to limit, and no single definition will cover every aspect of mysticism. Some have objected to the word itself and believed that "enlightenment" or "illumination" might be better. Though they meet, mysticism has to be distinguished from prophetic religions as well as from shamanism (a belief system built around psychic transformations). Working through chosen individuals—not necessarily saints and chosen for no other reason than God's will—prophetic religions emphasize action to a far greater extent than most forms of mysticism, with its penchant for inwardness and the beyond. Though in ecstasy the barriers seem to disappear, in prophetism God and man are rarely identified. Shamanism, a technique of ecstasy generally found in Siberia and Central Asia but with parallels in primitive society, provides a sort of correspondence with the purgative stage of mysticism (in which physical needs are negated). The closeness to paranormal (or supernatural) phenomena seems more pronounced, however, in shamanism. Both the shaman and the mystic, as communicants with a world beyond normal experience, reveal an identity of goal, if not of practice and content.

Basic patterns. Paradigmatic pronouncements in regard to mysticism pose problems of their own. The classic Indian formula—"that thou art," *tat tvam asi* (*Chāndogya Upaniṣad,* 6.9)—is hedged in with the profoundest ambiguity. The difficulty reappears in the thought of the medieval Christian mystic Meister Eckehart, who had the church raising questions for such unguarded statements as "The knower and the known are one. God and I, we are one in knowledge" and "There is no distinction between us."

Mysticism as a third kind of knowledge

Mysticism may be defined as the belief in a third kind of knowledge, the other two being sense knowledge and knowledge by inference. Adolf Lasson has written:

> The essence of Mysticism is the assertion of an intuition which transcends the temporal categories of the understanding. . . . Rationalism cannot conduct us to the essence of things; we therefore need intellectual vision.

This same view was held by the 3rd-century-AD Greek philosopher Plotinus. But the pattern misses the other dominant quality of mystical experience—love, or union through love. The medieval, theistic view of mysticism (as of religious life) was that it was "a stretching out of the

soul into God through the urge of love, an experimental knowledge of God through unifying love." Its other name was joy, and the endeavour of the mystic to grasp the divine essence or ultimate reality helped him to enjoy the blessedness of actual communion with the highest. This was considered both a science and an art. As a science (*i.e.*, intuitive knowledge, or the "science of ultimates"), mysticism is viewed as being able to help in "the overcoming of creatureliness," and also as being able to maintain "the tendency to stress up to an extreme and exaggerated point the non-rational aspect of religion."

Reality, a kingdom of values, is viewed not as a faceless infinite, an impersonal something or somewhat. If not an ego, it is a being, and most mystics would call it God. Mysticism arises when man tries to bring the urge toward a communion with God—a "Being conceived as the supreme and ultimate reality," according to the British scholar William Ralph Inge—toward a higher consciousness and being in relation with the other contents of his mind and total personality, when he tries to realize the presence of the living God in the soul and in nature or, more generally, in the attempt to realize (in thought and feeling) the immanence of the temporal in the eternal. A 19th-century scholar, Otto Pfleiderer, indicated that religious mysticism is "the immediate feeling of unity of the self with God; it is nothing, therefore, but the fundamental feeling of religion, the religious life at its very heart and centre." Against such exclusive concentration the British writer Richard Nettleship suggests a corrective element, that of wholeness and symbolism. "Mysticism is the consciousness that everything that we experience is an element, and only an element, in fact, *i.e.* that in being what it is, it is symbolic of something else."

Introvertive mysticism. Certain forms of mysticism, however, would seem to strive toward a naked encounter with the Whole or All, without and beyond symbols. Of this kind of direct apprehension of the absolute, introvertive mysticism offers examples from different times and traditions. Instead of looking out, the gaze turns inward, toward the unchanging, the undifferentiated "One without a second." The process by which this state is attained is by a blotting out or suppression of all physical sensations—indeed, of the entire empirical content of consciousness. *Cittavṛttinirodha* ("the holding or stopping of the mind stuff") was how the 2nd-century-BC Indian mystic Patañjali described it. The model of introvertive mysticism comes from the *Māṇḍūkya Upaniṣad:*

Unitary consciousness—One without a second

> The Fourth, [aspect of self] say the wise, . . . is not the knowledge of the senses, nor is it relative knowledge, nor yet inferential knowledge. Beyond the senses, beyond the understanding, beyond all expression is The Fourth. It is pure unitary consciousness wherein [all] awareness of the world and of multiplicity is completely obliterated. It is ineffable peace. It is the supreme good. It is One without a second. It is the Self. (From *The Upanishads, Breath of the Eternal;* trans. by Swami Prabhavananda and Frederick Manchester.)

Other definitions and experiences of mysticism. Such undifferentiated unity or union between the individual and the supreme self is unacceptable to certain traditions and temperaments. The Jewish philosopher Martin Buber emphasized an "I–Thou" relationship: "All real living is meeting," and one Thou cannot become It. But even his own "unforgettable experience" of union he would explain as "illusory." With a wider range, a British scholar, R.C. Zaehner, has tried to establish different kinds, or types, of mysticism: of isolation, the separation of spirit and matter, eternity from time; pantheistic, or "pan-enhenic," in which the soul is the universe—all creaturely existence is one; the theistic, in which the soul feels identified with God; and the beatific, with its hope of deification when "the perishable puts on the imperishable."

Definitions of mysticism include a bewildering variety, ranging from the biological through the psychological to the theological. The origin of the word and certain of its features strongly suggest the possibility that mysticism is the science of a hidden life. But there is also a growing belief among 20th-century scholars that "the people of the hidden" should not remain hidden too long and should come out in the open, befitting an era of "open development" and "open realization." Some 20th-century scientists, among them physicists, biologists, and paleontologists, have shown a marked mystical bias. A biologist, Ludwig von Bertalanffy, has confessed to "peak experiences" of a great unity and liberation from ego boundary: "In moments of scientific discovery I have an intuitive insight into a grand design." He finds no necessary opposition between the rational way of thinking and intuitive experience culminating in what the mystics have tried to express. Both have their place and may coexist. Earlier there had been a sharp dichotomy between scientific and mystical knowledge. The logic of levels may never be amenable to analysis or intellectual understanding, but that is not to deny the role of reason.

Attitudes toward mysticism since the middle of the 20th century have been considerably modified by an awareness of subliminal consciousness, extrasensory perceptions, and, above all, an evolutionary perspective. The Roman Catholic paleontologist Teilhard de Chardin asked if in an expanding universe mysticism would not burst the limits of narrow cults and religious rigidity and move toward an ecumenical future. In a larger view, mysticism has not so much to be defined as renewed and redefined.

20th-century views of mysticism

UNIVERSAL TYPES OF MYSTICAL EXPERIENCE

Intellectual and contemplative forms. Mystical experience, which is centred in a seeking for unity, admits of wide variations but falls into recognizable types: mild and extreme, extrovertive and introvertive, and theistic and nontheistic. Another well-known typology—corresponding to the faculties of thinking, willing, and feeling—employs the Indian formula, the respective ways of knowledge (*jñāna*), works (*karma*), and devotion (*bhakti*). Claims have been made on behalf of each, though maturer mystics have tried to accord to each its place and also to arrive at a synthesis, as in the *Bhagavadgītā* (Hindu sacred scripture). Depending on the powers of discrimination, the intellectual or the contemplative type tries to reach the Highest, the One, or the Godhead behind God. In its approach toward the supreme identity it tends to be chary of multiplicity, "to deny the world that it may find reality." Plotinus was "ashamed of being in the body." In the 17th century, Spinoza's nondenominational concept of intellectual love of God revealed a sense of aloofness or isolation reminiscent of the ancient Hindus.

Man, however, does not live by thought alone; to live is to work, and faith without works is dead. The mystic injunction is that works should be done in a spirit of nonattachment, with the ego sense (I, the doer) taken away. In a larger sense, not merely the doing of religious chores but all activity is offered to the Supreme. All life, according to many mystics, turns into a sacrament. "All life is *yoga* (meditation practice)."

Devotional forms. For the emotional type of person there is the mysticism of love and devotion. A theistic attitude, or devotional mysticism, depends upon mutual attraction. In the words of a Ṣūfī poet, "I sought Him for thirty years, I thought that it was I who desired Him, but no, it was He who desired me." The path of devotion includes the rituals of prayer, worship, and adoration, which—if done with sincerity, inwardness, and understanding—can bring some of the most rewarding treasures of the religious life, including ecstasy (or *samādhi*). There is a paradox and a danger here: the paradox of avoiding the loss of personality, the danger of self-indulgence.

Ecstatic and erotic forms. Also, in an unpurified medium, the experiences may and do give rise to erotic feelings, a fact observed and duly warned against by the wiser spirits and the Fathers of the Church. (Zen Buddhism avoids both the overly personal and erotic suggestions.) Sometimes the distinction between *eros* (Greek: "erotic love") or *kāma* (Sanskrit: "sexual love") and *agapē* (Greek: "a higher love") or *prema* (Sanskrit: "higher love") can be thin. In the Indian tradition the Vaiṣnava (devotional) and Tantric (sexual) experiments were, in their apparently different ways, bold and honest attempts at sublimation, though the majority of these experiments turned out to be failures and disasters.

The same fate is likely to overtake the craze for

Use of drugs as aids to visionary experience

psychedelic drugs and pharmacological aids to visionary experience—practices that are by no means new. A yogic writer, Patañjali, speaks of the use of *auṣadhi* (a medicinal herb) as a means to yogic experience, and the Vedas (Hindu scriptures) and *Tantras* (Hindu occultic writings) refer to wine as part of worship and the initiatory rites. The Greek Mysteries (religions of salvation) sometimes used sedatives and stimulants. Primarily meant to remove ethical, social, and mental inhibitions and to open up the subconscious no less than the subliminal, these techniques, as a rule, were frowned upon, even though those who took the help of such artificial aids had undergone prior training and discipline.

A whole new life-style and vocabulary have developed around medicinal mysticism in the 20th century. Peyote, mescaline, hashish, marijuana, *Cannabis indica,* LSD, and other similar products have become familiar to much of the world's population. The visions induced by such aids at best resemble the extrovertive type and cannot be easily equated with genuine mystical experience. According to taste, temperament, and tradition, the experience—a parody of creative spontaneity—may come from unexpected sources. In any case, utilizing such medicinal aids rarely achieves union with Self or God, and no permanent change of personality (in the mystical sense) has been known to occur.

GOAL OF MYSTICAL EXPERIENCE AND MYSTICISM

Experience of the divine or sacred. The goal of mysticism is "ghostly," a state or condition in which the soul is "one'd with God," according to the Western medieval work *The Cloud of Unknowing.* This "one-ing" is because all men, according to mystics, are called to their origin. Self-realization is basically one in intent with the injunctions of the Greek Mysteries: "Know thyself." This knowing, union, or communion with the divine and the sacred is of the essence of the ascent of man. As the only answer to the problem of identity, mystics look upon it as the final end, the *summum bonum.* At the journey's end waits the knowledge by identity. The direct, intuitive perception is more akin to revealed religion than to science and philosophy, though it is of itself a science, and philosophies spring from as well as lead to it.

Union with the divine or sacred. In the movement toward the goal there are, naturally, stages and processes, marked differently in different traditions. The discipline of prayer, purification, and contemplation culminates in the highest wordless union with the divine and the ultimate. As the process unfolds itself along the mystic way, an alteration of personality—a conversion, if not reintegration—occurs. The unregenerate "old man" (in Christianity) is replaced by the new being. The "twice born" (in Hinduism) becomes more than a metaphor or sacrosanct social arrangement. There is a change of level and mind. One of the aims or methods of mysticism is to make possible this change and conversion, a shift from the profane to the sacred, from "here" to "there": "Lead me from the unreal to the real, from darkness to light, from death to immortality" (*Bṛhadāraṇyaka Upaniṣad*). Before the transition, or the "great passage," is completed, however, the individual or pilgrim feels successively or simultaneously his oneness with nature, with people, and with things—an extension of awareness or expanded selfhood to which no limits can be assigned. Cosmic consciousness is thus a stage in a progressive self-discovery.

Conversion from the profane to the sacred

Experience of the universal. The nature of the goal, however, introduces a paradox. Like every other aim and activity, mysticism operates in a historical context. Yet, sooner or later, it also tends to reveal a timeless stance. The mystic is both in and out of time. The eternal now is a kind of release from the temporal order. Such a release may lead to a shift from the local to the universal, to a growing sense of unity of all experience. Though not a declared or conscious aim, this result could be looked upon as a not unworthy goal as well as a pragmatic standard.

To cure man of a provincialism of the spirit, from which more people suffer than either know or admit it, is one of the goals of a mysticism that has come of age. The true mystic is a cosmopolitan. In man's many-sided growth toward the real, a sane and mature mysticism leads to an ecumenical insight and obligation. Local colour, particulars, and uniqueness will not cease, but, in the perspective of the future and of wholeness, the universal alone will have survival value.

Experience of oneness with people. The apotheosized (divinized) field of consciousness is mysticism's ultimate goal and gift to the life of an evolving humanity. It alone is fitted to mediate between the anguish of existence and the serenity of essence, between saṃsāra ("cycle of birth and rebirth") and Nirvāṇa (the State of Bliss). According to an American Roman Catholic mystic, Thomas Merton, "The spiritual anguish of man has no cure but mysticism."

Though the mystic goal may seem to be tied to a transcendent reality, this does not mean a sundering of all relations and responsibilities. On the contrary, it is the guarantee of a set of altered relationships and a rehabilitation of what may be called the higher reason. Intuitions that sink into private fancy and morbidity have a short life to live. As for the mystic's "yonder," it is not spatially or posthumously remote but rather refers to a different order of reality and consciousness. The healthier forms of mysticism do not abjure action or the claims of love. It is an ancient maxim that one becomes what one loves. This is how the psychic birth repeats itself in the mystic soul, as stated, for example, by Meister Eckehart, a medieval German mystic: "It is more worthy of God that he should be born spiritually of every virgin, or of every good soul, than that he should have been born physically of Mary."

The claims of love versus mysticism

The mystic is not always amorous of the beyond, leaving an unredeemed world to its own ways. Not escape but, rather, victory is mysticism's inner urge and promise. The more sober among the mystics do not merely withdraw; they also return to the base and attempt the ancient alchemy, the transformation of men. A solitary salvation does not satisfy either head or heart.

MYSTICAL RELATIONSHIP BETWEEN MAN AND THE SACRED

Nature of the relationship. Within man is the soul of the holy, said Ralph Waldo Emerson in the 19th century. This is true of society, too. As the French sociologist Émile Durkheim saw it, the sacred is but a personified society. Mysticism, one might say, is the art and science of the holy. Theologically, it is but "the experience of the Holy Ghost, . . . the realization of the Spirit of Holiness." As the opposite of the profane and as a distinct and irreducible quality of the religious and mystical life, the sacred has always existed. It is indeed a mark of the real, and, when the German theologian Rudolf Otto isolated the sacred as a "quite distinctive category" of mystical apprehension, he had no lack of evidence. The emphasis, however, was not unanimously accepted. Some, like Inge, thought the sacred might as well be elicited from such ultimate values as "truth, goodness, and beauty."

According to the respective world view, the interpretation or emphasis varies, but the universal core remains unaffected. The sacred is in its own way a coherent system, though not rational. The dualists no less than the theists insist on the unqualified and irreducible "otherness," the unbridgeable gulf, even when one speaks of union or communion. It is the distance that preserves the sacred.

The "otherness" or unbridgeable gulf of the sacred

Christian mystics, who often speak of "union with God," generally do not imply identity with the divine, since this might lead to heresy. The 16th-century Spanish mystic St. Teresa of Avila could write with impunity: "It is plain enough what unity is—two distinct things becoming one." But most others could not be so plain and had to use special strategy to cover up traces of possible deviation from what was permissible. Even if there had been a semblance of interpenetration between man and the divine, there could be no substantial identity. "Each of these," wrote the medieval Dutch mystic Jan van Ruysbroeck, "keeps its own nature. There is here a great distinction, for the creature never becomes God, nor does God ever become the creature." The same doctrine is preached in the Middle Ages by the mystic Heinrich Suso:

> In this merging of itself in God the spirit passes away and yet not wholly; for it receives indeed some attribute of God, but it does not become God by nature. It is still something

that has been created out of nothing, and continues to be this everlastingly.

Identification of man with the divine, according to many the heart of mysticism, raises problems from other points of view as well. Pantheism, which asserts that all is God (or Nature), and God (or Nature) is all, is looked upon as a false doctrine in many religions. To John Calvin's leading question—"The Devil also must be God, substantially?"—the unsuspecting Spanish theologian and physician Michael Servetus had answered smilingly: "Do you doubt it?" The opinion cost him his life. The Hindus' *Upaniṣads,* however, insist on this identity in passage after passage. Closely looked at, this may not be simple pantheism but an identity in difference, a paradox present in even Vedānta (a Hindu monistic system). Islām has been fiercely critical of these claims of oneness and the medieval mystic al-Ḥallāj had to pay with his life (922) for making the unorthodox announcement of his identity with the divine: "*Anā al-ḥaqq*" ("I am the Truth"). He was not the only one to speak in this manner. The more moderate Maḥmūd Shabestarī had reported an experience (*c.* 1320):

> In God there is no duality. In that presence "I" and "we" and "you" do not exist. "I" and "you" and "we" and "He" become one. Since in the unity there is no distinction, the Quest and the Way and the Seeker become one.

But Muslim theologians as a rule tended to dismiss those who "boasted of union with the Deity" as merely "babblers." In the Jewish tradition, it is generally considered improper and indecorous for any man to give a personal account of his own mystical experience.

Behind these and other interpretations, the reality of the sacred—and its persistent ambiguity—appears to be too true to be denied or ignored. Awe may or may not be the best part of man, but without it a necessary dimension is left out of the image of man, the dimension of what Otto called the *mysterium tremendum et fascinans* ("the mystery that repels and attracts"). The mystics are loath to leave this dimension out and, directly or indirectly, insist on its inclusion. The reason was suggested in the 5th–4th centuries BC by the Greek philosopher Plato, who maintained that the divine was the head and root of man. The mystic's is the eye, the third eye, with which the world beholds itself and knows itself divine. Though the vision is partial and passes away, there could be an ideal state of unbroken awareness of the Real Presence, an epiphany (manifestation).

The dimension of attraction and repulsion

According to the mystical point of view, the rational content of religion is not enough; it is not of the essence of religion. The sense of the holy, the mark of man's encounter with the "other," is usually invested with an ethical aura or undertone. This is how most people understand it. But this lowers its potency considerably. There is clearly an overplus, below good and evil and beyond good and evil. The numinous (spiritual) is not altogether free of the ominous. Thus, though the holy may be discussed, it cannot be well defined. It can, however, be experienced and evoked, as part of that wordless mystery that man must face—even if he is not able to explain satisfactorily—in his journey toward the real. This may happen early or late in his mystical journey, and the notion of evolution may not be applied to it uniformly.

The holy is not always and altogether a pleasant experience. Often shrouded in a fear that is more than fear, it is an inward dread and shuddering. The holy as awe-inspiring can be found in the Indian pantheon in such figures as Rudra and Kālī, the dark and wrathful faces of the divine, in which—in a collapse of finitude—majesty and unapproachability are inexplicably blended. The feeling of being consumed in the presence of the divine is a profound expression of man's relation to the holy. As for the ultimate mystical identity with the Supreme, Self, God, or the Unknowable, that also confirms the nonrational and suprarational nature of the experience in which ego, logic, and grammar are shattered alike. A frightful and traumatic adventure, not unlike the Greek Mystery rites, it can erupt at every crisis, break through an insulated universe. A clergyman cited by William Starbuck, an American psychologist of religion, spoke of having experienced

> a [silent] presence [in the night], all the more felt because it was not seen. I could not any more have doubted that *He* was there than that I was. Indeed, I felt myself to be, if possible, the less real of the two.

Diminution apart, the holy generally gives rise to a sense of energy and urgency, which may take different forms. At a higher level, the consuming fire of love reported by mystics could be an extension or refinement of the same energy, for "Love is nothing else than quenched Wrath." The "nothing else" may be an exaggeration, but such paradoxes of the religious life—*e.g.,* the unity of opposites—meet man at every turn. The void in Buddhism, like the nothing in Western mysticism, may be a numinous ideogram of the "wholly other."

Means and modes of the relationship. As means to meet the divine, some mystics have taken recourse to fasting, breath control, meditation, ecstasy, simplification, autosuggestion, and monoideism (absorption in a single idea). Rituals, in some cases, provide contact. An old method is the *via negativa* ("negative way"): "the emptier your mind, the more susceptible are you to the working of the presence." In other words, the impediments have to be removed. Among other indirect—but no less effective—means would be the shock therapy of the blood-curdling images that one notes in Tibetan iconography and symbology, which have their links with the archaic and the chthonian (infernal). On more negotiable levels, works of art—as far apart as Sung (Chinese) paintings, Gothic cathedrals, medieval temple architecture in India, the Egyptian Sphinx, music such as the *Missa Solemnis* or Sanskrit (Hindu) hymns—are accredited conductors of the numinous. Darkness, solitude, silence, and emptiness are sometimes enough for the sensitive soul, and the doors of perception open to a wider world beyond. A wide stretch of land or cranes flying against a cloudy sky were enough to throw the 19th-century Indian saint Ramakrishna into transport. But, always, it is less the object than something seen through the object, a bodiless presence, that forms the essence. Without symbols in which the holy is embodied, the experience of the holy vanishes.

The *via negativa*

Though it creates a sense of awe and exaltation, the idea of the holy also produces a mood of dependence, leading to action aimed at appeasing the deity or the powers behind the universe. At first, the policy of appeasement may have been inspired by fear and hope of reward. But, since the deity is not ultimately malevolent, it could also evolve into an idea of grace. Mystical theology, both in the East and in the West, has sometimes been divided over the issue whether the union with the divine is the result of one's unaided effort or supernatural grace.

The approaches to the divine or sacred are various rather than uniform. Moving through physical, intellectual, devotional, and symbolic rituals and disciplines, it moves toward the ultimate goal: the annihilation of the self, *unio mystica* ("mystical union") in Western Christianity, *mokṣa* ("salvation") in Hinduism, Nirvāṇa (the State of Bliss) in Buddhism, and *fanā'* ("the snuffing out of self") in Islām. Though the words differ, the experiences are perhaps allied, if not the same. In a Ṣūfī (Islāmic mystical) poem the divine voice speaks exultantly:

The various approaches to the divine

> Annihilate yourself gloriously and joyously in Me, and in Me you shall find yourself; so long as you do not realize your nothingness, you will never reach the heights of immortality.

The description could as well be applied to the Buddhist *śūnyatā* ("void") and the self-negating of the Christian mystics.

The ultimate has been, as a rule, thought of as something "other" and apart, even if in mysticism what is sought is union or unity. Hierophany (manifestation of the holy) implies a choice and a distinction: between that which manifests the sacred and that which does not. Also, though a hierophany may represent a historic event that does not minimize its larger validity (and in any culture there may be local as well as general hierophanies), a hierarchy is not unlikely. On occasions, the sacred may manifest itself in something profane. Ideally, to a mystic, "the integrated quality of the cosmos is itself a hierophany." From this follows the possibility of consecrating the whole of life, so that by sacramental transformation, at any moment, "the

flash of a trembling glance" may be inserted into the great time and project the man amphibian (having dual life) into eternity. Deification, without doubt one of the goals of the mystical life and a fundamental concept of orthodox Christendom, is part of the dialectics of the sacred. The alchemic undertone, in the man–God idea, has never wholly been extinguished. But, as part of the continuing paradox, one should also mention a resistance to the sacred. Depending on the ambivalence of the response to the sacred, which at once repels and attracts, the resistance is ultimately a flight from reality.

SEMANTICS AND SYMBOLISM IN MYSTICAL EXPERIENCE

Role of symbols and paradoxes

Union of opposites. Mystical experience is flanked with a communication hazard, a "polar identity." The linguistic liberties and extravagances are part of the logical impossibility of having to describe one order of experience in terms of another. Hence, the rhetoric of mysticism is largely one of symbols and paradoxes. The most striking of the strategies, as the medieval Christian scholar Nicholas of Cusa put it, is *coincidentia oppositorum* ("union of opposites"). Since the opposites coincide without ceasing to be themselves, this also becomes an acceptable definition of God, or the nature of the Ground. God, said Heracleitus, is day and night, summer and winter, war and peace, and satiety and hunger—all opposites. A 5th- to 6th-century-AD Christian mystical writer called Dionysius the Areopagite advised people to

> strip off all questions in order that we may attain a naked knowledge of that Unknowing and that we may begin to see the superessential Darkness which is hidden by the light that is in existent things.

This use of language or view of things is obviously not normal.

Old myths and archetypes are full of examples of such dichotomy. The Zoroastrian tradition has Ormazd (the Good Lord) and Ahriman (the Lie); the Gnostic myth speaks of Christ and Satan as brothers; and the same idea is found in the Vedas, where the *suras* ("good spirits") and *asuras* ("bad spirits") are shown to be cousins. In a different context there is the androgyne ("man–woman"), the *ardhanārīśvara* in Indian myth. As for the Hindu *jivanmukta,* the liberated individual, he is liberated from duality. This is also part of what the Lord Kṛṣṇa (Krishna) said, when he asked the hero Arjuna to rise above the three *guṇas* ("modes"). The Tantras refer to the union of Śiva (a Hindu god) and Śakti (Śiva's consort) in one's own body and consciousness and provide appropriate practices to this end. The Chinese had their Yang and Yin (opposites), the Tibetans their Yab and Yum (opposites), and Buddhism its saṃsāra and Nirvāṇa as aspects of the Same. In *Prajñāpāramitā,* a Mahāyāna (northern Buddhist) text, the Illumined Ones are supposed to engage in a laughter in which all distinctions cease to exist.

Emptiness and fullness. Mystical experience permits complementary and apparently contradictory methods of expression: *via affirmativa* ("affirmative way," or fullness) as well as *via negativa* ("negative way," or emptiness). For fullness and freedom both are needed. This is because the reality affirmed contains its own opposite. In fact, the apparent negations—*neti-neti,* ("not this, not that") of the *Upaniṣads,* the *śūnyatā* ("void") of the Buddhists, or the Darkness beyond Light of Dionysius—perform a double function. They state a condition of being as well as its utter freedom from every determination. As Dionysius explains it, "While God possesses all the attributes of the universe, being the universal Cause, yet in a stricter sense He does not possess them, since He transcends them all." The "negative way," a way of turning the back upon the finite, is part of an old, positive, verified insight, at once the last freedom and, as far as many men are concerned, perhaps a lost freedom.

Symbolism of divine messengers. Experiences relating to these realities could not at any time have been common or widespread and must have come mainly through consecrated channels: yogis (Hindu meditation practitioners), gurus (Hindu teachers), prophets, mystics, saints, and spiritual masters of the inner life. This channelling through human agents has given rise to a host of divine messengers: a hierarchy of angels, intermediaries, and incarnations, singly or in succession. This manner of approaching or receiving the divine or holy is the justification of avatars (incarnations of God) and the man–God in various religions. "God was made man in order that man might be made God."

Participation in two worlds at once

The mystical experience is a renovation of life at its root; that is, of the forgotten language of symbols and symbolism. The mystic participates in two worlds at once, the profane and the sacred. Rituals and ceremonies become the means of integration with a higher reality and consciousness. But symbols cannot be deliberately manufactured, nor do they make an arbitrary system. "Being for ever communicating its essence" is the source of their abundance, potency, and unity. Even a nontheistic mysticism, such as Buddhism, has deployed symbols freely, of which perhaps the most well-known is the formula *oṃ maṇi padme hūṃ* ("the jewel in the lotus").

Symbols point beyond themselves, participate in that to which they point, open up levels of reality that are otherwise closed to man, unlock dimensions and elements of the soul that correspond to reality and cannot be produced intentionally or invented. Symbols may be inner or outer. To some, nature symbolism comes easily.

Symbolism of love and marriage. A far more risky but inescapable mode of symbolism than pantheism has been the use of the analogy of human love and marriage. Not all the mystics have been deniers or champions of repression. The soul, it may be added, is always feminine. The Christian mystics St. Bernard and St. John of the Cross, the Islāmic Ṣūfī poets, and the Hindu Dravidian and Vaiṣṇava saints could teach lovers. Not only the church but the faithful are viewed to be among Christ's brides and speak the language of love. "O that you would kiss me with the kisses of your mouth!" The speaker is the bride, thirsting for God. St. Bernard has shown that through carnal, mercenary, filial, and nuptial love the life of man moves toward the mystery of grace and union.

The hermeneutics (critical interpretation) of "the Bridegroom-Word" is that "the soul's return is her conversion to the Word, to be reformed through Him and to be conformed to Him." In the West, the roots of the tradition go back to the Song of Solomon in the Bible, not, perhaps, the best of models. The Hindu *līlās* ("love plays") of Rādhā and Kṛṣṇa have been freely misunderstood in spite of the repeated disclaimer that the events described are not facts but symbols. The charge of immorality has been loudest against the Tantras, which had made a subtle, bold, and strict experiment in sublimation, whose inner sense may fail to be intelligible even to those who are attracted by it. That the marriage symbol should find a readier response among the brides of Christ is only to be expected. In *The Interior Castle,* St. Teresa has been fairly outspoken: "He has bound Himself to her as firmly as two human beings are joined in wedlock and will never separate Himself from her." But this was not a monopoly of nuns. The medieval theologian Richard of Saint-Victor has described as well as explained the "steep stairway of love" made up of betrothal, marriage, wedlock, and fruitfulness. In a slightly different set of symbols, St. John of the Cross states that after the soul has driven away from itself all that is contrary to the divine will, it is "transformed in God by love."

The notion of the "stairway of love"

Symbolism of the journey. Another prominent mystical symbol is the way, quest, or pilgrimage. Having lost the paradise of his soul, man, as the 16th-century physician and alchemist Paracelsus says, is a wanderer ever. A Christian monk, St. Bonaventure, has written about the mind's journey to God, and an English mystic, Walter Hilton, has described the Christian journey thus:

> Right as a true pilgrim going to Jerusalem, leaveth behind him his house and land, wife and child, and maketh himself poor and bare from all that he hath, that he may go lightly without letting: right so, if thou wilt be a ghostly pilgrim, thou shalt make thyself naked from all that thou wouldst be at Jerusalem, and at none other place but there. (From *The Ladder of Perfection.*)

According to the Ṣūfīs, the pilgrim is the perceptive or intuitional sense of man. Aided by attraction, devotion,

and elevation, the journey leads, by way of many a wine shop (divine love), to the tavern (illumination), "the journey to God in God." In his *Conference of the Birds,* the 12th-century Persian Ṣūfī ʿAṭṭār refers to the seven valleys en route to the king's hidden palace: the valleys of quest, love, knowledge, detachment, unity, amazement, and, finally, annihilation. Others have gone further and spoken of "annihilation of annihilation." In the symbolic universe, denudation may be viewed as a way of fullness.

Men are called to the journey inward or upward because of a homing instinct. Eckehart put the matter simply: earth cannot escape the sky. All men are called to their origin, which implies God's need of man. A mutual attraction, the tendency toward the Divine cannot be stifled indefinitely, since it returns after every banishment. For some, paradise is not enough; it is too localized and perhaps perishable. They strain toward eternity, a leap beyond history into the incommunicable forever. A white radiance to some, to others it is "a ring of pure and bright light." The Veda speaks of the *kālahaṃsa* ("the swan of time") winging back to the sky and nest, eternity.

Essentially a way of return, *ricorso,* the final aim of mysticism is transfiguration. But

> by what alchemy shall this lead of mortality be turned into that gold of divine Being? But if they are not in their essence contraries? If they are manifestations of one Reality, identical in substance? Then indeed a divine transmutation becomes conceivable. (From Sri Aurobindo, *The Life Divine.*)

This is a clue to the Vedas, those hymns to the mystic fire and the inner sense of sacrifice, burning forever on "the altar Mind." Hence the abundance of solar and fire images: birds of fire, the fire of the sun, and the isles of fire. The symbol systems of the world religions and mysticisms are profound illuminations of the human-divine mystery. Be it the cave of the heart or the lotus of the heart, "the dwelling place of that which is the Essence of the universe," the third eye, or the eye of wisdom—the symbols all refer back to the wisdom entering the aspiring soul on its way toward progressive self-understanding. "I saw my Lord with the Eye of the Heart. I said, 'Who art Thou?' and he answered, 'thou'." Throughout the ages man, *homo symbolicus,* has been but exploring the endless miracle of being. Mystical experience is a living encyclopaedia of equations and correspondences, pointer readings that partly reveal and partly conceal.

Solar and fire symbolism

PSYCHOLOGICAL ASPECTS OF MYSTICISM

Awareness. Mysticism has been accused of passing off psychological states for metaphysical statements. But the psychological base has never been questioned seriously. It would, however, be proper to call it autology (the science of self). If the word psychology is to be retained, it must be in the original sense of the word now discarded. The contrast between the old and the new has been well expressed by the Russian philosopher P.D. Ouspensky:

> Never in history has psychology stood at so low a level, lost all touch with its origin and meaning, perhaps the oldest science and, unfortunately, in its most essential features, a forgotten science, the science of [man's] possible evolution.

Mysticism is that science in which the psychology of man mingles with the psychology of God. The major change or orientation is from the level of the profane to the sacred, an awareness of the divine in man and outside. The source and goal of such a psychology was revealed in the 18th-century Methodist leader John Wesley's dying words: "The best of all is this, that God is with us."

A mark of the mystic life is the great access of energy and enlarged awareness, so much so that the man who obtains the vision becomes, as it were, another being. Mansions of the mind, *maqām* (Arabic: "place"), and *bhūmi* (Sanskrit: "land"), open up to the gaze of the initiate, a wayfarer of the worlds. This means a renewal or conversion until one knows that the earth alone is not man's teacher. The mystic begins to draw his sustenance from supersensuous sources. He has "drunk the Infinite like a giant's wine," and a hidden bliss, knowledge, and power begin to sweep through the gates of his senses.

Role of identification. The state of energizing is facilitated by controlled attention. It is customary to fix the mind on some object or idea, some focus of contemplation. According to the Indian formula, to worship God one must become like him (*devam bhutva devam yajet*). Exercises, physical no less than mental, including methods of worship and prayer, have been developed to this end until one becomes what one contemplates. The ranges and creative aspects of the mind are part of the psychology of the mystics and one of the oldest traditions of mankind. The old Indian psychology divided consciousness into three provinces: waking state (*jāgrat*), dream state (*svapna*), and sleep state (*suṣupti*), and added a fourth (*turīya*), which is the consciousness of man's pure self-existence or being. The fourfold scale represents the degrees of the ladder of being by which man climbs back to the source, the absolute divine. The change, from "here" to "there," is not an uneventful process. There come dry periods, deviations, violent alterations, and temptations. If there are raptures and blue heavens, there are python agonies and absolute abandonments, howling deserts and "dark nights of the soul" to go through. Tears of joy, horripilation (bristling of the hair), stigmata (bodily marks or pains), and parapsychological phenomena have been known to develop.

Indian states of consciousness

The earlier phases of a naturalistic psychology had no qualms in relegating most of these experiences to the scrap heap of obsolete and archaic vanities, disorders, and morbidities—in a word, hallucination. One reason for such overall denigration was that complacent aliens to the mystical life did not care to distinguish between abnormal and supernormal phenomena. To them all were the same, at best some kind of religious sport. An American Quaker philosopher, Rufus Jones, has noted that psychology, as 20th-century man knows it, is empirical and possesses no ladder by which it can transcend the empirical order.

According to mystics, most men live in a prison, the walls thick with ego, the senses, and restricting interests. But some prisoners develop a passion to scale the walls and move toward an unwalled horizon, an adventure of ideas, if nothing more. Thus, the hypothesis that there might be cherubs and seraphs (angels of knowledge and of love) who call and guide men in the upward way is difficult to ignore. But if the distinction between love and knowledge is at all valid, the achievements of men would seem to be the products of love, since, as Aristotle maintained, the intellect by itself moves nothing. Without "the driving and drawing that we feel in the heart," mysticism would lack power and might sink into quiescence, as has sometimes happened. To will what the Supreme wills is the supreme secret, the *primum mobile.* "Nothing burns in hell but self-will" (*Theologia Germanica,* ch. 34). The mystic approaches this knowledge and mobility even when he is compelled to withdraw from society for long or short periods. But withdrawal without return is not complete. As scientists of the psyche, the mystics insisted on the primacy of the inner factors. Modern psychoanalysis claims to have made available to man's knowledge areas of darkness beneath the conscious levels. However revealing these evidences of the ape and the tiger, psychoanalysis is debarred from understanding the superconscious, and it is viewed by mystics as being less than correct in its reading of the irrational in man. The inescapable pessimism of the psychoanalytic conclusion stands in contrast to the possibilities of self-development and sublimation to which mystics have always pointed.

Among other discoveries on the mystical way is that of ambivalence, or the alternate ways of looking at the world: temporal as against eternal. The double vision characterizes the saint whose life forms a point of intersection between time and timelessness. Mystical psychology assumes a transcendental faculty, in the hiddenness, beyond the threshold. It is committed to a breakthrough and could never have sustained itself without constant verification. In many ways a guarded secret, meant for the competent few, the experiment has hazards and could upset any but the most disciplined. The rousing of energy, the infusion of grace, and confrontation of the levels of reality create tensions and difficulties. Hence, the insistence on moderation and balance on all hands. "The higher the love, the greater the pain," a voice had consoled a 13th-century German mystic, Mechthild von Magdeburg. "Believe me,

children," wrote a 14th-century German mystic, Johann Tauler, "one who would know much about these matters would often have to keep to his bed, for his bodily frame could not support this."

These upheavals of "mystic ill health" are part of a developing consciousness that has to move through and adjust to habits of inertia and resistance in the system and to an inability to support the emerging powers and their demands. A little imbalance now and then should take no one by surprise. The possibility of ranges of consciousness without thinking is one of the basic premises of yogic or mystical psychology. It constitutes a confutation of the formula of the 17th-century French philosopher René Descartes: *cogito ergo sum* ("I think; therefore, I am"). Being can exist without *cogito* (or *ratio,* "reason") in a direct awareness of things that is the function of intuition, *prajñā.*

Ranges of consciousness without thinking

SYSTEMATIC EXPOSITION OF MYSTICAL EXPERIENCE

Attempts of mystics to record the nature of their experiences. The theory or interpretation of mysticism is not mysticism. Generally, there are two sides to the theory: philosophical and practical. There may be another: confessional and justificatory. Though some mystics have been content to record what happened, others have worked out manuals of praxis (techniques), or *sādhāna.* As a rule, mystical method, experience, and exegesis cannot be sharply set apart from one another. However ineffable, raids on the inarticulate and expositions of the same have not ceased. The expositions have formed part of a particular framework of culture, tradition, and temperament. The 8th- to 9th-century-AD Indian philosopher Śaṅkara and the 16th-century Spanish mystic St. John of the Cross are not likely to talk in the same tone or accent. However universal in intent, all expositions tend to be localized.

The study of comparative mysticism as well as the spirit of the age make it possible and perhaps mandatory for modern man to move toward an open and untethered mysticism, the "ocean of tomorrow." Indications of this change in attitude and emphasis are not wanting, especially in the 20th-century writings of the Indian mystic philosopher Sri Aurobindo and Teilhard de Chardin, who represent something totally new but allied. R.C. Zaehner has explained that both, though unknown to each other, not only accepted the theory of evolution, but enthusiastically acclaimed it, indeed were almost obsessed with it. Both were profoundly influenced by Bergson, both were deeply dissatisfied with organized religion, and both were vitally concerned not only with individual salvation or "liberation," but also with the collective salvation of mankind.

The value and meaning of mystical experience. Among the attempts to explain the value and meaning of mystical experience, a few features may be indicated. Claimed to be a guarantee for order and reconciliation, mysticism does not take away mystery from the world, nor is it essentially irrational. Though in their penchant for the beyond or God-intoxication some mystics have inclined to reject the world, the maturer variety has not divided the world of spirit and matter but has tried to mediate between spirit and matter with the help of emanations, correspondences, and a hierarchy of the real. As a giver of life, mysticism is meant to fulfill and not to destroy. Thus, it need not be world negating.

Pointing to a scale of senses and levels of mind, mysticism provides an escape from a life of uninspired existence. It magnifies man and gives him a hope and destiny to fulfill. With its abiding sense of the "more," mysticism may be called the religion of man or the religion of maturity. It offers not irrational developments or inducements but the working out of inherent potentials. Evolution, according to mystics, is not yet ended.

The mystical life is not for those who are well adjusted and other oriented. In Ramakrishna's homely phrase, at some point or other one has to "take the plunge." A change so radical calls for a kind of attention other than what most people seem prepared to give. To make it his supreme business one must have a call to holy living. He who seeks the divine must consecrate himself to God and to God only.

Problems of communication and understanding. The problem of communication, of tidings from another country, is obvious. Transvaluation of values is not easy to accept, adjust to, or express. The dialogue between mystical and other pursuits is an unsolved problem. After he had undergone a spiritual experience, the 13th-century Christian philosopher St. Thomas Aquinas is reported to have said, "I have seen that which makes all that I have written and taught look small to me. My writing days are over." This, from the author of the voluminous *Summa theologica* ("Summary of Theology"), is not without its importance.

Even if it is difficult to describe visions and dangerous to systematize, the direction in which mysticism points is clear: relational transcendence. The 20th-century crises and the mass media suggest the possibility of a mysticism brought up to date that will serve "the Creative Intention that past ages have called God." Whether it comes through symbols, systems, paradigmatic examples, or extreme situations, there will probably always be some response to the call of the real.

MYSTICISM AS A SOCIAL FACTOR

Mystical experience is no doubt solo, the experience of a singular person. But more than "a flight of the alone to the Alone," it could also be a redemption of solitude no less than of society. In the mystic experience, as Jakob Böhme said, the world is not destroyed but remade. At times a protest against heteronomy (*i.e.,* external authority and ecclesiastical machinery), mysticism has expressed itself in diverse backgrounds and flourished during dark periods of history.

Because of its other-worldly bias, the belief still persists that the solitary mystic, absorbed in a vertical relation with God or reality, owes no social responsibility. Altogether an outsider, he has deliberately undergone a civil death. This is not an ideal or wholly accurate picture. "A Mystic who is not of supreme service to the Society is not a Mystic at all" (from preface to R.D. Ranade, *Mysticism in Maharashtra*). According to Zen Buddhism, the great contemplative—even when "sitting quietly, doing nothing"—has been a man of action, perhaps the only kind of action that leaves no bitter residue behind. The less extravagant forms of mysticism represent attitudes and principles of charity, detachment, and dedication, which should guide the relation of the individual to the group. The mystics have fought the inner battle and won, creating themselves and their world.

The mystic as a man of action "doing nothing"

Mysticism proves the individual's capacity to rise above the conditioning factors of nature, nurture, and society and to transform collective life, though this has not been generally recognized. With a hidden and potent force, mystics have tried, as best as circumstances permitted, to mend the universal ill. As in the classic resolve of the *bodhisattva* ("buddha-to-be"), they have looked forward to universal enlightenment. If the attempt by mystics to create a new order or a better society has failed, the incapacity or defection of the majority may be the reason for the failure.

"Revolution" is a word too often profaned. The change suggested is mainly, if not wholly, from without. In such contrived salvation by compulsion, the inner core is hardly touched. "But it is an eternal law that there can be no compulsion in the realm of the spirit. It is essentially a world of free creative choices" (Rufus Jones, *Some Exponents of Mystical Religion*). Mystics insist on a change of consciousness, a slower and more difficult process, and also on a scrupulous equation between ends and means. Impatience, deviations, and subterfuges in this respect can be costly, ironic, and instructive. According to mystics, the individuals who will most help the future of humanity will be those who recognize the unfinished and ultimate revolution—the evolution of consciousness—as the destiny and therefore the great need of all men as of society.

Holiness does not mean a retreat from or a rejection of the world. To be a mystic or a seer is not the same thing as being a spectator on the fence. As the Swedish secretary-general of the United Nations, Dag Hammarskjöld, proved with his life, in the modern era the road to holiness

necessarily passes through the world of action. Many with a mystical frame of mind look beyond what mystics call quasi-revolutions to a great life—an entire civilization, the civilization of consciousness. The need of synthesis places its stake on the future and the All.

> The outcome of the world, the gates of the future, the entry into the super-human—these are not thrown open to a few of the privileged nor to one chosen people to the exclusion of all others. They will open only to an advance of *all together.* (From Teilhard de Chardin, *The Phenomenon of Man.*)

According to mystics, here may be the outline of a revolution whose message has reached but a few. The hope of a Kingdom of Heaven within man and a City of God without remains one of mysticism's gifts to what many mystics view as an evolving humanity. (S.Gh.)

BIBLIOGRAPHY

The nature of religious experience: WILLIAM JAMES, *The Varieties of Religious Experience* (1902), a classic philosophical and psychological study; RUDOLF OTTO, *Das Heilige,* 9th ed. (1922; Eng. trans., *The Idea of the Holy,* 1923; 2nd ed., 1950), a study of the nonrational in religious experience; J.M. MOORE, *Theories of Religious Experience, with Special Reference to James, Otto and Bergson* (1938); JAMES A. MARTIN, JR., *Empirical Philosophies of Religion* (1945); and C.C.J. WEBB, *Religious Experience* (1945), contain valuable appraisals and good bibliography; H.E. BRUNNER, *Wahrheit als Begegnung* (1937; Eng. trans., *The Divine-Human Encounter,* 1943); and MARTIN BUBER, *Ich und Du* (1923; Eng. trans., *I and Thou,* 1937 and 1970), express the view that authentic religion is based on personal encounter between man and God; JOHN DEWEY, *A Common Faith* (1934), argues for the "religious" in experience; E.S. BRIGHTMAN, *A Philosophy of Religion* (1940), represents the Personalist view that personhood is the most basic quality of reality; W.E. HOCKING, *The Meaning of God in Human Experience* (1912, reprinted 1963); J.E. SMITH, *Experience and God* (1968), emphasize the experiential basis of the question of God; and ALISTER HARDY, *The Spiritual Nature of Man: A Study of Contemporary Religious Experience* (1979), a collection of 3,000 personal reports.

Religious experience and other experience: H.D. LEWIS, *Our Experience of God* (1959); and J.E. SMITH, *Religion and Empiricism* (1967), deal with the bearing of different conceptions of experience on religion; JOSIAH ROYCE, *The Sources of Religious Insight* (1912); W.G. DE BURGH, *From Morality to Religion* (1938); and PAUL TILLICH, *Morality and Beyond* (1963), treat the relation between religion and morality; GERARDUS VAN DER LEEUW, *Vom Heiligen in der Kunst* (1957; Eng. trans., *Sacred and Profane Beauty,* 1963), treats the relation between art and religion.

The structure of religious experience: JOHN MACMURRAY, *The Structure of Religious Experience* (1936, reprinted 1971); J.B. PRATT, *The Religious Consciousness* (1920); PAUL TILLICH, *The Dynamics of Faith* (1957); and A.N. WHITEHEAD, *Religion in the Making* (1926), deal with psychological, theological, and metaphysical aspects; JOACHIM WACH, *The Sociology of Religion* (1944), is an indispensable study of the social expression of religious experience; WILLIAM A. CHRISTIAN, *Meaning and Truth in Religion* (1964); F. FERRE, *Basic Modern Philosophy of Religion* (1967); NINIAN SMART, *Philosophers and Religious Truth* (1964); and J.E. SMITH, *Reason and God* (1961), deal with the issue of the cognitive import of religious experience; J.H. LEUBA, *The Psychology of Religious Mysticism* (1925), argues against its cognitive import; W.E. HOCKING, *Science and the Idea of God* (1944); W.T. STACE, *Religion and the Modern Mind* (1960); and H.N. WIEMAN, *The Wrestle of Religion with Truth* (1927), discuss the relation between religion and science; J. MACQUARRIE, *God-Talk* (1967); I.T. RAMSEY, *Christian Discourse* (1965) and *Models and Mystery* (1964), represent the linguistic approach to religious experience; MIRCEA ELIADE, *Le Mythe de l'éternel retour* (1949; Eng. trans., *Cosmos and History: The Myth of the Eternal Return,* 1954; rev. ed., 1965) and *The Sacred and the Profane: The Nature of Religion* (1959), interpret religious experience through myth, symbol, and ritual.

Situational contexts and forms of expression: EVELYN UNDERHILL, *Worship* (1936, reprinted 1957), invaluable for the meaning of worship and its forms; P. EDWALL *et al.* (eds.), *Ways of Worship* (1951), treats the liturgies of the major Christian communities; EMILE DURKHEIM, *Les Formes élémentaires de la vie religieuse* (1912; Eng. trans., *The Elementary Forms of the Religious Life,* 1915, paperback 1961), presents the "group theory" of religion; C.C.J. WEBB, *Group Theories of Religion and the Individual* (1916), a critique of Durkheim; MIRCEA ELIADE, *Birth and Rebirth* (1958); ARNOLD VAN GENNEP, *Les Rites de passage* (1909; Eng. trans. 1960); and GERARDUS VAN DER LEEUW, *Phänomenologie der Religion* (1933; Eng. trans., *Religion in Essence and Manifestation,* 1938), on initiation rites and the cycle of sacred life; JOACHIM WACH, *The Sociology of Religion* (1944), the best source for the relation between religious and nonreligious groupings.

Types of religious experience and personality: GERARDUS VAN DER LEEUW, *Religion in Essence and Manifestation (op. cit.);* and JOACHIM WACH, *The Sociology of Religion* (1944) and *Types of Religious Experience* (1951), invaluable for the analysis of religious roles and personalities; ALFRED GUILLAUME, *Prophecy and Divination Among the Hebrews and Other Semites* (1938); RUDOLF OTTO, *Religious Essays* (1931); and JOHN SKINNER, *Prophecy and Religion* (1922; paperback ed., 1961), deal with the meaning of prophecy in the Semitic traditions.

Mystical experience: RICHARD M. BUCKE, *Cosmic Consciousness: A Study in the Evolution of the Human Mind* (1905, many later editions), introduced two important ideas as one, ideas that would recur, with modifications, in later writings. RUFUS M. JONES, *Studies in Mystical Religion* (1908, reprinted 1970), provided a balanced and liberal attitude that emphasized its experiential quality, its value as a practical guide, and the presence of a mystical brotherhood through the centuries. EVELYN UNDERHILL, *Mysticism,* 12th ed. rev. (1957), has been a pioneer work, though its insistence on the Mystic Way has been questioned. REYNOLD A. NICHOLSON, *The Mystics of Islam* (1913, reprinted 1966), is one of the earliest studies in Ṣūfism that still holds interest. SRI AUROBINDO, *The Synthesis of Yoga* (first published, serially, 1914–21, later in book form), with much collateral illumination, explains the idea of Integral Yoga. HENRI BERGSON, *Les Deux sources de la morale et de la religion,* 3rd ed. (1932; Eng. trans., *Two Sources of Morality and Religion,* 1935), forms part of a general thesis on creative evolution and, paradoxically, on the need for mysticism in an age of mechanization. GERSHOM G. SCHOLEM, *Major Trends in Jewish Mysticism,* 3rd rev. ed. (1954), clearly brings out the distinction that the concept of union is not an essential of mystical experience as understood in the Jewish tradition. ALDOUS HUXLEY, *The Perennial Philosophy* (1946), is an anthology with sophisticated, sometimes cynical, commentary with an ascetic bias. JACQUES DE MARQUETTE, *Introduction to Comparative Mysticism* (1949), is a fair and straightforward survey in which its relevance to modern life and thought is brought out and an awareness of possibilities hinted at. R.C. ZAEHNER, *Mysticism, Sacred and Profane* (1957), beginning as a caveat against the use of drugs for transcendental experience, goes on to make critical distinctions between four types of mysticism. D.T. SUZUKI, *Mysticism: Christian and Buddhist* (1957), offers a sympathetic study of contrasts as well as some resemblances between two traditions. RADHAKAMAL MUKERJEE, *The Theory and Art of Mysticism* (1960), is an overall study, particularly good with regard to the Eastern material. WALTER T. STACE, *Mysticism and Philosophy* (1960), is balanced and analytic but singles out introvertive mysticism as more genuine and superior. SIDNEY SPENCER, *Mysticism in World Religion* (1963), is a helpful anthology with a reliable introduction to the field of comparative mysticism. PIERRE TEILHARD DE CHARDIN, *Le Phénomène humain* (1955; Eng. trans., *The Phenomenon of Man,* 1959), though its scientific accuracy has been questioned, its poetic and impassioned attempt to mediate between religious insights and a hope for man and the future has made it the object of much attention. See also LOUIS DUPRÉ, *The Deeper Life: An Introduction to Christian Mysticism* (1981); and RICHARD WOODS, *Mysterion: An Approach to Mystical Spirituality* (1981).

Religious Symbolism and Iconography

Symbolism, the basic and often complex artistic forms and gestures used as a kind of key to convey religious concepts, and iconography, the visual, auditory, and kinetic representations of religious ideas and events, have been utilized by all the religions of the world.

In the 20th century the symbolical character of religion has often been stressed over attempts to present religion rationally. The symbolic aspect of religion is even considered by some scholars of psychology and mythology as the main characteristic of religious expression. Scholars of comparative religions, ethnologists, and psychologists have gathered and interpreted a great abundance of material on the symbolical aspects of religion, especially in relation to Eastern and primitive religions. In recent Christian theology and liturgical practices another revaluation of religious symbolical elements has occurred.

The importance of symbolical expression and of the pictorial presentation of religious facts and ideas has been confirmed, widened, and deepened both by the study of primitive cultures and religions and by the comparative study of world religions. Systems of symbols and pictures that are constituted in a certain ordered and determined relationship to the form, content, and intention of presentation are believed to be among the most important means of knowing and expressing religious facts. Such systems also contribute to the maintenance and strengthening of the relationships between man and the realm of the sacred or holy (the transcendent, spiritual dimension). The symbol is, in effect, the mediator, presence, and real (or intelligible) representation of the holy in certain conventional and standardized forms.

This article is divided into the following sections:

The nature of religious symbols and symbolization 591
- Concepts of symbolization 591
- Varieties and meanings associated with the term symbol 592
- The symbolic process 592
- Symbols in the religious consciousness 592

The relation of the symbol and the sacred 593
- Symbols as the incarnate presence of the sacred or holy 593
- Symbols as indicators of the sacred or holy 593
- Symbols of sacred time and space 593
- Ceremonial and ritualistic objects as indicators or bearers of the sacred or holy 593
- Other relations between the symbol and the sacred 594

Relation of religious symbolism and iconography to other aspects of religion and culture 594
- Relation to myth and ritual 594
- Relation to meditation and mysticism 594
- Relation to the social realm 594
- Relation to the literary and visual arts 594
- Relation to other areas of culture 594
- Changes in symbolical relations and meanings 594

Modes of symbolic expression 595
- Diagrammatic and emblematic 595
- Pictorial 595
- Gestural and physical movements 595
- Verbal symbolism 596
- Musical symbolism 596
- Conjunction and combination of various modes 596

Icons and systems of iconography 596
- Iconographic forms 596
- Iconographic themes 597
- Iconographic types 597

Influence of man's environment on religious symbolism and iconography 598
- Influences from nature 598
- Influence of human relationships 598
- Symbolism of sex and the life cycle 599
- Cultural influences 599
- Conceptual influences 599

Influence of religion on symbolism and iconography 599
Conclusion 599
Bibliography 599

THE NATURE OF RELIGIOUS SYMBOLS AND SYMBOLIZATION

Nature and function of symbols

The word symbol comes from the Greek *symbolon*, which means contract, token, insignia, and a means of identification. Parties to a contract, allies, guests, and their host could identify each other with the help of the parts of the *symbolon*. In its original meaning the symbol represented and communicated a coherent greater whole by means of a part. The part, as a sort of certificate, guaranteed the presence of the whole and, as a concise meaningful formula, indicated the larger context. The symbol is based, therefore, on the principle of complementation. The symbol object, picture, sign, word, and gesture require the association of certain conscious ideas in order to fully express what is meant by them. To this extent it has both an esoteric and an exoteric, or a veiling and a revealing, function. The discovery of its meaning presupposes a certain amount of active cooperation. As a rule, it is based on the convention of a group that agrees upon its meaning.

Concepts of symbolization. In the historical development and present use of the concepts of symbolization, a variety of categories and relationships must necessarily be differentiated. Religious symbols are used to convey concepts concerned with man's relationship to the sacred or holy (*e.g.*, the cross in Christianity) and also to his social and material world (*e.g.*, the *dharma-cakra*, or wheel of the law, of Buddhism). Other nonreligious types of symbols have achieved increasing significance in the 19th and 20th centuries, especially those dealing with man's relationship to and conceptualization of the material world. Rational, scientific–technical symbols have assumed an ever increasing importance in modern science and technology. They serve partly to codify and partly to indicate, abbreviate, and make intelligible the various mathematical (*e.g.*, $=$, equality; $\equiv$, identity; $\sim$, similarity; $\parallel$, parallel; or $<$, less than), physical (*e.g.*, $\sim$, alternating current), chemical (*e.g.*,  , benzene ring), biological (*e.g.*, ♂, male; ♀, female), and other scientific and technical relationships and functions. This type of "secularized" symbol is rooted, to a degree, in the realm of religious symbolism. It functions in a manner similar to that of the religious symbol by associating a particular meaning with a particular sign. The rationalization of symbols and symbolical complexes as well as the rationalization of myth have been in evidence at least since the Renaissance.

The concept of the religious symbol also embraces an abundantly wide variety of types and meanings. Allegory, personifications, figures, analogies, metaphors, parables, pictures (or, more exactly, pictorial representations of ideas), signs, emblems as individually conceived, artificial

symbols with an added verbal meaning, and attributes as a mark used to distinguish certain persons all are formal, historical, literary, and artificial categories of the symbolical. If one looks for a definable common denominator for the various types of symbols, one could perhaps choose the term "meaning picture" or "meaning sign" to best describe the revealing and at the same time the concealing aspects of religious experience. The symbol (religious and other) is intended primarily for the circle of the initiated and involves the acknowledgment of the experience that it expresses. The symbol is not, however, kept hidden in meaning; to some extent, it even has a revelatory character (*i.e.,* it goes beyond the obvious meaning for those who contemplate its depths). It indicates the need for communication and yet conceals the details and innermost aspects of its contents.

Varieties and meanings associated with the term symbol. Different forms and levels of the experience of and relationship to reality (both sacred and profane) are linked with the concepts of symbol, sign, and picture. The function of the symbol is to represent a reality or a truth and to reveal them either instantaneously or gradually. The relationship of the symbol to a reality is conceived of as somewhat direct and intimate and also as somewhat indirect and distant. The symbol is sometimes identified with the reality that it represents and sometimes regarded as a pure transparency of it. As a "sign" or "picture" the representation of the experience of and relationship to reality has either a denotative or a truly representative meaning. The doctrine of the eucharistic (sacramental) presence of Christ in the teachings of Eastern Orthodoxy, Roman Catholicism, and the Protestant Reformers concretely demonstrate the various and extensive levels of symbolical understandings. These levels extend all the way from the concept of physical identity in the transubstantiation theory of Roman Catholicism (in which the substance of breadness and wineness is believed to be changed into the body and blood of Christ, though the properties of the elements remain the same) through Luther's Real Presence theory (in which Christ is viewed as present, though the question of how is not answered because the question of why he is present is considered more important), and Zwingli's sign (symbolic or memorial) theory, to the concept of mere allusion. The concept of the symbol, however, includes all these interpretations.

Denotative and cognitive aspects

Furthermore, a symbol in its intermediary function has aspects of epistemology (theory of knowing) and ontology (theory of being). As a means of knowledge, it operates in a characteristically dialectical process of veiling and revealing truths. It fulfills an interpretative function in the process of effectively apprehending and comprehending religious experience. In doing so, the word, or symbol—with its meaning, contextual use, relationship to other types of religious expression, and interpretative connection with the various forms of sign, picture, gesture, and sound—plays an important part in the process of symbolical perception and reflection. Although the symbol is an abbreviation, as a means of communication it brings about—through its connection with the object of religion and with the world of the transcendent—not only an interpretative knowledge of the world and a conferral or comparison of meaning to life but also a means of access to the sacred reality. It may possibly even lead to a fusion, or union of some sort, with the divine. To this extent, the sacrament of the Lord's Supper, the liturgical and ritualistic mystery in Christianity—with its many symbolical signs, pictorial representations, significative actions, interpretative words, and various levels of approach to the divine reality—is an example of a highly developed form of a complex symbolic action. Here, the concept of analogy is important; the symbol functions in these ways because it has an analogous cognitional as well as existential relationship to that which it signifies.

The symbolic process. To trace the origin, development, and differentiation of a symbol is a complicated process. Almost every symbol and picture in religion is at first either directly or indirectly connected with the sense impressions and objects of man's environment. Many are derived from the objects of nature, and others are artificially constructed in a process of intuitive perception, emotional experience, or rational reflection. In most cases, the constructions are again related to objects in the world of sense perception. A tendency toward simplification, abbreviation into signs, and abstraction from sense objects is quite evident, as well as a tendency to concentrate several processes into a single symbol. A good example of this last tendency may be seen in ancient Christian portrayals of the triumphant cross before a background of a star-filled heaven that appear in the apses of many basilican churches. In these representations the Crucifixion, Resurrection, Ascension, exaltation, and Transfiguration of Christ are joined to apocalyptic concepts (centring on sudden interventions by God into history) inherent in the doctrine of the Last Judgment. An excellent example of such an apse mosaic is to be found in the S. Apollinare in Classe, near Ravenna (in Italy). On the other hand, there is a tendency to accumulate, combine, multiply, and differentiate symbolical statements for the same thought or circumstance, as seen, for example, on the sarcophagi (stone coffins) of late Christian antiquity—especially in Ravenna. Here, the same idea is symbolically expressed in various manners; *e.g.,* by means of persons, objects, animals, and signs, all appearing side by side.

Development and differentiation of a symbol

The forms and figures of symbolical thought can change into exaggerations and rank growths, however, and lead to transformations and hybrids—figures with several heads, faces, or hands—as exemplified in the statues and pictorial representations of the deities of India (*e.g.,* the multi-armed goddess Kālī) and of Slavic tribes (*e.g.,* the four-headed Suantevitus). The meaning of individual symbols can change and even be perverted. The lamb that in ancient Christian art symbolizes Christ may also symbolize the Apostles or mankind in general. The dove may symbolize the Holy Spirit or the human soul. The wheel or circle can symbolize the universe, the sun, or even the underworld. The encyclopaedic Christian allegorism (symbolism) of the Middle Ages offers many interesting examples, as noted in the writings of Isidore of Seville, a 6th- to 7th-century Spanish theologian, and Rabanus Maurus, a 9th-century German abbot and encyclopaedist.

The foundations of the symbolization process lie in the areas of the conscious and the unconscious, of experience and thought, and of sense perception, intuition, and imagination. From these arises the structure of religious symbolism. Sensation and physiological and psychological processes participate in the formation of the symbol structure. Extraordinary religious experiences and conditions, visions, ecstasy, and religious delirium brought about by intoxication, hallucinogenics, or drugs that produce euphoria and changes in consciousness must also be taken into consideration. The symbol itself, however, is intended as an objective concentration of experiences of the transcendent world and not as a subjective construction of a personally creative process. In cultic and mystical visions and trances, the forms and processes of the external world and of the religious tradition are condensed and combined with mythical images and historical events and take on a life of their own. The process of rational conceptualization and structuralization, however, also plays a part in the origin and development of many symbols. There is a correlation between sense perception, imagination, and the work of the intellect.

The foundations of the symbolization process

Symbols in the religious consciousness. The formation of religious symbols that occur when unconscious ideas are aroused or when a process of consciousness occurs is principally a matter of religious experience. Such symbols usually become intellectual acquisitions, and, as religious concepts are further elaborated upon, the symbols may even finally become subjects of major theological questions. In Christian theology, for example, summaries of dogmatic statements of faith are called symbols (*e.g.,* the Apostles', Nicene, and Athanasian creeds and the confessional books of Protestantism, such as the Augsburg Confession of Lutheranism). This particular use of the term symbol is exceptional, however.

In the development of the symbol, religious experience, understanding, and logic are all connected, but each places different accents on the individual categories and species

Religion as both origin and product of symbols

of symbol. Occasionally, religion is regarded as the origin and the product of certain established (or fundamental) symbols. In such cases the outcome of the process of the structuralization of religious consciousness would then be the establishment of a symbol that is generally applicable to a particular historical species of religion. Conversely, one could ask whether the experience and establishment of an individual or collective symbol by a creative personality or a community is not itself the establishment of a religion. If so, the classical symbol that was developed at the time of the foundation of any one particular religion would then be constitutive for its origin and further development (*e.g.*, the T'ai Chi or the combination of Yin and Yang for the Chinese, the cross for the Christian religion). In any event, the symbol belongs to the essence of man's coming of religious consciousness and to the formation of history's institutional religions. It plays a fundamental and continual part in the further growing of such religions and in the mental horizons of their followers.

THE RELATION OF THE SYMBOL AND THE SACRED

Symbols as the incarnate presence of the sacred or holy. Whatever the experience of reality that lies behind the religious symbol may be, it is above all the experience of the sacred or holy, which belongs essentially to any concept of religion. The historical study of religions has shown that it is fundamentally the symbol that mediates and forms for man's religious consciousness the reality and the claim of the holy. Religion is a system of relationships, a system of reciprocal challenges and responses the principal correspondents of which are the sacred or holy and man. Though there are many forms of experience in which the sacred or holy is distinctly known and felt, the experience is often acquired in worship, in which this system of relationships is realized and continually renewed and in which the sacred or holy supposedly makes itself present. The details of worship serve to objectify and regulate in a perceptual and material manner the presupposed presence of the sacred or holy, of which the symbol and the picture are intended to be its materialization. In its material manifestation the sacred or holy is adapted to man's perceptual and conceptual faculties. Viewed from the aspect of its holiness, the symbol originates in a process of mediation and revelation, and every encounter with it is supposed to bring about a renewed actualization and a continual remembrance of this revelation.

Incarnation and identity

The actualization of the presence of the holy by means of symbolic representation can, in extreme cases, lead to an identification of the physical manifestations with the spiritual power symbolized in them. The symbol, or at least an aspect of it, is then viewed as the incarnated presence of the holy. The sacred stone, animal, plant, and drum and the totem symbol or the picture of ancestors all represent the sacred or holy and guarantee its presence and efficacy. The origin of many such symbols clearly indicates the identity that was presumed to have existed between the symbol and the sacred or holy. The Greek god Dionysus as a bull, the Greek goddess Demeter as an ear of corn, the Roman god Jupiter as a stone, the Syrian god Tammuz-Adonis as a plant, and the Egyptian god Horus as a falcon all are viewed as manifestations of the deities that were originally identified with these respective objects of nature.

Symbols as indicators of the sacred or holy. The symbol is understood to have a referential character. It refers to the reality of the sacred or holy that is somewhat and somehow present. When the symbol is an indicator of the sacred or holy, a certain distance exists between them, and there is no claim that the two are identical. Short of actual identification, various degrees of intensity exist between the symbol and the spiritual reality of the sacred or holy. The symbol is a transparency, a signal, and a sign leading to the sacred or holy. The objects, gestures, formulas, and words used in meditation—for example, the Buddhist *mudrās* (gestures), *pratimās* (images), *mantras* (magic formulas)—and in mysticism—for example, the crystal or the shoemaker's ball in the contemplative experience of Jacob Böhme, a 16th- to 17th-century German cobbler and mystic; the navel in Omphalicism (a method [called Hesychasm] of contemplating the navel in order to experience the divine light and glory in medieval Greek Christian mysticism of the monks of Mt. Athos); and the pictures of the deity in the language of Hindu, Islāmic, and medieval Christian mysticism—all of these are truly symbols, but nonetheless they have at most only an indirect mediating relationship to the divine, a purely noetic (intellectually abstract) significance with regard to the reality and presence of the sacred or holy.

Symbols of sacred time and space. The symbolical forms of representation of the sacred or holy are to be understood as references to or transparencies of the sacred or holy. The sacred manifests itself in time and space, so that time and space themselves become diaphanous indications of the holy. The holy place—a shrine, forest grove, temple, church, or other area of worship—is symbolically marked off as a sacred area. The signs, such as a stake, post, or pillar, that delimit the area themselves are endowed with sacred symbolic meanings, which often can be noted by their particular designs. The ground plan of the sacred building and its orientation, walls, roof, and arches are all utilized to symbolize the sacred or holy. Prehistoric places of worship—*e.g.*, Stonehenge (in England) and other megaliths of Europe and the shrines and holy places of ancient Egypt, Babylon, China, and Mexico—were invested with symbolical meanings.

Sacred places are often pictorial reflections of the universe and its design and partake of its holiness. The domes of Christian churches are symbols of heaven, the altar a symbol of Christ, the Holy of Holies of the Temple in Jerusalem a symbol of Yahweh, the Holy of Holies in Shintō shrines (*honden*) a symbol of the divinity, and the prayer niches in mosques a symbol of the presence of Allāh. In many instances shoes may not be worn on holy ground (*e.g.*, Shintō temples), and hands and feet are to be washed before entering into a holy place. The woodwork of demolished Shintō shrines, when taken to private homes, makes the sacred or holy present in the homes of pious Japanese families.

The cycle of the sacred year

Time as a transparent symbol of the sacred may be represented by means of the cycle of the sacred year and its high points—*e.g.*, New Year's (as in ancient Near Eastern religions), the times of sowing and reaping, and the solstices and equinoxes. Or the lapse of time may be represented in signs and pictures. Cosmic, mythical, and liturgical time and destiny are portrayed, for example, in the Buddhist symbol of the wheel of life, *bhava-cakra*, with its causal chain of human deeds and succession of existences, entwined by the claws of a devouring monster; the figures of Aion (Time) in late Greco-Roman and Persian antiquity show a figure with a winged lion's head standing on a globe and encircled by a snake. Time itself, its course, division, and fixed points, is both an allusion and the bearer and mediator of the sacred or holy.

Ceremonial and ritualistic objects as indicators or bearers of the sacred or holy. Liturgical and ceremonial objects can also indicate or lead to the sacred or holy. Not only holy pictures and symbols (*e.g.*, the cross in Christianity or the mirror in Japanese Shintō) but also lights, candles, lamps, vessels for holy materials, liturgical books, holy writings, vestments, and sacred ornaments are indicators of the sacred or holy. Liturgical vestments and masks are intended to transform the wearer, to remove him from the realm of the this-worldly, and to adapt him to the sphere of the sacred or holy; they help him to come into contact with the divine—for example, by obscuring his sexual characteristics. The vestments may be covered with symbols, such as those worn by Arctic shamans (medicine men with psychic transformation abilities). They are signs of the function of the wearer and his relationships to the sacred or holy and to the profane world. Such vestments are frequently derived from those of rulers or from ceremonial court dress; *e.g.*, Japanese Shintō and Roman Catholic and Eastern Orthodox Christianity. They are supposed to create a fitting atmosphere of solemnity and dignity. In Western Christianity, the liturgical vestments have a very specific symbolism: the alb (a tunic) symbolizes purity of heart; the stole, the raiment of immortality; and the chasuble (an outer eucharistic, or holy communion,

vestment), the yoke of Christ. The liturgical vestments of the Eastern Christian churches have a similar symbolism. The ritual headdress and the crown express the sacred dignity of the wearer. The vestments of the various religious orders (Oriental and Occidental) express the holiness of the members of the community, their nearness to the sacred or holy, and the significance of religious life for them. In the reception ritual of Jainism and Buddhism, the monastic vestments are put on as a sign of an entrance in a new state of life. This ritual in Jainism resembles that of a wedding ceremony. The taking over of the monastic garb is an essential part of becoming a *sādhu*. The monks of the Jainistic Śvetāmbara sect wear five objects (*e.g.*, shells) as symbols of the five monastic virtues. In early Christianity the white baptismal vestment was a symbol of rebirth, new life, and innocence.

Other relations between the symbol and the sacred. The sacred or holy as represented or manifested in the symbol has, generally speaking, a sanctifying function (elevating one to a closer relationship to the sacred or holy) and an exorcising function (decreasing or eliminating those aspects that hinder one's relationship to the sacred or holy). Remembrance (*anamnēsis*) and imitation (*mimēsis*) are the analogous and associative means of representing the reality and indestructibility of the sacred or holy and its power, which defends, protects from injury, bans evil, and guarantees salvation. Symbolic signs and pictures (*e.g.*, masks; sex, animal, or plant symbols, such as the skulls or horns of animals) are placed on houses and sacred places to make present the saving and sanctifying power of the sacred or holy.

RELATION OF RELIGIOUS SYMBOLISM AND ICONOGRAPHY TO OTHER ASPECTS OF RELIGION AND CULTURE

Relation to myth and ritual. The symbol has a long-established relationship with myth (sacred stories that define the human condition and man's relation to the sacred or holy). Often containing a collection of symbolic forms, actions, expressions, and objects, myths describe gods, demons, men, animals, plants, and material objects that are themselves bearers of symbolical meanings and intentions. Thus, it is sometimes difficult to distinguish between a myth and a coherent complex of symbols brought together in story form. Examples are myths of cosmogony (origin of the world), theogony (origin of the gods), and anthropogony (origin of man). The details and contexts of religious teaching, dogma, and theology also produce or form symbolic values or refer to traditional symbolic representations. Symbol structures and pictorial representations are brought into connection with dogma and theological statements—*e.g.*, the Buddhistic *karma-saṃsāra* (law of cause and effect and reincarnation) theory and the *bodhisattva* (buddha-to-be) theory or the Christian teaching of the Last Judgment, punishment of sin, hell and purgatory, and eternal reward (Paradise). In worship, individual actions and objects used in the ritual are given a symbolic meaning that transcends their immediate practical purpose. Magic, in its ritual, also uses various formations of symbols, pictures, and symbolical actions that may be seen as parallels to the distinctively religious use of symbols.

Relationship of symbolism and doctrine

Relation to meditation and mysticism. The spiritualization of religious experience in forms of meditation and mysticism assimilate and rework the existing symbols and pictures of an older historical period of religion, giving to some symbols a higher value and placing others in the centre of focus. At the same time it develops new forms the appearances of which stem especially from the visionary experiences of the mystic and from his need for a suitable means of expression and from the objects of meditation training; *e.g.*, holy sounds and words (*om*), the lotus flower, the *vajra* (ritual object shaped like a thunderbolt), and the wheel in Buddhist meditations or the ladder, the heart, and the letters IHS (the first three letters of the Greek word for Jesus) in Christian mysticism. In contemplation, colours, forms, sounds, signs, and pictures become ways and means of penetrating to the centre of the mystical union. Jacob Böhme's work is characteristic of the development of an especially rich mystical language of symbols. Mysticism supplies conventional and customary religiosity with new pictures and symbols.

Relation to the social realm. The field of symbolism and iconography shows a strong interdependence that existed between religion and other areas of culture that were later to become autonomous and profane (or secularistic). The social domain under the influence of religion develops its own symbolism for expressing its values and objectives. Conversely, religion often draws its symbols and pictorial forms from the social, political, and economic domains. Persons (*e.g.*, king, father, mother, child, slave, brother) and conditions and structures in society and the state (*e.g.*, government, a people, family, marriage, occupation) all receive meaning as symbolical and pictorial motifs in myth and cult. Examples of such motifs are throne, crown, sceptre, standard, arms, instruments, the figures of the father, mother, and child, and symbols of familial relationships. The morals, law, administration of justice, and the customs and habits of a society contain religious symbols and symbolical actions, as in the anointing of a king and in the administering of the oath or ordeal or in the observance of traditions and customs associated with birth, marriage, and death.

Reciprocal influences of culture and religion

Relation to the literary and visual arts. Religious symbols and pictures may be identical with, related to, or similar to those of language (metaphors) and to pictorial expressions in prose and poetry. They are related in allegory, parable, fairy tales, fables, and legends in which they can appear in a form that is closely related to that of religious symbolism. Religious symbols are used in the plastic arts, in architecture, and in music. Symbols also have been developed in those arts and then introduced into religion. A few examples of such symbols are house, room, door, column, sound, harmony, and melody (as when Christ was viewed as the "new melody" in the words of Clement of Alexandria, a 2nd-century philosophical theologian). Here, also, the interdependence and the continual reciprocal influence of religion and culture may be observed.

Relation to other areas of culture. The formation of religious symbols and pictures has been stimulated by numerous other areas of human culture—such as the philosophy of nature, the natural sciences (especially botany and zoology), alchemy, and medicine (including anatomy, physiology, pathology, and psychiatry). In the works of Jacob Böhme, alchemy (*e.g.*, the elements, fire, salt, sulfur, mercury, tincture, gold, essence, the philosopher's stone, and the transmutation) found an all-inclusive symbolical use; and in the works of Robert Fludd, an English physician and mystical philosopher of the 16th and 17th centuries, medical, cosmological, alchemical, and theosophical (esoteric religious) symbols were fused together (*e.g.*, the contrast of light and darkness and the idea of man as a microcosm). Symbols, also religious and mythological (such as signs of astral gods for the planets in astronomy), have achieved new importance in the conceptual presentations of distinctively scientific systems; *e.g.*, in physics, cosmology, psychiatry, and psychology. Even spaceships bear symbolic or mythical names. Psychoanalysis and depth psychology have reevaluated the role of the religious symbols and have used them in interpreting psychological processes, such as in the works of the Swiss psychiatrist Carl Jung. Jung interprets religious processes as symbolic ones and emphasizes the growing of individual and social symbols in the unconscious. According to his interpretations, many of the symbols, transforming the archaic libido into other functions, come out of dream experiences in a kind of intuition or revelation. Important symbols are duality (male–female, animus–anima), trinity, and quaternity.

Relationship of religious symbolism to scientific symbols

Changes in symbolical relations and meanings. Symbols emerge and disappear and change in their value and function. Although symbols have a tendency to be normative, stable, and to have a fixed meaning value, the demise of old symbols and the genesis of new ones or changes in the meaning of existing symbols nevertheless occur. Many ancient Christian symbols (*e.g.*, the fish) had long lost their recognition value or had been pushed into the background. With the renewal of ancient Christian

symbolism in modern times, they have had a re-evaluation. The triangle and the eye as recently used in Christianity are relatively new symbols for God. The old and formerly very meaningful religious symbolism of the axe and hammer has almost disappeared. The symbolism of kingship and sovereign authority has, on the other hand, been maintained in religious language and in the religious conceptual framework, although the political structures from which it originated have disappeared or lost their relevance. The disintegration of individual symbols and the change in the emphasis on the role of symbolism in general are partly consequences of cultural, intellectual, social, and economic transformations.

MODES OF SYMBOLIC EXPRESSION

In the long history of the forms of symbolical expression a narrower (exclusive) and broader (inclusive) idea of what a symbol is has gradually evolved. This evolution is reflected in the various manners of symbolical expression that may influence and combine with one another. Many scholars question whether a picture or a verbal expression, for example, strictly corresponds to the idea of a symbol. Just as the ideology and terminology of the ancient Greek mystery (salvatory) religions distinguished between that which is shown and seen (*deiknymenon*), that which is done (*drōmenon*), and that which is said (*legomenon*), so also can one make distinctions among three types of symbols: the visual symbol, the symbolical action that is dramatically enacted in worship, and the linguistic symbol, which includes music and other sounds. Viewed in these various aspects, the complex character of the symbol becomes apparent.

Diagrammatic and emblematic. Symbolic representations are usually depicted in diagrammatic or ideographic modes as signs, abbreviations, images, and objects of all kinds that indicate a larger context. In this category belong the simplified or abstract forms of objects of nature or other objects and geometrical forms, as well as colours, letters, and numbers. The circle, the disk, the rosette, or the swastika, for example, may symbolize the sun, universe, or a star. The square and the cross may symbolize the Earth or the four cardinal points; the wreath, the labyrinth, the spiral, the plait, and the knot may indicate eternity, the flow of time, or a magical spell.

Geometric, numerical, chromatic, and alphabetical symbolism

Ornamental designs in primitive artwork, those of the American Indians, for example, frequently have a symbolical meaning and embody fundamental figures, such as the straight line, circle, rectangle, rhombus, or ellipse. The cross in its varying forms: the Latin (†), Tau (T), ankh (☥), Saint Andrew's (X), and the forked (Y) may symbolize man and his extremities. Among various peoples and in different religions a number of basic colours have at times different and sometimes even opposite meanings. White, for example, may signify joy and festivity or death and sadness. Red has the most pronounced symbolical value: it refers to the liturgical, priestly sphere and also to life and death. In Christianity, colour symbolism is associated with the sacred year; in Buddhism with the picture of the universe, the regions of which are classified according to particular colours; and in the religion of the Maya of Mexico and Central America with the four world directions—east (red), north (white), west (black), and south (yellow). The symbolism of metals and precious stones also is related to their colours (*e.g.*, emerald with green).

The symbolism of the letters of the alphabet varies according to the alphabet (*e.g.*, A and Ω in Greek or *A* to *Z* in English) and is often connected with magic and prophecy—which is also true of the symbolism of numbers. In the picture writing of hieroglyphic systems and in the ideographic (idea-sign) writing of earlier times there is a direct relation between the word-sign and the object to which it refers. In alphabetic writing, numbers and letters are interchangeable if the letter has a number value, as, for example, in Greek and Hebrew alphabets. In some religions, the world of the gods is arranged according to a number system; *e.g.*, into enneads (nines), triads (threes), or dyads (pairs). The idea of oneness was then extended to a numerically arranged pantheon. Gnosticism (a Hellenistic esoteric dualistic system) and Kabbala (a Jewish esoteric mystical system) developed number symbolisms. The letters received a symbolic character in two ways: first, as components of a word for which they stood—*e.g.*, the Hebrew tetragram (four-letter) YHWH (Yahweh) for God, the Latin IOM for Jupiter Optimus Maximus (Jupiter the Best and Greatest), and the Greek IHS for Jesus—or, second through their numerical value, as in A and Ω, the beginning and end of the Greek alphabet, signifying Christ. They then became means of abbreviation—signs possessing a specific content and meaning.

Pictorial. Pictorial symbolism in its many forms is a further development of nonrepresentational, ideographic symbolism and also, to some extent, its origin. In depicting the world of nature, pictorial symbolism captures and mediates the religious experience of reality. The picture shows plainly and clearly the rich and intricate connections of its symbolic content. It may present a part for a whole (a head, a hand, a foot, or an eye for a complete figure) or the whole itself. Symbolic expression of religious experience by means of painting has had a long history.

Sculptural representations of the sacred or holy have their origin in cult. They range from Stone Age idols to the sacred sculpture of early Mesopotamian and Egyptian cultures, from the statues and reliefs of Greco-Roman gods, divinized heroes, and their deeds to the symbolic sculpture of India with its Hindu gods and demons, from the sacred sculpture of China and Japan with their respective pantheons to that of Mahāyāna (Greater Vehicle) Buddhism with its *bodhisattvas* (buddhas-to-be), saints, and spirits. These sacred figures, which may appear in statue or relief form, are sculptured out of various materials. The reliefs on the interiors and exteriors of temples have a decorative function similar to that of wall painting. They narrate a myth or tell a sacred history. Particular parts of the body and symbolical objects may also be sculpturally represented. They may be the male and female sexual organs (*e.g.*, the *liṅga-yoni* in Hinduism); the hand of Sabazius, a Greek god sometimes identified with Dionysus (the god of wine), whose hand is portrayed as raised in blessing and encircled by a number of rather bizarre appendages; and human limbs used as votive offerings for the cure of the part of the body represented. Representative symbolic sculpture tends either to simplify the figure in an abstract, geometrical, or expressionistic style or to imitate nature realistically.

Sacred figures in sculptured or relief forms

Gestural and physical movements. Gestures and bodily movements play an important part in religious ritual and in the religious conduct of man. Such behaviour derives its meaning from its relationship to the holy.

In proceeding to and from a holy place, a worshipper generally proceeds according to certain symbolic patterns: rectilineal, circular, and vertical. Rectilineal movement to and from a holy place is intended to gradually prepare the worshipper for the spatial encounter with the holy and after the encounter to remove him from the sacred sphere. Special streets for processions often are marked off or built to a temple or holy place, such as in ancient Egypt, Mesopotamia, and China. The great procession from Athens to Eleusis for participants in the mysteries possessed a symbolic meaning. Worshippers not only enter a holy place but may also walk around it. Rectilineal and circular movement thus complement each other. Movement to and from a holy place may also be vertical as well as horizontal, as to and from a holy place on top of a mountain or pyramid. All these various types of movement give expression to the symbolism of the holy way or path.

The sacred dance combines rectilineal and circular movements and may also include hopping, jumping, and hand movements. Hand and finger movements in temple dances in Indian and other Asian cultures are strictly regulated and have a precise symbolic meaning. The liturgical dance in a rudimentary form was maintained for a long time in Christianity, as has been the procession. Dancing has not only a significative but also a magical function. It seeks to enchant the holy power.

Hand movements are widely used in ritual and liturgical actions; the touching of holy objects, materials, or men is performed according to a canon (rule) that precisely reg-

Symbolism of hand movements

ulates these gestures and their accompanying prayers and blessings. The gesture of blessing may imitate a symbolic form, such as that of the cross in Christianity. Here the position of the fingers is regulated and has a special meaning, as is also true in the Hindu and Buddhist practice of meditation (*mudrās*). Stroking, thrusting, striking, pushing, waving, and hand clapping also can be symbolical gestures. By raising his hands in prayer, the worshipper approaches the realm of the heavenly gods; by kneeling, the realm of the underworld. This apparently was the original meaning of kneeling before it became an expression of humility. The bow as an intimated genuflection generally indicates respect. The kiss and the embrace—and sometimes also the actions of breathing or spitting upon someone or anointing a person with spittle—were originally magical manipulations; in later usage, they indicated union with or a strengthening of the community or the transferral or communication of power. The holy kiss, whether practiced or only verbally depicted, plays an important part in many religions. Standing is a posture of respect; sitting expresses the reception and acceptance of the sacred word or teaching. It is also the position for meditation as it is practiced in Buddhist monasteries. Symbolic gestures may be either individually or collectively performed.

Verbal symbolism. Gestures are usually accompanied by words. The spoken and written word in religion generally is not thought of primarily as symbolic but rather as a form of rational communication, of communication of thought. Despite its predominantly rational character in modern times, however, language does develop expressions that extend into the area of the symbolical. In its origin, language most likely was richly symbolical. Linguistic symbolism, however, has always had a certain tendency toward rational transparency and logical coherence, and thus words, objects, and pictures—in their origin as symbols—are very closely related. The visual value of the object and picture is later translated into language and enhanced by it.

Linguistic symbolism generally is metaphorical; the allegory, a particular development of the metaphor, symbolically represents an idea by means of a coherent complex of metaphors. Specific genres of narration and literature, such as myth, belong in this category. In a figurative, interpretative, and cryptic sense, names and metaphors denote the person or thing in question. God sometimes is metaphorically called a "spring" or a "rock"; Christ, "the Beloved"; Mary (the mother of Jesus), "the Rose"; and Vardhamāna Malāvira (the founder of Jainism) and the Buddha (Siddhārtha Gautama, the founder of Buddhism) are called "the Conqueror."

Symbolic syllables and names

Individual syllables or sounds may also have a symbolic quality. The *om,* which is used to introduce the holy texts of Hinduism and is a meditation syllable used in Buddhism, provides one example. Understood magically as an emanation of the divine, the word or a name or a part of a word can become an independent hypostatized (substantial) object, a representation, or even the incarnation of the divine, such as the Logos (Word) in the Gospel According to John or *hū* ("he") and *al-ḥaqq* ("the truth") in Islāmic mysticism or the name Metatron in Kabbala. A holy writing or book in its entirety may represent the divine in the same way, as the Bible in Christianity, the Qur'ān in Islām, and the *Ādi Granth* in Sikhism.

Musical symbolism. Music, like the word, also may have symbolic meaning. The basic elements out of which musical symbolism is built are sounds, tones, melodies, harmonies, and the various musical instruments, among which is the human voice. Sound effects can have a numinous (spiritual) character and may be used to bring about contact with the realm of the holy. A specific tone may call one to an awareness of the holy, make the holy present, and produce an experience of the holy. This may be done by means of drums, gongs, bells, or other instruments. The ritual instruments can, through their shape or the materials from which they are made, have symbolic meaning. The Uitoto in Colombia, for example, believe that all the souls of their ancestors are contained in the ritual drums.

The relationship between religious ideas and music is of special importance when the sacred word is set to music or when the music supports or interprets the sacred word by orchestral accompaniment. Medieval and modern Christianity in the West has made important contributions in this area. The symbolic word may be enriched, intensified, and increased in meaning when it is given a musical form. In the medieval technique of motet composition, different but parallel texts from the Bible or the liturgy would be simultaneously sung in various voices to appropriate but different melodies. This is an example of the structuralization of symbols into a coherent whole, a process that may sometimes also be encountered in the visual arts.

Conjunction and combination of various modes. In ritual, liturgy, liturgical and devotional art, and in religious literature and experience, many different types of symbolical expression are frequently combined. Pictorial art may be symbolically interpreted or its present meaning may be reinforced by the addition of a verbal explanation or possibly even by music. In ritual, symbolical words, tones, noises, gestures, signs, odours (*e.g.,* the odour of the sacrifice or the fragrance of incense as an expression of prayer and offering), colours, and pictures are combined. Pictorial art often depicts religious texts and ideas; in so doing it not only uses the human form but also objects of nature, scenery, sacred architecture, and particular symbols. A picture or sign on an emblem often receives its interpretation by the inscription of a verbal explanation. Conversely, in an illustration of religious texts, the picture or sign interprets the text. Over against verbal and musical symbols stands the sacred value of silence. It may indicate devotion, contemplation, or the presence of God.

ICONS AND SYSTEMS OF ICONOGRAPHY

Relationship of symbolism to experience

Throughout the history of their development, religious iconography and symbolism have been closely interrelated. Many religious symbols can be understood as conceptual abbreviations, simplifications, abstractions, and stylizations of pictures or of pictorial impressions of the world of sense objects that are manifested in iconographic representations. In conceiving, describing, and communicating the experience of reality, the realistic picture and the nonrepresentational sign both have as their primary function the expression of this experience in religious terms. In religious pictures that are of a compound or complex nature, particular symbols occasionally reappear. These pictures may also include other types of symbolic representation, such as words, tones, gestures, rituals, and architecture.

Iconographic forms. *Temples and other sacred places.* The architectural iconography of sacred buildings and places of worship is a field of its own. The place of worship, insofar as it is understood as the image of the universe and its centre, must be architecturally patterned according to a specific design of the universe. The place of worship may be considered to be the navel of the world; *e.g.,* the *omphalos,* a round stone in the temple at Delphi (in Greece), the holy stone in the Church of the Holy Sepulchre, in Jerusalem, or the rock in the temple area of the Dome of the Rock (Mosque of Omar), in Jerusalem. A holy place usually is built around these holy points.

The cross-shaped ground plan of the Christian transept church is sometimes interpreted as an architectural portrayal of the crucified Christ, the apse with its altar representing Christ's head. The holy place as a structural creation together with its natural setting may create an idyllic or overwhelming effect, evoking in the beholder an experience of religious awe or devotion. The Shintō and Buddhist temples of Japan and the beauty of the landscape in which they are set, the mountain temples of ancient Greece, and Christian churches and chapels built in such dramatic settings as Le Mont-Saint-Michel in France all inspire a sense of wonderment. The Buddhist temple in all the splendour and richness of its form, trappings, and surroundings or the *stūpa* (a building containing relics of the Buddha) represents the presence of the Buddha.

Great importance, therefore, is often attached to the exterior form of the holy place, and its construction is governed by a canon of symbolical and iconological principles. The individual parts of the building—the walls,



Triumphant cross, symbolizing Christ, before a star-filled heaven, the eternal, surrounded by the symbolic words or letters ΙΧΘΥΣ (ICHTHYS; fish), SALUS MUNDI (salvation of the world), and Α and ω (Alpha and Omega; beginning and end); one set of symbols bringing together several Christian ideas of the relationship between man and God, the temporal and eternal, earth and heaven, the transfiguration and ascension of Christ, the Last Judgment. Apse mosaic in the church of S. Apollinare in Classe, near Ravenna.



Avalokitesvara, the compassionate *bodhisattva*, shown as a composite figure with eleven heads and eight arms, symbolic of his ability to sense man's needs everywhere in the universe. In the Rijksmuseum voor Volkenkunde, Leiden, The Netherlands. Height 35 cm.

The symbol: unity and multiplicity



Buddha's footprint, the "sign of a grand man," containing the symbols of the swastika, the wheel of the law, the *tri-ratna* (the "three jewels" of the Buddhist creed), and the lotus. Stone relief from Gandhara, 2nd century AD. In the Prince of Wales Museum of Western India, Bombay.



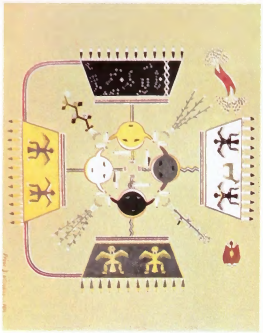
Two-headed Janus, who sees forward and backward, personification of the month of January; romanesque high-relief sculpture of stone. In the Museo del Duomo, Ferrara, Italy.



The Trinity represented by Christ as man, the Holy Spirit as a dove, and God as a hand; Armenian miniature of the baptism of Jesus, 1273. In the Topkapı Museum, Istanbul. 16 × 11 cm.

Plate 1: By courtesy of (top right) the Rijksmuseum voor Volkenkunde, Leiden, The Netherlands; (centre) the Prince of Wales Museum of Western India, Bombay; photographs (top left and bottom left) SCALA—Art Resource/EB Inc.; (bottom right) Ara Guler

Symbols of space and time



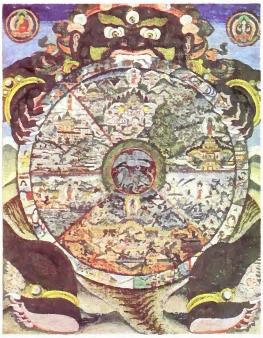
"The Skies," watercolour copy of a Navajo sand painting. Four holy plants radiate from the central four storm figures. The dawn, midday, evening, and night skies are shown in the east, south, west, and north. Paired guardians in the east are a medicine pouch and a bat. In the Museum of Navaho Ceremonial Art, Santa Fe, New Mexico.



Cathedral of Notre Dame, Chartres, France, completed mid-13th century. The nave and apse form a cross; the steeple and bell tower may be interpreted symbolically as the "fingers of God."



Borobudur *stupa*, Buddhist monument in central Java, built in the form of a *mandala*, late 8th century. The surrounding galleries are decorated with reliefs representing scenes of the life and religious development of the Buddha.



Bhava-cakra ("wheel of life," circle of existences), Tibetan *tanka*, probably 19th century. In the Newark Museum, New Jersey. 109.2 × 86.3 cm.

Stonehenge, circular earthwork and stone religious site, Wiltshire, England; late Neolithic to Early Bronze Age (1800–1400 BC).



Plate 2: By courtesy of (top left) the Museum of Navaho Ceremonial Art, Inc., Santa Fe, New Mexico, Newcomb Collection, (centre left) the Newark Museum, New Jersey; photographs, (top right) Alain Perceval, Paris, (centre right) Brian Brake—Rapho, Photo Researchers from EB Inc., (bottom) Aerofilms Ltd.

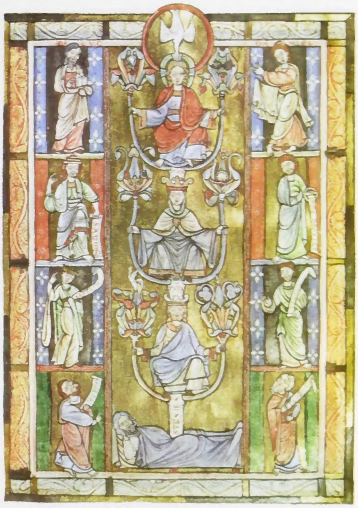


Poseidon, Apollo, and Artemis, detail from the marble east frieze of the Parthenon, Athens, *c.* 440 BC. In the Acropolis Museum, Athens. Height 109.2 cm.

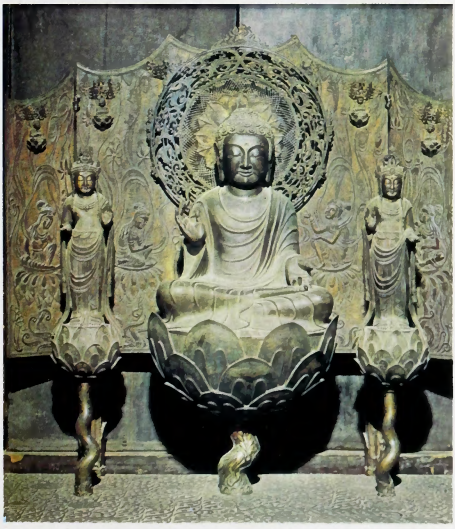
The sacred or holy in the image of man, animals, and plants



Horus as a falcon, Egyptian bronze, 26th dynasty to Ptolemaic dynasty (7th–3rd centuries BC). In the Brooklyn Museum, New York. Height 28.8 cm.



"Tree of Jesse," symbolizing Jesus' descent from the house of David, illuminated page from Rabanus Maurus's *De laudibus sanctae crucis*, from Anchin, mid-12th century. In the Bibliothèque Municipale de Douai, France. 29 × 20 cm.



Buddha Amitabha and two Bodhisattvas on lotus blossoms, the "Amida-Trias." Japanese bronze known as the Tachibana Shrine, 7th–8th century. In the Horyu-ji, Nara Prefecture, Japan.



The four Evangelists represented by their symbols—Matthew by a man, Mark by a lion, Luke by an ox, and John by an eagle—illuminated page (f.27v) from the Book of Kells, Hiberno-Saxon, 8th century. In Trinity College Library, Dublin.

Plate 3: By courtesy of (top left) the Acropolis Museum, Athens, (centre) the Brooklyn Museum, New York, Charles Edwin Wilbour Fund, (top right) the Bibliothèque Municipale de Douai, France, (bottom right) the Board of Trinity College, Dublin; photographs, (top left) G. Nimatallah, (top right) Giraudon—Art Resource, Ltd. Inc., (bottom left) Benrido, Kyoto

Symbols from the realm of family and society



Krsna (right) with his beloved, Rādhā, painting in the Kangra style from India, 18th century. In the Prince of Wales Museum of Western India, Bombay.



The child Krsna, bronze from India, 16th–17th century. In the Detroit Institute of Arts. Height 11.4 cm.



The warring archangel, shown in "St. Michael Triumphing Over the Devil," bronze by Sir Jacob Epstein, 1959. Coventry Cathedral, England. Height 7.6 m.



Christ enthroned as Lord of All (Pantocrator), with the explaining letters IC XC, symbolic abbreviation of "Iesus Christus." Mosaic in the cathedral of Monreale, Palermo.



"The Virgin Adoring the Child," painting by Fra Filippo Lippi, *c.* 1460. In the Staatliche Museen Preussischer Kulturbesitz, West Berlin. 127 × 116.8 cm.

Plate 4 By courtesy of (centre) the Detroit Institute of Arts, (bottom right) the Staatliche Museen Preussischer Kulturbesitz, Berlin—EB Inc.; photographs, (top left) Pramod Chandra, (top right) A.C. Cooper Ltd., (bottom left) Farrell Grehan—Photo Researchers, EB Inc.

columns, ceilings, vaults, and towers—usually have pictorial and symbolic functions. Generally, the ceiling or vault presents a picture of heaven. Special accent is placed on the portals and the paths leading to them, on the position of the tables of offering, altars, sacred pictures, and relics. The bell tower, or campanile, is characteristic of Christian churches and is popularly interpreted as the finger of God. Ancient Christian basilicas (large, roofed buildings, generally with aisles) were viewed as images of the heavenly Jerusalem. The pictorial aspect of the place of worship extends not only to the building in the entirety of its architectural form but also to the painted, sculptured, and mosaic artwork that decorates it. The exteriors of Hindu and Buddhist holy places, such as the famous terrace temple of Borobuḍur on Java, and the pediments and friezes of Greek temples utilize an abundance of figures and reliefs representing scenes from myth and sacred history. The facades of Egyptian temples are covered with tableaus of the gods and depictions of ritual ceremonies. The facades and portal walls and sometimes the outside walls of Christian churches portray the main figures and events in the history of salvation, legends of the saints, and the Last Judgment. Inside the holy place, this pictorial and interpretative function is continued in the figures and scenes on its walls, capitals, and vaults. The adytum (sanctuary), the apses, and the altar may be decorated with symbols or pictures of the divinity or of other gods and saints.

Icons and images. Pictures are the main subject matter of iconography, which also includes free-standing sculptured forms and reliefs. Free-standing figures or statues are important in ritual as well as in partly serving magical purposes, which cannot always be separated from religious ritual. Such figures, which later became objects of personal devotion and meditation, include representations of the gods and demons in various prehistoric religions and of Buddha, Christ, and the various Buddhist and Christian saints. Generally, Judaism, Islām, and ancient Shintō have rejected any representation of the divine.

Pictorial representations of myths and historical events

Painted or sculptured tableaus of historical or mythical events originally belonged in a ritual setting. The function of a wall painting, wall or floor mosaic, or relief was or is to establish the ritual actions as authentic re-enactments of their mythical or historical prototype and to make these mythical or historical events continually present. These tableaus also may be found on the interiors and sometimes the exteriors of houses and on cemetery monuments. They are made to serve private devotion and a personal confession of faith. In the form of a framed picture, Oriental roll picture, print, or book illustration, such an iconographic tableau contains religious information, mediates, and stimulates contemplation and devotion.

Iconographic themes. In the religions of highly developed cultures and in the universal religions, complicated systems of iconography have been developed. In the course of time, however, these systems have been subject to change. Icons (images) may depict the divine in its oneness and in the plurality of its differentiations, emanations, and incarnations, as well as man in his various relationships to the sphere of the holy. They may also depict the world as the stage of divine action, as the realm of the diabolical, or as the battleground of these two warring forces. They may portray evil, the diabolical, and the Satanic (the negatively sacred); or, more positively, they may depict the offer of salvation, redemption, and damnation. Furthermore, icons may portray the ritual means of attaining salvation or moral relationships and duties. Icons may borrow from myths and other religious narrative material to depict the historical past and the present, as well as the future and the afterlife. Icons, finally, may represent religious doctrine and the theological treatment of dogmatic themes, as well as other religious beliefs, religious experiences, and conceptions of a more individualistic nature.

Iconographic types. There are many fundamentally different points of departure in the ways of conceiving the contents of religious pictures and of forming them. These differences, which go back to very early times, continue to exist side by side throughout the history of religions, some dominating at one time while others recede in importance.

Anthropomorphic motifs. The object that generally is depicted in religious pictures or sculpture is an anthropomorphic (human-form) representation. Man is shown as the image and likeness of the holy and as engaging in typically religious behaviour; conversely, the divine appears with anthropomorphic characteristics. This tendency is found quite early in the history of religions. Examples include the religious pictures used in ancestor worship; the spirit and soul idols of various primitive cultures in animism; the fetish, or charm, figures of West African fetishism; and the magical objects of hunter and agrarian cultures. This type of anthropomorphism reaches its high point in the ritual and mythical pictures of the great polytheistic religions and is especially characteristic of ancient Greek religion and also of Jainism in its pictures of the Tīrthaṅkaras (saviours).

Importance of human forms in representations of the holy

In universal religions, such as Buddhism and Christianity, anthropomorphic pictures of the divine were maintained despite criticism. They were not intended to be interpreted realistically but rather as symbolically representing the divine. Buddhism adapted the gods and anthropomorphic myths of the then popular Asian religions and developed the figure of the *bodhisattva* (buddha-to-be) to represent the attainment of Nirvāṇa (the state of extinction or bliss). In Christianity, the picture of Christ usually serves as a representation of the divine. God the Father also is anthropomorphically depicted, usually as an old man wearing papal or imperial insignia. Individual parts of the body may be depicted and serve as symbols of the divine: *e.g.,* the hand of God may stand for Christ, the creative power of God, God's covenant with man, or for God's fidelity and truth; the phallus or foot may symbolize Śiva (a Hindu god). Man may be portrayed as a miniature copy of the universe or as the recipient of salvation and also the bearer of the divine, as in the Christian iconography of Mary and the saints.

Theriomorphic, or zoomorphic, motifs. Besides animal demons in primitive religions and totemism (a belief system and social system based on animal symbolism), animal images frequently occur in other more sophisticated religions. The animal form as a representation of the divine (theriomorphism, or zoomorphism) is characteristic of polytheism. It has been maintained in Hinduism, to some extent in Buddhism, and occasionally in Christianity. Besides the theriomorphic (animal-form) representations of the holy (*e.g.,* the ancient Egyptian gods and animals that are symbols of the divine or the lamb symbolizing Christ in Christianity), there are also theriomorphic (animal-form) pictures of the universe and its powers and of the world of the demons. In many religions the animal kingdom is depicted as a part of creation, as in the portrayals of creation in ancient Greek myths and in the Bible. Animals also play important roles in allegories. Various forms of the shepherd–flock motif have been developed to describe God's relationship to man.

Besides being represented in human form, the Christian Evangelists Mark, Luke, and John are symbolically depicted in animal form (lion, ox, and eagle, respectively). Byzantine iconography sometimes depicts St. Christopher (patron of travellers) with a dog's head. Parts of animals (skulls, horns, wings, and feet) also serve as symbols of the power of the divine or diabolical.

Phytomorphic motifs. Phytomorphic, or plant-form, representations of the divine also are rich in diverse examples and often enigmatic. Holy plants and plants considered to be divine are represented in connection with gods in human form. The god sometimes is the plant itself, as the Egyptian god Nefertum is the lotus, or begets the plant, as the Egyptian Osiris or the Greek Demeter as deities of corn, or the deity comes forth from the plant, as the Egyptian goddess Hathor from the sycamore or the *bodhisattva* from the lotus, or the god unites with or is transformed into the plant, as the Greek heroine Daphne changed into the laurel tree, which thus became sacred to Apollo. The genealogy of Christ from "the root of Jesse," the father of the Israelite king David, is represented as a tree the last blossom of which is Christ. The biblical story of creation describes the vegetative surroundings of man and his dependence on plants (*e.g.,* the tree of knowledge). The tree of life, the world tree, and the primeval cosmic

Plants as representations of the holy

plant all have characteristics related to the nature and origin of the cosmos.

The grapevine is a prominent ritual motif. It is found, for example, in representations of Dionysus and Christ. Painted and sculptured leaf, flower, and plant motifs decorate Christian churches and many religious and funeral monuments. Plants bound into a wreath symbolically promise victory over death and the joys of heaven. In such instances, the simple forms of nature may sometimes be depicted in a nonrepresentational and ultimately abstract and stylized manner.

Hybrid motifs. In religious iconography, anthropomorphic, theriomorphic, and phytomorphic motifs may be combined. The result of this fusion of forms may be seen in the numerous hybrid figures of primitive culture (*e.g.,* totem poles, *uli* figures of New Ireland, and ancestral tablets). Such combined motifs occur also in ancient Near Eastern figures of winged demons with human heads and animal bodies or in winged beings with animal heads and human bodies and in the winged Greek goddesses, as well as in the winged protectresses of the dead in ancient Egypt and the angels and demons in Christian art. In Christianity, the snake in the Garden of Eden is sometimes portrayed with a human head (the face of Satan). In the Middle Ages, representations of the living cross with its arms depicted as hands appear. The cross also has been combined with various other anthropomorphic and phytomorphic elements.

A composite picture of plants, animals, and men together with other natural objects and architectural structures often becomes a sacred scenic background against which the mythical and ritual action takes place. Such scenic depictions were developed in Hellenism and adopted by early Christianity. Paradise scenes including plants, animals, men, Christ, and the saints are later enriched by symbolic and diagrammatic elements. Renaissance painting and East Asian Buddhist and Taoist art also use such combinations when depicting sacred, mythological, and allegorical scenes.

Chrematomorphic motifs. Objects that are used, or chrematomorphic objects, provide another form of pictorial representation. Holy objects, especially those used in worship, fall in this category. The holy book, the cross, the throne and other insignia of power and majesty, lights, lamps, and canopies become representatives of the holy. Garments also may have a symbolic meaning of their own apart from their wearer, as, for example, the veil or the blue mantle of Mary as symbols for the tent of heaven.

Absence of representational forms. The absence of an expected object, person, plant, or animal in a picture or the absence of all pictorial representation may also represent the holy or divine. In the Holy of Holies of the Jewish Temple in Jerusalem there was no picture of Yahweh in or on the ark of the Covenant, although it was supposed to be a sort of portable throne for God. Ancient Christian art often depicted an empty throne on which perhaps lay a folded purple robe or a book (*hetoimasia*) as a symbol of the invisible presence of God. In mosques the empty prayer niche (*miḥrāb*), which is oriented toward Mecca, represents the presence of Allāh. Buddha apparently was not iconically represented in early Buddhist art in accordance with the theory of "emptiness" (*śūnya*) and the radically negative transcendence of the aim of salvation, Nirvāṇa. The rejection of a picture as a means of representing the holy also is a symbolical way of positively asserting the presence of God.

Depiction of the presence of the holy by nonrepresentation

Hostility toward and prohibition of pictures are found in ancient Shintō, Judaism, Islām, the various radical movements (*i.e.,* the iconoclasts, or image destroyers) of 8th-century Christianity that were influenced by Islām, and, centuries later, in some elements of Reformed Protestantism.

INFLUENCE OF MAN'S ENVIRONMENT ON RELIGIOUS SYMBOLISM AND ICONOGRAPHY

Influences from nature. The main streams of the influence from nature are derived from man's experience of nature itself, his position in the universe, and his attempt to master his world in religious terms. Man's sense of the holy influences the way he perceives and understands nature. The space that surrounds man provides him with the dimensional coordinates of his religious experience. Height, depth, breadth, direction, proximity, and distance are the spatial forms in which the holy manifests itself. The holy may reside on a mountaintop, in heaven, in a chasm, in the underworld, in watery depths, or in a desert. The holy way or path provides man with his direction to the divine and a means of approaching it. The spatial position of the holy and the direction to it may also be abstractly expressed—*e.g.,* by means of symbolical numbers or coordinates. The infinity of space may be represented by geometrical and linear figures.

Emptiness or fullness may characterize the utilization of spaces and surfaces that are usually intended for the reception of symbols and signs. Works of art may be totally absent in certain architectural structures; or all available space may be filled with a dense profusion of all kinds of figures and objects, all of which may sometimes be encircled by an ornamental network or web of branches, vines, leaves, and blossoms; an example of such embellishment is Islāmic art. The ebb and flow of, time and things, the flow of water, and the cyclic recurrence of time are pictorially expressed—in symbols such as the wheel, spiral, wave, and circle. Time appears as the god of destiny—*kāla* in ancient Indian and Zurvān in ancient Persian religions. In late antiquity time takes the form of a demon entwined with snakes (Aion). The snake biting its own tail, the ring, and the spiral are frequently recurring symbols of fate and eternity; in Christianity, eternity is represented by the A and Ω, and the wreath.

Other physical, chemical, and physiological facts of nature also serve as sources of symbolic and iconographic concepts. Examples include the experiences of seeing, hearing, tasting, smelling, and touching; the myriad forms of plant and animal life; heaven and its astral and meteorological phenomena, which may be represented realistically or abstractly through symbols or personifications; and the colours and various colourful natural occurrences such as the rainbow (often symbolizing God or Christ) or the sunrise and sunset or minerals and precious stones. In the depiction of natural happenings, plant, animal, and human forms may blend into one another, as in the symbolic circle of the Mesopotamian god Tammuz in which the tree of life is combined with figures of fertile wild and domestic animals and the figure of a shepherd. All these symbols represent the preservation and regeneration of life. They also represent nature as a holy power.

Another area of nature symbolism is that of the microcosm and macrocosm of heaven and Earth. Heaven and Earth are depicted as a dually or polarly related pair, which generally are theistically personified as a man and a woman. These roles may sometimes be reversed, as in ancient Egypt: the heavenly divinity a goddess, Nut, and the Earth divinity a god, Geb. In Greek myths on the origin of the gods (theogony), the world of the gods and man results from such pairing. Mother Earth is a central figure of many myths: she is the mistress of fertility and death.

Influence of human relationships. Another group of pictures and symbols that are especially significant in depicting the relationship of God and man are those drawn from the area of family and social relationships, especially the roles of the father and mother. These relationships to some extent are determined by the structure of the society and its economy. The mother image is closely bound up with Earth symbolism, vegetation, agriculture, fertility, the reappearance of life, and the lunar cycle. The father image usually is associated with the sphere of heaven, authority, dominion, age, wisdom, and struggle. Love, betrothal, marriage, sexual union, family, and friendship also are significant in symbolization. The relationships between brothers and sisters are of importance, especially in the structure of religious communities and in the various fraternal groups and secret organizations of modern societies. The images of the child, the subject, or slave again indicate man's relationship to God; those of the ruler, king, or master express the power and authority of the deity. Even the structure of the world of the gods is explained in terms of family.

The importance of social relationships in concepts of the holy

Symbolism of sex and the life cycle. The symbols of sexuality and the life cycle perform a function similar to those of time and eternity in the higher religions. They indicate the permanence of the cycle of sexual functions and the return and renewal of individual and collective physical life. The endless renewal of life is variously represented. It may be as realistic depictions or diagrammatic and stylized abbreviations of man and woman, god and goddess, masculine and feminine animals in the act of love and sexual union, as in reliefs on Hindu temples, or as depictions of sex characteristics (*e.g.,* in Indian *liṅga-yoni* symbolism). The theme of renewal also may be depicted in representations of woman with emphasis on her function as mother, as in the nursing-mother figures of ancient Greece. The life cycle also is represented by figures portraying the ages of man or by depictions of pain and suffering, as in pictures of the Buddha's death, which also indicate his breaking out of the endless chain of existence.

Cultural influences. Other cultural, political, social, and economic institutions and conventions also influence religious symbolism and iconography. Work and leisure, war and peace, and the myriad things associated with them—occupations, positions in society, classes and their functions, the tools of domestic and professional life, technical equipment, forms of international relations and strife—all play an important part in man's interpretation and understanding of religious reality and hence in his symbolization of this experience. Hunters, farmers, shepherds, warriors, artisans, and merchants and their activities are represented in religious pictures and appear in the verbal symbolism of religion. In the universal and missionary religions, such as Buddhism, Christianity, and Islām, the believer is summoned to take up spiritual arms and fight for salvation. In Judaism, Christianity, and the religion of ancient Rome, the relationship between God and man is regulated according to the model of a peace treaty. In ancient German and Indian religions the military virtues of loyalty, duty, and comradeship are stressed. Man's religious activities may also be expressed in terms of play and sport, training, competition, and victory.

Conceptual influences. Ideas, theories, and structured systems of thought also are incorporated into religious symbolism. Abstract ideas—such as wholeness, unity, and the absolute—and the power of the spirit are concretely expressed in religious terms. The idea of unity plays an important part in expressing the oneness of the divinity. Mathematical principles expressed in number symbolisms are used to organize the world of the gods, spirits, and demons, to describe the inner structure of man, and to systematize mythology and theology. The concepts of duality or polarity find expression as the body and soul of man: the divine pair; the syzygy (paired emanations) in Gnosticism; the dualism of God and the devil, of good and evil; and, finally, as the two natures of Christ. The number three, or triplicity, is represented in divine triads, the trinity, and the body–soul–spirit structure of man; as is number four, or quaternity, in the four cardinal points, the picture of the cosmic whole, the divine quaternity. Time and eternity may be expressed in abstract symbolical terms as well as concretely in picture form.

Symbolism of numerical concepts

INFLUENCE OF RELIGION ON SYMBOLISM AND ICONOGRAPHY

Religious figures and spiritual authorities themselves form a vast complex of symbols: gods, saviours, redeemers, heroes, the avatars (incarnations) and the Īśvaras (manifestations) of Hinduism, the heroes and gods of epics, the founders, lawgivers, saints, and reformers of the great religions. The biblical prophets, apostles, and evangelists and the Christian saints are characterized by a very complicated system of symbols. Theologians, mystics, and contemplatives may also be symbolically and pictorially represented; the doctors (teachers) of Roman Catholicism and Eastern Orthodoxy and fathers of the early church have standard iconic forms, attributes, and symbols (*e.g.,* St. Augustine is represented by the heart; St. Jerome by the lion). Persons connected with ritual and representatives of the religious institution (*e.g.,* hierarchs, priests, assistants in the liturgy, male and female dancers, and musicians) may also be symbolically and iconographically depicted.

The offering, the place of offering, the altar and its trappings, the instruments that prepare and destroy the offering, the fire that consumes it, the liquids and drinks used in the rite, the sacred meal, and the rites of communion all are objects of iconography and symbolism. The offering symbolizes the idea of submission to the ideals of a religion, the giving up of valuables and possessions for religious purposes and for the service of human brotherhood, and the giving up of one's life for religion.

A religious community recognizes itself and its ideas by symbols. Examples are the *yin-yang* (union of opposites) symbol bound by the circle of stability (*t'ai-chi*) in Chinese universalism; the swastika in Hinduism and Jainism; the wheel of the law in Buddhism; the *khaṇḍā* (two swords, dagger, and disk) in Sikhism; the star of David or *menora* (candelabrum) in Judaism; and the cross in its various forms in Christianity.

CONCLUSION

The further development of symbolism and iconography in the higher religions of the modern world is an open question. During much of the 20th century in the Christian communities, revivals of the liturgical traditions and of ritual symbolism were in progress, though criticized vigorously by many theologians. Liturgical symbolism became valued anew and stabilized. Theological systems, like that developed by Paul Tillich, were based on the concept of the symbol. On the other hand, during the 1960s some indifference toward symbols and pictures developed because of an emphasis on the moral and social tasks of religion. Symbols, myths, pictures, and anthropomorphic ideas of God were rejected by many theologians, and philosophical structures (*e.g.,* the theory of the demythologization of the Bible, or the "God-is-dead" theology) became substitutes for them. In the great non-Christian religions, this process seems to be less acute. Within the horizons of a secularized, skeptical, and agnostic society, religious symbols seem to be dispensable, but nonetheless a new and increasing interest in symbols was appearing, especially among the younger generations who came into contact with both Eastern and Western religious and cultural traditions with their rich sources of symbolic images and modes of thinking. Thus, a resurgence of an understanding for the specific values of symbolism and iconography has been recognized in the latter part of the 20th century, in spite of all apparently opposite trends.

Modern uses of religious symbols

BIBLIOGRAPHY

General works: Bibliographie zur Symbolik, Ikonographie und Mythologie, 4 vol., ed. by MANFRED LURKER (1968–71), contains fundamental and comprehensive bibliographies; *Symbolik der Religionen,* ed. by FERDINAND HERRMANN (1958–), a series of compendious and important monographs; JUAN E. CIRLOT, *A Dictionary of Symbols,* 2nd ed. (1971, reprinted 1983; originally published in Spanish, 1962), with an excellent bibliography; RENÉ ALLEAU, *La Science des symboles* (1976), a study of the principles of symbolism and methodology of interpretation; J.C. COOPER, *An Illustrated Encyclopaedia of Traditional Symbols* (1978); AD DE VRIES, *Dictionary of Symbols and Imagery,* 3rd rev. ed. (1981). See also *Visible Religion* (annual), started in 1982 by the Institute of Religious Iconography, State University, Groningen.

The idea and nature of symbols, symbolization, and culture: HAROLD BAYLEY, *The Lost Language of Symbolism: An Inquiry into the Origin of Certain Letters, Words, Names, Fairy-Tales, Folklore, and Mythologies,* 2 vol. (1912, reprinted 1968); ERNST CASSIRER, *The Philosophy of Symbolic Forms,* 3 vol. (1953–57; originally published in German, 1923–29), a pioneer work; HUGH DALZIEL DUNCAN, *Symbols in Society* (1968, reprinted 1972); MIRCEA ELIADE, *Images and Symbols: Studies in Religious Symbolism* (1961, reissued 1969; originally published in French, 1952); ERICH FROMM, *The Forgotten Language: An Introduction to the Understanding of Dreams, Fairy Tales, and Myths* (1951, reissued 1974), discusses psychoanalytical aspects; DÉSIRÉE HIRST, *Hidden Riches: Traditional Symbolism from the Renaissance to Blake* (1964); HANS JENSEN, *Sign, Symbol and Script,* 3rd rev. ed. (1969; originally published in German, 1925); FREDERICK ERNEST JOHNSON (ed.), *Religious Symbolism* (1955, reissued 1969); CARL GUSTAV JUNG, *Psyche and Symbol,* ed. by VIOLET S. DE LASZLO (1958), and *Man and His Symbols* (1964, reprinted 1979); GYORGY KEPES (ed.), *Sign, Image and Symbol* (1966); SUSANNE K. LANGER, *Philosophy in*

a New Key: A Study in the Symbolism of Reason, Rite, and Art, 3rd ed. (1957, reprinted 1979); LUCIEN LÉVY-BRUHL, *L'Experience mystique et les symboles chez les primitives* (1938); ROLLO MAY (ed.), *Symbolism in Religion and Literature* (1960); RUDOLF OTTO, *The Idea of the Holy*, 2nd ed. (1950, reprinted 1970; originally published in German, 1917; this translation is from the 9th German ed., 1922); JOSEPH R. ROYCE *et al.*, *Psychology and the Symbol* (1965); THEODORE THASS-THEINEMANN, *Symbolic Behavior* (1968); PAUL TILLICH, "Existential Analyses and Religious Symbols," in HAROLD A. BASILIUS (ed.), *Contemporary Problems in Religion*, pp. 35–55 (1956, reprinted 1970); JOACHIM WACH, *The Comparative Study of Religions* (1958, reprinted 1966); HEINZ WERNER and BERNARD KAPLAN, *Symbol Formation; An Organismic-Developmental Approach to Language and the Expression of Thought* (1963); ALFRED NORTH WHITEHEAD, *Symbolism; Its Meaning and Effect* (1958); EDWARD C. WHITMONT, *The Symbolic Quest: Basic Concepts of Analytical Psychology* (1969, reissued 1978); S. FOSTER DAMON, *A Blake Dictionary: The Ideas and Symbols of William Blake* (1965, reprinted 1979), a generously documented encyclopaedia; JOSEPH CAMPBELL, *The Mythic Image* (1974), an analysis of mythologies from different cultures embracing Buddhism, Christianity, and Islām; and JOHN SKORUPSKI, *Symbol and Theory: A Philosophical Study of Theories of Religion in Social Anthropology* (1983), a highly scholarly work.

Recurrent themes in history of symbolism and iconography: LEROY H. APPLETON and STEPHEN BRIDGES, *Symbolism in Liturgical Art* (1959); JITENDRA NATH BANERJEA, *The Development of Hindu Iconography*, 2nd ed. rev. (1956, reprinted 1974); BENOYTOSH BHATTACHARYA, *Indian Buddhist Iconography*, 2nd ed. rev. and enlarged (1958); E. DOUGLAS VAN BUREN, *Symbols of the Gods in Mesopotamian Art* (1945); MAURICE FARBRIDGE, *Studies in Biblical and Semitic Symbolism* (1923, reissued 1970); GEORGE FERGUSON, *Signs and Symbols in Christian Art*, 2nd ed. (1955); ANTOINETTE K. GORDON, *The Iconography of Tibetan Lamaism*, rev. ed. (1959, reprinted 1967); DONALD A. MACKENZIE, *The Migration of Symbols and Their Relations to Beliefs and Customs* (1926, reprinted 1970); CHARLES P. MOUNTFORD, *Art, Myth and Symbolism* (1956); JOSEPH M. KITAGAWA and CHARLES LONG (eds.), *Myths and Symbols* (1969, reprinted 1982); DORA and ERWIN PANOFSKY, *Pandora's Box: The Changing Aspects of a Mythical Symbol*, 2nd rev. ed. (1962, reprinted 1978); H. DANIEL SMITH, K.K.A. VENKATACHARI, and V. GANAPATHI, *A Source Book of Vaiṣṇava Iconography According to Pāñcarātrāgama Texts* (1969); CHARLES ALFRED SPEED WILLIAMS, *Encyclopedia of Chinese Symbolism and Art Motives* (1960); HEINRICH ZIMMER, *Myths and Symbols in Indian Art and Civilization* (1946, reprinted 1972); GERSHOM G. SCHOLEM, *On the Kabbalah and Its Symbols* (1965, reissued 1969; originally published in German, 1960); GLADYS A. REICHARD, *Navaho Religion: A Study of Symbolism*, 2nd ed. (1974, reprinted 1983); BEATRICE L. GOFF, *Symbols of Prehistoric Mesopotamia* (1963), and *Symbols of Ancient Egypt in the Late Period: The Twenty-First Dynasty* (1979); JAMES A. AHO, *Religious Mythology and the Art of War: Comparative Religious Symbolism of Military Violence* (1981). (K.M.A.G.)

Rembrandt

For most modern observers Rembrandt Harmenszoon van Rijn remains synonymous with 17th-century Dutch painting, and his art has attained a kind of universal familiarity and popularity. Yet the biblical scenes and the self-portraits that today form the hallmark of his art were by no means typical of Dutch pictures then; more commonly, his contemporaries produced landscapes, still lifes, or genre scenes of daily life that never held great interest for Rembrandt. In his own era Rembrandt achieved greatest fame as the most fashionable portrait painter of Amsterdam during the 1630s, but he was eventually eclipsed even during his own lifetime by younger rivals, including some of his own students. Another major field of accomplishment lay in the medium of etching. Rembrandt commanded high prices for his prints even during his lifetime, and his technical mastery had a lasting effect on printmakers for centuries.

By courtesy of the Iveagh Bequest, Kenwood House (Greater London Council)



Rembrandt, self-portrait, oil on canvas, *c.* 1661–62. In the Iveagh Bequest, Kenwood House, London.

If any quality typified the works of this great artist, especially in his youth, that quality would be a personal ambition to rival the dominant artists of Europe, particularly Peter Paul Rubens from nearby Antwerp. But the tides of fashion in Holland and Rembrandt's own temperament seem to have frustrated much of his ambition and left him increasingly isolated and idiosyncratic in his final years. There is actually a kernel of truth to the apocryphal legend of Rembrandt's rejection by the leading patrons of Amsterdam, although this loss of favour was gradual and never total. As a result of his increasing isolation, however, Rembrandt achieved a particular personal independence that doubtless contributed to his distinctive and evocative suggestion of the timeless human world of quiet yet deep emotional states. The silent human figure remained the central subject of Rembrandt's art and contributed to the sense of a shared dialogue between viewer and picture, which still is the foundation of Rembrandt's greatness as well as of his popularity today.

Early years in Leiden. Rembrandt's youth does not help much to explain either the derivation or the character of his art. The artist was born on July 15, 1606, in the university city of Leiden. His father, Harmen Gerritszoon van Rijn, was a miller, a reasonably prosperous man; the family of his mother, Neeltje van Zuytbroeck, were bakers, but more important, they remained Catholics at a time when Leiden had adopted the Protestant creed. Indeed, Rembrandt's father was the only member of his family who became a Calvinist rather than remaining a Catholic. According to a Leiden chronicle written during the artist's lifetime (by Jan Orlers, 1641), the young Rembrandt was sent to a Latin school and directed toward the local university, the very first to have been established in Holland (1575) and a major centre of learning. But because the young man's proclivities led toward art, he was apprenticed during the period 1619–22 to the local painter Jacob Isaacszoon van Swanenburg. Little of the work of van Swanenburg can be identified today, and his art seems to have left scant influence on Rembrandt, but the fact that he, too, was a Catholic might have affected the choice of van Swanenburg as a master. Moreover, van Swanenburg's father had also been a highly successful painter in Leiden and had trained Rubens' master, Otto van Veen. Thus, this tutor held out a potential set of connections for the young Rembrandt.

Apprenticeship with Pieter Lastman

But Rembrandt's chief training came from the Amsterdam painter Pieter Lastman (1583–1633), who had spent time in Italy (1603–1606/07) and had returned to Amsterdam to become the leading painter of biblical, mythological, and historical pictures. Although Rembrandt seems to have spent only about half a year with Lastman around 1623, he fully absorbed the lessons of his master. From Lastman he learned the importance of painting lofty subjects in a broad format with careful attention to the ancient costumes, dramatic gestures, and compositional groupings of the full-length figures. The earliest Rembrandt pictures, including "Stoning of Saint Stephen" (1625), "Palamedes Before Agamemnon" (1626), and "Baptism of the Eunuch" (1626), clearly derive closely from both the themes and the pictorial formulas of Lastman. The baptism of the eunuch, for example, had already been painted by Lastman in a broad format a few years before (1623; Karlsruhe); Rembrandt's version of the scene from Acts is transposed into a vertical format, but it retains most of the same figures, costumes, and accessories, yet condensed into a tighter, more dramatically lighted mass. Another close comparison of both theme and form is provided by Lastman's 1622 panel and Rembrandt's denser, vertical 1626 panel of the same subject, "Balaam's Ass and the Angel." Recent research links the "St. Stephen" and the "Palamedes" with commissions from the young Rembrandt in Leiden by a local humanist named Petrus Scriverius, whose estate cites two large pictures by Rembrandt; otherwise the early patrons of these pictures are unknown today.

The early Rembrandt paintings already reveal the artist's ambition to rival the leading painters in Europe. Not only did he concentrate on the most learned and morally serious subjects but he also strove for the historically plausible settings and costumes that distinguished the pictures of Lastman and such painters in Rome as the German émigré Adam Elsheimer. Also evident in these early paintings are Rembrandt's nascent fascination with dramatic personal responses and with spotlight effects of light and shadow. If anything, these elements came to dominate his art in the succeeding decade. In particular, Rembrandt's exposure to a group of artists from nearby Utrecht led to an abrupt emulation of their sharply drawn chiaroscuro, or painting in light and dark. These Utrecht painters, led by Gerrit van Honthorst, had recently returned from Rome, and their art enjoyed not only local popularity but also strong favour in the courts of northern Europe. Hence, when Rembrandt painted such religious works as "The Presentation in the Temple" (*c.* 1627–28) or "Christ at Emmaus" (1628), he sought to emulate the drama of lighting and gesture of Elsheimer, Caravaggio, and, now,

van Honthorst and to place himself firmly into the international world of art. A measure of the self-concept of Rembrandt around this time is the small but dramatic "Young Painter in the Studio" (*c.* 1629), which shows a full-length shadowy figure of an artist situated against the back wall and dwarfed by a massive panel lying on its easel in the foreground. This panel, seen from behind, lies in shadow, with only its near edge glowing with light. The overall effect is one of heroic confrontation within the very act of creation.

That Rembrandt had attained eminence as an artist by the end of the 1620s can be discerned from a famous reference, dating from 1629/30, in the autobiography of Constantijn Huygens, the secretary of the Prince of Orange. Huygens singles out Rembrandt as well as his young Leiden friend and colleague, Jan Lievens (1607–74), for special praise in terms of their future promise as artists. Rembrandt is lauded for his penetration to the essence of his subjects and for his effects in small format. In particular, the 1629 panel "Judas Returning the Thirty Pieces of Silver" is held up as a model for moving gesture and emotion, worthy of the finest works of Italy or even of antiquity. Huygens' chief regret is that Rembrandt and Lievens never traveled to Italy for further study of the past masters.

Only in recent years has Lievens begun to receive attention commensurate with that paid to Rembrandt, although the careers of the two artists developed in tandem for many years. Lievens, too, journeyed from Leiden to Amsterdam for a two-year apprenticeship (1618–19) with Lastman. Indeed, it may well have been the example of Lievens that led Rembrandt to study with Lastman, and the influence of Lievens remained essential. Probably through Lievens came the exposure to Utrecht painting, which was to influence Rembrandt's art. In the late 1620s Lievens' art so closely resembled Rembrandt's that scholars are still debating the proper attribution of some panels. For example, Lievens' "Capture of Samson" (*c.* 1627–28; Amsterdam) appears to have been the stimulus for Rembrandt's "Capture of Samson" (1628), and both works emulate the same subject as painted by Rubens (1610; London) and circulated throughout Europe in prints from an engraved version. In similar fashion Rembrandt and Lievens maintained a pictorial dialogue concerning the subject of the raising of Lazarus, beginning with Rembrandt's *c.* 1630 panel (Los Angeles), followed by Lievens' 1631 canvas (Brighton) and etching, and ending with Rembrandt's masterful, dramatic, and large etching of about 1632. Rembrandt even seems to have predated some of his works to make them seem earlier than the comparable Lievens compositions. In 1632, however, Lievens departed for England, where he most likely became acquainted with Anthony Van Dyck, whose art redirected his own and led him to a later career in Antwerp between 1635 and 1644 before he returned to Amsterdam.

Friendship and rivalry with Jan Lievens

As part of the same ambition to paint historical pictures, both Rembrandt and Lievens also experimented with studies of heads, or what the Dutch call *tronies*. Often these figures wear exotic millinery and receive dramatic poses and lighting, but they are not portraits. Rather, they seem to have served as possible models or practice pieces for the character heads to be included within larger histories. Many of the pictures with the same models that were known in the 19th century as Rembrandt's "father" or "mother" are actually such studies of heads, with special attention to the rendition of stuffs, of lighting, and of facial expressions or features. Many of the early self-portraits also seem to have been variants of the *tronies* formula, in which Rembrandt simply used his own features in lieu of those of another model and dressed himself up in military or fashionable garb: plumed hats, golden chains, armour gorgets. Some of the heads of the older models reappear virtually without change on the numerous prophets and apostles (including the luminous 1630 "Jeremiah") that Rembrandt produced in 1630–31 in his later years in Leiden; this was a kind of picture that he left off doing until his final decade in the 1660s.

Rembrandt already enjoyed the attention of pupils and followers during his early years. His first disciple was Gerrit Dou, who emulated still another category of pictures from Rembrandt's oeuvre: his genre scenes, or depictions of everyday activities. Rembrandt had already created such scenes in his 1626 "Music Lesson," a work that also features archaic costumes and suggestions of lustfulness. Dou, also a Leiden native, the son of a glass engraver, became a pupil of Rembrandt in 1628 and continued this kind of subject but with overtones of seriousness and moral instruction and with an enamel-like fineness on a minute scale that was highly prized by collectors.

Having attracted the attention of the influential Huygens at court in The Hague, Rembrandt made inroads with the ruling House of Orange, chiefly with Prince Frederick Henry, for whom he painted in 1632–33 two scenes of Christ's Passion, the "Raising of the Cross" and the "Descent from the Cross" (both in Munich) as well as a portrait of the princess Amalia van Solms that was to have been the pendant of a van Honthorst portrait of Frederick Henry. The Passion scenes were ordered for the Prince by Huygens and are closely linked to the model of Rubens, again known to Rembrandt chiefly through an engraving. At the time Rubens was the leading artistic force in Europe, and as a cultivated diplomat as well as a consummate painter he was especially favoured at princely courts. Thus, to emulate Rubens' "Descent from the Cross" for his own princely patron was for Rembrandt the highest act of artistic self-assertion. Rembrandt even went so far as to produce his own 1633 etching of his picture in emulation of Rubens. One striking feature about both of Rembrandt's Passion scenes is that the artist gave his own features to participants within the scene; in the "Raising of the Cross" he even employed modern dress and a focused light to underscore this personal involvement, meant perhaps to express his own meditative spirituality.

Emulating Rubens

The letters from Rembrandt to Huygens concerning the Passion series survive, and they document a second phase of artistic production between 1636 and 1639, when three more pictures were made for Frederick Henry. The letters document the progressive disenchantment with Rembrandt by Huygens and the Prince, but one of them also contains a rare personal testimonial from Rembrandt concerning his artistic aims. The letter underscores the artist's commitment to evoking "the greatest and most natural emotion" for his religious subjects. In this respect he is close to Rubens, who also was dedicated to the evocation of energy, drama, and emotion. Rembrandt's works in comparison present less of the heroism and beauty of Rubens' scenes but emphasize instead dramatic nocturnal lighting, humble figures, and intimate, lifelike reactions of his religious actors. These were basically the same elements that Huygens had already singled out for praise in Rembrandt's earlier pictures, and they continued to inform his religious art during the 1630s.

Early years in Amsterdam. Rembrandt moved to Amsterdam in late 1631. He already had a dealer in that city, Hendrick Uylenburgh, and his prospects at court were eclipsed by the domination of van Honthorst. Thus, the prosperity of Amsterdam, a capital of capitalism and a virtual city-state, drew him inexorably. In part through his introductions from Uylenburgh, Rembrandt quickly became one of the most fashionable and well-paid portraitists in Amsterdam. He was able to impress the regents of his adopted city, that clan of mercantile patricians who formed the centre of political power and influence. A mark of Rembrandt's early success was his commission to paint "Anatomy Lesson of Dr. Nicolaes Tulp" (1632), a commemoration of the annual anatomic demonstration to the city's guild of surgeons by its praelector, or chief surgeon. This large-scale group portrait by Rembrandt has been justly celebrated ever since for its departure from the rule of showing a coordinated row of portrait heads. In contrast Rembrandt animated his subjects through a pyramidal composition and his mastery of dramatic lighting to focus attention on the actual process of the lecture itself. At the same time he enlivened the faces of the listeners with a rich variety of expressions of attention, investing them with the same suggestive pictorial psychology that would remain his trademark. Many of the same features can be found in Rembrandt's portraits of individuals or

"Anatomy Lesson of Dr. Nicolaes Tulp"

of husbands and wives painted shortly after his arrival in Amsterdam. Although Rembrandt had painted very few portraits while at Leiden, his first four years in Amsterdam brought him some 50 portrait commissions, most of them quite well paid. Inasmuch as Nicolaes Tulp was not only a surgeon but also an alderman and a member of the Amsterdam town council, he was an influential man within the regents' group. Also popular with the regents was Uylenburgh, the art dealer with whom Rembrandt lived briefly and also entered into commercial partnership. Many of Rembrandt's portrait sitters (*e.g.,* Marten Looten, 1632) appear to have been Mennonites, religious conservatives, whom he met through Uylenburgh and who were well connected with the Amsterdam regents.

Rembrandt also portrayed a number of religious leaders of Holland during his first decade in Amsterdam: the Remonstrant Johannes Uytenbogaert (1633 panel and 1635 etching), the Calvinist Johannes Elison (1634), and the Mennonite Cornelis Anslo (1641 double portrait panel and etching). This last figure was a renowned preacher, and Rembrandt's portrayal emphasizes Mennonite reliance on the spoken word. In general his renditions take up the traditional challenge to the pictorial arts to render life without the aid of the spoken or the written word, as if in response to the challenge written in verse by the greatest of 17th-century Dutch poets, Joost van den Vondel:

> That's right, Rembrandt, paint Cornelis's voice!
> His visible self is second choice.
> The invisible can only be known through the word.
> For Anslo to be seen, he must be heard.

Yet, in addition to these portraits and the numerous pendant pairs of portraits during these early Amsterdam years, Rembrandt also clearly yearned for recognition, after the model of Rubens, as a painter of both mythologies and biblical stories. About the time of his move from Leiden, he produced his most extensive group of mythologies, beginning with "Andromeda" (*c.* 1630), which stresses the pathos rather than either the beauty or the heroism of the nude victim. A large "Pluto and Proserpina" (*c.* 1632) was clearly made for Frederick Henry at the same time as the Passion pictures, and both its scale and its frenetic energy attest to its relationship to the idiom of Rubens, although the refined execution still harks back to the Leiden of Dou. More typical, however, of Rembrandt's tendency to demythologize is the way he renders such subjects as "Rape of Europa" (*c.* 1632) or "Rape of Ganymede" (1635). The former places a seraglio of exotically clad, small-scale women in front of a shoreline that includes a Dutch harbour scene. The latter scene is even more prosaic, showing a mewling toddler instead of the seductively beauteous youth of legend. Not only is the eagle elevating the child upward against a leaden-gray sky but also the frightened boy urinates reflexively in his horror. The artist seems almost to have taken pains to violate conventions of beauty and decorum in such a work, as if to engage in persiflage rather than homage to the classical heritage. Scholars still debate whether this work holds Neoplatonic significance as a mythic analogue to the union of the Christian soul with the divine or whether its irreverence lies closer to the homophilic traditions of the subject.

Rembrandt's mythological paintings

Rembrandt's masterpiece of mythology followed a year later: "Danaë" (1636). Again, the artist avoided the standard formulation of the subject; instead of the shower of golden rays or coins that signal the arrival of Jupiter in his metamorphosed state, this picture simply bathes the heroine in a rich, luminous light. The "Danaë" also provided Rembrandt with the occasion for one of his most splendid nudes (reworked during the 1650s). He portrays the young woman waiting expectantly for her approaching lover despite the mournful, bound cupid above her head. The theory has been advanced that this was the large-scale picture sent as an unsolicited gift to Huygens and referred to in Rembrandt's letters: "My lord, hang this piece in a strong light and so that one can stand a distance from it, then it will show at its best."

If Rembrandt's mythologies reveal a turn toward the unconventional, his biblical subjects offer a range of themes and stagings. His paintings include a number of unusual subjects from the Old Testament, drawn from the story of Samson in the Book of Judges or from the Book of Daniel. One large picture of this kind is "Belshazzar's Feast" (*c.* 1635), a work filled with Oriental costumes, dramatic gestures, and horror-stricken facial expressions. (In addition, the presence of accurate Hebrew orthography testifies to Rembrandt's connections with the Sephardic Jewish community of Amsterdam through yet another religious leader, Menasseh ben Israel, an ecumenical and millenarian author. Rembrandt also etched a portrait of Menasseh in 1636.) Some of the same vitality and energy informs Rembrandt's "Sacrifice of Isaac" (1635; St. Petersburg), where the angel's sudden intervention stills the hand of Abraham and causes the fatal knife to hang suspended in midair, giving the scene photographic immediacy. What is striking about the St. Petersburg picture is that it has a slightly modified mate in Munich, inscribed with the ambiguous signature *Rembrandt verandert en overgeschildert* ("Rembrandt altered and overpainted"), suggesting that the documented Uylenburgh workshop practice of copying works in stock might well have been a practice of Rembrandt's own store of pictures during the 1630s.

Another feature of Rembrandt's output during the 1630s is his development of ambitious etchings of biblical subjects. A beginning to this trend may already be seen in the 1633 "Descent from the Cross," adapted from the painting for Frederick Henry, and in the mid-1630s Rembrandt and Uylenburgh seem to have had the ambition to create a series of large etchings at premium prices. One etching, "Christ Before Pilate" (1636), was carefully modeled in advance in an oil sketch of grays and browns (1634). Another oil sketch on canvas, "John the Baptist Preaching" (*c.* 1634), was never completed as a print. Its broad expanse of figures in picturesque costumes against a ruined setting still recalls the formulas of Lastman, who had died in 1633. Another ambitious etching that shows an ostentatious mastery of rich, dark tones is the 1634 "Annunciation to the Shepherds." Here Rembrandt once more added a native Dutch element in transforming the shepherds into cowherds, who evince their own primal fear before the glowing angelic apparition. The character of experimentation that would typify Rembrandt's etchings emerges from this print in the conscious reversal of lights and darks. Here the forms and highlights come forward out of the dense, dark lines, partly through relative lightness and partly through actual removal of lines with a burnisher. This technique, which creates the effect of light emerging out of darkness, became a hallmark of some of Rembrandt's most moving late religious prints.

Etchings of biblical subjects

Some of Rembrandt's finest pupils can be documented in Amsterdam in the late 1630s: Govert Flinck from Kleve, Jacob Backer of Harlingen, and Ferdinand Bol of Dordrecht. Flinck (1615–60) in particular may well have reproduced several of Rembrandt's pictures for collectors or else collaborated with him in the workshop. But the replication of Rembrandt's formulas for historical pictures, portraiture, and *tronies* remained a staple of his pupils for many years to come.

Rembrandt married on June 22, 1634. His bride was Saskia Uylenburgh (born in 1612), the cousin of Hendrick Uylenburgh and the daughter of the burgomaster of Leeuwarden. Saskia brought a substantial dowry as well as patrician status with her, so this marriage represented a substantial climb in social status for Rembrandt. At the time of the marriage, Rembrandt's court connections with The Hague were at their zenith; hence, the match might have seemed well balanced. One of Saskia's cousins and her guardian when the couple first met was the Reformed clergyman Johannes Cornelisz. Sylvius, whose posthumous etched portrait of 1646, emerging dynamically out of an illusionistic oval window frame, offers one of Rembrandt's most moving likenesses. Yet Saskia herself was to be the subject of the largest number of single portraits during the 1630s. Rembrandt posed her in mythological dress, particularly in the flower-draped abundance of the goddess Flora (1634), and in formal attire with fastidious profile. But he also delighted in using her as his subject in a cluster of spontaneous domestic drawings, such as those catching her leaning out of a window or lying in bed. Some of the unhappy biographical data make the

Marriage to Saskia Uylenburgh

scenes in bed poignant: Saskia lost three children before the birth of a son, Titus, in 1641. Saskia herself died only a year later.

With his marriage Rembrandt seemed to be at the summit of his potential, and his self-portraits of the 1630s radiate confidence and prosperity. The climax of this trend is the stately, pyramidal self-portrait of 1640, which at once borrows its nonchalant pose upon a balustrade from Titian's portrait, then thought to be the poet Ariosto, plus its rich costume and golden colour harmonies from Raphael's portrait of the quintessential courtier Baldassare Castiglione. Both of these pictures had passed through prosperous Amsterdam's art market in recent days, a fact which gave support to Rembrandt's response to Huygens that he need not go to Italy to study the great masters. In addition, this image after the Renaissance model shows not only Rembrandt's successful assimilation of the forms but also the courtier aspirations of Raphael, Titian, and, most recently, Rubens himself. Just as he had previously done with the Anslo portrait, Rembrandt also issued an etched version of this self-portrait on a stone sill in 1639.

If the velvet berets and fur-trimmed coats show Rembrandt to be a man luxuriating in his successes, a unique double portrait of Rembrandt and Saskia (*c.* 1635–36) seems to offer an ironic and reflective gaze at his life. Here, too, an etching echoes the subject of the painting, but in the 1636 etched double portrait Rembrandt shows himself at work, drawing as he looks up at the viewer. In the Dresden painting Saskia sits on the lap of a foppishly dressed Rembrandt, who gaily holds up a flagon of ale as he twists to offer a silly grin out of the picture. This tavern setting at once indulges a current "Arcadian" fashion for showing fashionable ladies as courtesans (yet another incarnation of the goddess Flora, with whom Rembrandt had already identified Saskia) as well as draws upon the pictorial tradition of the Prodigal Son with the tavern harlots. It is worth noting that the lavish dress of this couple offers an echo of the finery in the Kassel profile of Saskia, but here the suggestion of loose living and of future repentance from the Prodigal Son analogy also sounds a note of self-criticism.

As if to underscore the conspicuous consumption of the Dresden double portrait, Rembrandt purchased a large house in 1639 on the Jodenbreestraat. (This house is still known as the Rembrandt House, and today many of his etchings are kept there as a memorial.) In the same year Rembrandt painted a full-length portrait identified as Andries de Graeff, the brother of a leading Amsterdam patrician, Cornelis de Graeff, and son of the recent burgomaster, Jacob de Graeff. Cornelis' brother-in-law and fellow patrician, Frans Banning Cocq, headed the civic guard that was to grant Rembrandt the commission for what is one of his grandest and most famous pictures: the group portrait nicknamed the "Night Watch" (Amsterdam).

The "Night Watch"

The proper title of the "Night Watch" ought to be "The Militia Company of Captain Frans Banning Cocq." Executed between 1640 and 1642, it was one of six such group portraits of militia companies designed for the new hall of the Kloveniersdoelen, the musketeer branch of the civic militia (alongside crossbowmen and archers). Two of the other group portraits were painted by Rembrandt's pupil Flinck, and a third was entrusted to his emulator Backer. Most of the sitters in the "Night Watch" lived near one another in the drapers' section of Amsterdam and were men of wealth, headed by Banning Cocq and his lieutenant Willem van Ruytenburgh. These are the two striding figures who lead the march in the spotlighted foreground of the large painting. Like the anatomy lesson group portrait of a decade earlier, Rembrandt's militia group portrait differs from other Dutch instances by showing the entire scene animated and energized into a single, dynamic action. Emerging from in front of a large triumphal arch, the militia company seems to be setting out on a specific mission, evoking the glory days of the Dutch Revolution at the end of the preceding century. Scholars have debated whether in fact a specific ceremonial event was commemorated by this picture, such as the triumphal entry of the dowager French queen-mother, Marie de Médicis, in September 1638, one of the first state occasions for the young Dutch Republic. In any case, the militia company's action elicits comparison of the armed guards at the city gates in a past era with the present protection of the city's independence and privilege by her own citizens. Thus did Rembrandt in this largest of his pictures (even though trimmed seriously on its left side; originally about $4\frac{1}{2} \times 5$ metres) fuse history and portraiture, action and description. In the process he realized within an authentic Dutch pictorial tradition the same glories that Rubens had rendered in his equestrian portraits of reigning princes.

Experimentation and exploration. The contradictory experiences of Saskia's death and the completion of the "Night Watch" in 1642 produced a watershed in Rembrandt's life and artistic career. His connections at court had come to nothing late in the 1630s, and during the decade of the 1640s neither the sitters for his portraits nor the documented owners of his pictures reveal the kinds of social eminence of those of a decade earlier after his sudden arrival in Amsterdam (except for a couple of later submissions of religious pictures in 1646 to the court in The Hague). Legend has it that Rembrandt's patronage suffered on account of the boldness of invention in a work like the "Night Watch"; while this is mere fable, the taste, especially in portraits, for greater refinement and detail of execution was exactly the opposite of Rembrandt's own developing predilection. If he had delighted in the meticulous in his earlier portraits and in the later Leiden histories, Rembrandt now experimented with bolder brush strokes and stronger shadows with dark backgrounds. His followers, chiefly Bol, Flinck, and Backer, were becoming more successful and popular than their master, even among the highest levels of patronage.

Rembrandt's pupils

Another cluster of talented pupils worked with Rembrandt around 1640: Gerbrand van den Eeckhout (1621–74) from Amsterdam; Carel Fabritius (1622–54), later a leading painter in Delft; and Samuel van Hoogstraten (1627–78) of Dordrecht, later better known for his art theory, *Inleyding tot de hooge schoole der schilderkonst* (1678; "Introduction to the High School of Painting"). Each of these made his own variation on the theme of the master, by adding colour to the history groups (van den Eeckhout), by working with a lighter background and experimenting with perspective effects (Fabritius), and by producing illusionistic experiments (van Hoogstraten).

Dutch scenes

Rembrandt's art in general took on a more private and modest ambition during the 1640s. One measure of this development is provided by the comparison between his landscapes of the succeeding decades. During the 1630s the chief landscape efforts by Rembrandt took the form of paintings, which often show stormy skies and blasted trees. These wild pieces of nature respond to a heritage of powerful, stormy vistas, represented by several Rubens landscapes with narratives and in Dutch art by Hercules Seghers. During the early 1640s, however, Rembrandt turned increasingly to images with a more local, Dutch flavour, and he increasingly produced drawings and etchings rather than paintings of such scenes. Examples include the painted "Landscape with a Stone Bridge," the filigree etching of the Amsterdam skyline (*c.* 1641), the etched "Windmill" (1641), and several etched "Cottages" (1641/42). A climax of landscape etching was produced in 1643 with the "Three Trees." Here a stormy sky at the upper left provides the dramatic contrasts of light and dark that had characterized painted landscapes, but the silhouetted trees of the hillside offer a topographic complement to a rural Dutch scene dotted with miniature windmills and human figures at their labours (including a lonely artist at the upper right, sketching on top of the hill!). This is a virtuoso performance that includes reworkings of the plate in drypoint for the delicate clouds.

One cliché of scholarship avers that this turn toward landscapes in both drawings and etchings indicates Rembrandt's desire to find a peaceful alternative to the stresses in his life. More likely, these works meant a shift of marketing strategy away from the portraits and the kinds of histories that had been his staples during the halcyon days of the 1630s. When Rembrandt now created biblical histories, he omitted the Old Testament topics that had fascinated him earlier in favour of more conventional

themes from the infancy or mission of Christ. Some of the pictures, such as the 1648 "Christ at Emmaus" (Paris), already exemplify the quiet dignity and human vulnerability of Rembrandt's later spirituality, but a number of his other pictures exist in several versions, many of which were doubtless copies by associates, such as van den Eeckhout. Quite a few other paintings show sentimental single figures of young girls or old, bearded men without any of the force of the earlier figures and without their refined technique.

This fallow period of paintings is accompanied by a relative silence in the documents concerning Rembrandt during the 1640s. The exceptions are unpleasant. Despite Rembrandt's posthumous tribute to Saskia's former guardian, Sylvius, in the 1646 etching, the Uylenburghs made legal inquiries concerning Rembrandt's administration of Saskia's estate. In addition Rembrandt was having an affair with Titus' nurse, Geertghe Dircx, but at the end of the decade she sued him for breach of promise of marriage. The episode dragged on and ended badly, with a cash settlement by Rembrandt, followed by the commitment of Geertghe to a reformatory and her early death in 1656.

In light of the importance of painted portraits and the relative scarcity of etched portraits in Rembrandt's output, it surely must be significant that in the late 1640s and the 1650s the artist produced a series of spectacular etched likenesses of friends and associates: the landscape painter Jan Asselijn (*c.* 1647), the Jewish physician Ephraim Bonus (1647), the patrician and then burgomaster Jan Six (1647), the printseller Clement de Jonghe (1651), the municipal auctioneer Pieter Haaringh (1655), and the apothecary Abraham Francen (*c.* 1657). The personal nature of some of these etchings emerges when viewed against the fact that Rembrandt filed for bankruptcy in 1656, for Francen, an art collector, stood by Rembrandt during his financial crises, while Haaringh must have been the auctioneer of Rembrandt's goods in the winter of 1655/56. But the interest in these plates is not just biographical; they are among the richest and most complex of Rembrandt's etched work. The "Jan Six," in particular, is worked up with meticulous care; velvety darks, reworked in drypoint to allow the subtlest nuances, are broken only by the glaring white of the paper for the light-filled window where the elegant young man stands. A decade later this same formula with a window was repeated in the Francen etching. Similar control of lights emerging out of darks can be found in the portrait of Bonus. In many of these portraits numerous details in the plates were reworked to produce subsequent states, and both etching and drypoint were combined with burnishing to alter the linear or tonal qualities of figures and settings.

The "Hundred Guilder Print"

Etchings of religious subjects also came to occupy increasing importance in the output of Rembrandt in the later 1640s and early 1650s. The largest, most celebrated religious etching, now known as the "Hundred Guilder Print" because of its high prices, was produced over an extended period of time, *c.* 1643–49, and it shows the full richness of Rembrandt's etching technique and experimentation. The subject, as in many of the paintings of the 1640s, is Christ's ministry—healing the sick, receiving the little children, and preaching to the multitude. It is spread out across a huge wall like the one for the "Night Watch" or the multi-figural biblical episodes of the 1630s (*e.g.*, the "Doubting Thomas," 1634). In this large print, almost 30 × 40 centimetres, there are enormous variations in the techniques of different segments, creating a range of lights and darks, of descriptive detail or sketchiness, and of narrative focus. At the centre of the entire composition, almost dissolving in his delicately worked halo before the dark wall, is the tall figure of Christ. Rembrandt had made a number of painted sketches for the face of Christ from a single model, possibly a young Jewish man from his neighbourhood; replicas of these sketches doubtless served the same purpose as the *tronies* of the 1630s, that is, as models for biblical histories, and the 1656 inventory of Rembrandt's collection lists three such heads, two by him. This is also the same Christ featured in the 1648 "Christ at Emmaus," where the quiet and reflective apostles take the place of the crowd of the "Hundred Guilder Print."

The human component of religious stories became an increasing concern of the artist, and some of his most moving visions were developed in etchings of the early 1650s. As if a complement to the "Hundred Guilder Print," the etching of "Christ Preaching" (*c.* 1652), known as "La Petite Tombe," carries out the same theme within a more intimate and structured space, now marked by deep, dark passages of drypoint. Other etchings carry to the ultimate Rembrandt's fascination with figures emerging out of darkness. In contrast, however, to the drama of earlier biblical etching experiments, such as the 1634 "Annunciation to the Shepherds," these are sombre, meditative environments of quiet inwardness, such as the "Adoration of the Shepherds" (with lamp, *c.* 1654; at night, *c.* 1652). Most of these nocturnal visions concern themselves with the issue of divine revelation as a personal epiphany, often around the body of Christ in either infancy or Passion: "Entombment" (*c.* 1654) and "Descent from the Cross" (*c.* 1654), plus the dark "Presentation in the Temple" (*c.* 1654). In addition Rembrandt produced a pair of large-scale etchings that he reworked almost continuously during the mid-1650s as a kind of meditative exercise that produced increasingly focused attention on the figure of Christ and on the spiritual content of the scene—at the expense of the narrative richness of the earlier states. These two prints also depict key moments of Christ's Passion: "The Three Crosses" (1653; four states) and "Christ Presented to the People" (1655; six states). Most scholars seem to think that Rembrandt continued to work on the plate of the "Three Crosses" until around 1660, after which time he abandoned this most widespread and easily available of mediums entirely, except for a single commissioned portrait of 1665.

The other chief burst of pictorial activity during the 1650s was a continuation of the interest in landscape in both etching and painting. Again, most of the etchings show local, Dutch rural settings of gabled cottages and farm buildings, but some reflect renewed interest in monumental or picturesque buildings, such as towers or obelisks. One masterwork of etching is the oblong 1651 panorama, known as the "Goldweigher's Field," which was actually the estate, near Haarlem, of Christoffel Thijsz., the man who had sold Rembrandt his house 12 years earlier and was thus his creditor. Like the rendition of the delicate skyline of Amsterdam a decade earlier, this print has a refined sense of spatial gradation achieved through scale, tonal change (including rich foreground drypoint accents), and a curving perspective; its broad format replicates in etching the formulas used in Dutch landscape paintings by artists such as Jan van Goyen. In paintings Rembrandt's landscapes ("The Mill"; "River Landscape with Ruins") incorporate the bluer skies and more careful structures of the Italianate Dutch landscape painters, one of whom, Asselijn, was the subject of a portrait etching of about 1647.

Return to self-portraits

After a gap of nearly a decade Rembrandt also returned to self-portraits at the end of the 1640s, including his final etched self-portrait, which shows him emerging out of the shadows as he draws beside a window (1648; five states). This is a fleshier figure than the jaunty courtier of 1639, but the presence of his own portrait amid those of his friends and patrons might point to a renewed confidence. At a time when portrait patrons began to return Rembrandt also produced some self-portraits, notably the large, three-quarter-length 1652 likeness.

Rembrandt's art of the 1640s did not inspire many pupils or followers, although Nicolaes Maes of Dordrecht (1634–93) seems to have apprenticed in the early 1650s and to have assimilated some of the single figure types of young or old women in ruddy colours, but used in this case for moralizing on their behaviour.

Beginning in 1653 the documents reveal a steadily worsening financial picture for Rembrandt, and in 1656, after transferring most of his assets to Titus, he applied for bankruptcy. Because of the auction of his possessions an inventory of his collections was drawn up in 1656 prior to liquidation. Analysis of his collections reveals not only quite a number of his own works, including many

landscapes and animal pictures, but also a vast collection of prints and drawings by other artists, usually mounted in books. Of the Dutch masters owned by Rembrandt, Seghers and Lievens predominate, but the Fleming Adriaen Brouwer, the Dutchman Jan Porcellis, and numerous Italians are also present. The variety of other items, including *naturalia*, Orientalia, and weapons, have been related to the contemporary princely tradition of the encyclopaedic collection, or *Kunst- und Wunderkammer* (arts and natural wonders room), a kind of personal museum. Here, as in the very subjects of his pictures and his early education in Leiden, Rembrandt is seen as a cosmopolitan, learned, and ambitious individual with broad interests.

Final years. Rembrandt's fame extended far beyond Holland. His "Aristotle Contemplating the Bust of Homer" (1653) was produced for a Sicilian, Don Antonio Ruffo. The imaginative classical subject, incorporating the worlds of philosopher, poet, and prince into a single work, seems to have been Rembrandt's own idea. This work not only seems to glow with its own inner light, coming out of a murky darkness, but it also shows Rembrandt's strength at painting character heads in meditative reverie. In 1660 Ruffo asked for a companion picture, and Rembrandt delivered an "Alexander" along with a "Homer" (1663; preserved only in a fragment).

"Aristotle Contemplating the Bust of Homer"

Closer to home Rembrandt received an important large commission for another anatomy lesson by Dr. Jan Deyman (fragment, 1656). The well-known "Syndics of the Drapers' Guild" (1662) was a final group portrait, painted for the sampling officials closely connected with the musketeers' company shown in the "Night Watch." This work is at once filled with the lively animation of Rembrandt's other groups, but it conforms more closely to the conventions of clustered yet balanced rows of posed heads that were the rule for such assignments.

The grandest commissions in Amsterdam, however, were the pictures for the new and grand Town Hall, and these were going to Rembrandt's circle—to Lievens, Bol, and especially Flinck, rather than to Rembrandt himself. Rembrandt's only venture, "The Conspiracy of Julius Civilis" (fragment, *c.* 1661–62), depicted a heroic moment of Dutch revolt against the ancient Romans rendered as a nocturnal dinner oath by suffused candlelight. But Rembrandt's painting of the one-eyed chief conspirator, Julius Civilis, and his coarse band was only briefly hung in 1662 and never paid for, probably because it lacked the proper heroic decorum to be found in works by Flinck and Bol. Thus, two decades after the triumph of the Kloveniersdoelen militia commissions shared with his former students Rembrandt lost out to them on the success and the rewards of his hometown's principal adornment.

Rembrandt did share his later years with someone: Hendrickje Stoffels (*c.* 1615–63) had become his live-in companion after Geertghe Dircx, and she was even more featured as his subject than Saskia had been. The full-bodied nude of the 1654 "Bathsheba" has traditionally been identified with Hendrickje and accords well with her numerous portraits and studies (*e.g.*, "Woman Bathing," 1655). In 1654, shortly before she bore a daughter, Cornelia, Hendrickje was officially censured by the church council for living in sin. At the same time Titus, who held Rembrandt's assets in his name, became the subject of numerous portraits before he predeceased his father in 1668, just prior to his 27th birthday.

Hendrickje Stoffels

Despite financial strains and lack of prominent commissions Rembrandt was able to summon his energies in his final decade of creativity to produce the kinds of soulful pictures for which he is best loved today. He returned to biblical histories but gave them the portrait-like, stilled power of personality that was already present in the "Aristotle." Beginning with "Jacob Blessing the Sons of Joseph" (1656), family drama became the subject, as it did for the "Return of the Prodigal Son." In this later work, unfinished at Rembrandt's death, the sombre figures seem to emerge with their own glow out of a gloomy background, and each turns his gaze inward in quiet reflection. Some works so closely resemble portraits that their religious identity is submerged or even lost; such is the case of the "Jewish Bride" (*c.* 1664; Amsterdam), where the husband's tender embrace of his wife dominates over the thickly painted exotic, rich, gold and vermilion costumes of the figures. A rare family portrait in modern costume from the late 1660s (Braunschweig) shows close affinity with the Amsterdam couple and prompts the question of whether the couple of the "Jewish Bride" might not be biblical but rather a costumed double portrait in historical guise, akin to the portrait of the musketeers of the "Night Watch."

Personal isolation and internal anguish dominate Rembrandt's interpretations of other historical figures: "Peter Denying Christ" (1660) and "Lucretia" (1664, 1666). A series of half-length, portrait-like apostles of 1661, especially the "Self Portrait as the Apostle Paul" (1661), completes the overlap between contemporary individuals and historical subjects in Rembrandt's oeuvre. These were painted at the same time as the "Homer" and the "Alexander" exported to Ruffo in Sicily and culminate the *tronies* tradition of model heads of historical figures.

In light of his own financial woes during this later period, Rembrandt could not sustain a large workshop or circle of pupils. Yet Aert de Gelder of Dordrecht (*c.* 1645–1727) spent at least a couple of years with Rembrandt around 1660 after training initially with van Hoogstraten. De Gelder, too, chiefly created half-length historical pictures in the manner of Rembrandt in the 1660s, and his "Self-Portrait as the Painter Zeuxis" (1685; Frankfurt) has been related to the late, laughing self-portrait by Rembrandt (*c.* 1668–69; Cologne), where the aged artist seems to smirk at his destiny as well as at the stern figure alongside him.

Rembrandt was not without patrons for portraits during his final decade, and some of his most moving likenesses were produced then with the kind of stillness and empathy that he also brought to his single-figure, biblical characters. Included among his sitters were the wealthy Trip family, also originally from Dordrecht and patrons of Bol. Rembrandt had painted mother and daughter of the Elias Trip clan in 1639, and in 1661 he painted Elias' brother, Jacob, and his wife, Marguerite de Geer. Most other sitters for the late works of Rembrandt are unknown, except for the classicizing artist Gerard de Lairesse (1665) and the religious poet Jeremias de Decker (1666), a close friend. Yet there are impressive but anonymous pendants (especially in New York and Washington) plus a family group (Braunschweig) showing thick, colourful impasto applied with a palette knife.

Finally, the last dozen years of his life were a fertile period for self-portraits. In addition to using his face for St. Paul in the apostles series of 1661, Rembrandt shows himself in another larger half-length portrait of around that time (Kenwood House) with palette in hand and two large arcs on the background to record his mastery of geometry (possibly also an allusion to the hemispheres of a world map in the conventions of Dutch cartography). Two further canvases record the artist in his final year, 1669. The first (London) presents the pyramidal dignity of his early maturity; the second (The Hague) combines an exotic turban, as seen in the early *tronies*, with a dispassionate, almost clinical directness in recording his firm gaze amidst his sagging flesh.

Last self-portraits

Rembrandt died on Oct. 4, 1669, and was buried in Amsterdam's Westerkerk.

Reputation. With the ascendance of a classicizing taste for clear compositions and handsome protagonists, Rembrandt was eclipsed even during his lifetime by his own pupils, particularly by Flinck and Bol. After an ambitious period, when he emulated the models of Lastman, van Honthorst, and Rubens, Rembrandt himself gave up the attempt to follow international currents, although he never renounced the serious themes of biblical subjects. Rembrandt has been lauded for his spiritual qualities, and his art shows contact with orthodox Calvinists as well as with Mennonites, Remonstrants, and even Jews in tolerant Amsterdam. Although he enjoyed patrician patronage from the de Graeff circle and painted important group portraits throughout his career, his art never held sway at the peak of Amsterdam fashion after the 1630s, though it never was fully eclipsed.

Rembrandt was never really forgotten, but he was eagerly rediscovered by artists during the 19th century, who found

in his works echoes of many of their own Romantic strivings for independent formal means and experiments with the depiction of character and passion. His etchings, which continued to exert considerable influence over later practitioners of that medium, such as Giovanni Castiglione in 17th-century Genoa or Giovanni Battista Tiepolo in 18th-century Venice, also came back into prominence with the etching revival in the mid-19th century. His emphasis on the natural rather than the beautiful accorded well with the credo of a painter such as Courbet, who strove to be "modern" in protest against French academic training. Thus, Rembrandt acquired the mantle of the Romantic hero—of the individual following his own inner light.

The issue of attribution

Modern scholarship has followed several leads in evaluating Rembrandt. Debates continue over precisely which works belong to him and which can be assigned to pupils, followers, imitators, or even later forgers. Scientific investigations in the laboratory have added to the data necessary for drawing such conclusions, but the recent studies of Rembrandt's pupils have at once given their own characters sharper focus while also serving as a reminder of how little one can define their precise relationship to Rembrandt while they were working in his studio. The study of workshop working method remains a crucial task for sorting out authentic Rembrandt paintings or drawings from the inauthentic works or works that seem to have at least partial contributions from his students. This latter category—consisting of works formerly ascribed to Rembrandt but not yet fully attributable to the defined oeuvres of his individual students—remains a growing corpus that will occupy scholars for years to come.

The integration of Rembrandt's works into a larger and longer visual tradition has also been a recent project of scholarship, which has determined that many of Rembrandt's religious subjects conform to earlier models, especially from Dutch prints. In addition, the less familiar Dutch tradition of ambitious historical pictures for town halls, palaces, and wealthy patrons has been clarified, in light of which Rembrandt no longer seems like quite such an anomaly.

Finally, major discoveries concerning Rembrandt's relationship to his patrons and to contemporary politics, economics, and social structure in Amsterdam have provided insights into this artist's role within the culture of his day. It is now possible to compare his achievement with that of rivals and colleagues, such as Lievens, Flinck, and Bol, by considering patrons and status as well as by drawing on modern evaluations of their respective ambitions. In this light the fact that Rembrandt was himself a collector and a student of past art from both Italy and Holland acquires greater meaning, even as it dispels the persistent myth of the isolated genius.

Thus, the current image of Rembrandt ties him more closely to his environment in various ways: through working method, favoured subjects, and patronage within the Dutch urban culture. Without denying the creativity and the distinctive personal sensitivity that has always been the basis of Rembrandt's popularity, modern interest focuses on Rembrandt and his art in terms of their meaning for 17th-century Holland, for it is only by comparing his accomplishments to those of his fellow artists that one can truly grasp the achievement of this beloved Dutch master.

MAJOR WORKS

Paintings

PORTRAITS: "Self Portrait" (*c.* 1628; Rijksmuseum, Amsterdam); "Self-Portrait" (1629; Isabella Stewart Gardner Museum, Boston); "Old Woman Praying" (*c.* 1630; Residenz Gallery, Salzburg, Austria); "Old Man with a Gold Chain" (*c.* 1631; Art Institute of Chicago); "Nicolaes Ruts" (1631; Frick Collection, New York City); "Amalia van Solms" (1632; Jacquemart-André Museum, Paris); "Anatomy Lesson of Dr. Nicolaes Tulp" (1632; Mauritshuis, The Hague); "Marten Looten" (1632; Los Angeles County Museum of Art); "Johannes Uytenbogaert" (1633; private collection, England); "Johannes Elison" (1634; Museum of Fine Arts, Boston); "Saskia as Flora" (1634; Hermitage, St. Petersburg); "Saskia in Profile" (*c.* 1634; State Art Collections, Kassel, Ger.); "Rembrandt and Saskia as the Prodigal Son" (*c.* 1635–36; Picture Gallery, Dresden, Ger.); "Andries de Graeff" (1639; State Art Collections, Kassel); "Self-Portrait" (1640; National Gallery, London); "Cornelis Anslo and Aeltje Schouten" (1641; Picture Gallery, Dahlem Museums, Berlin); "The Militia Company of Captain Frans Banning Cocq" ("Night Watch," 1642; Rijksmuseum); "Self-Portrait" (1652; Kunsthistorisches Museum, Vienna); "Anatomy Lesson of Dr. Jan Deyman" (1656; Rijksmuseum); "Self-Portrait" (1658; Frick Collection); "Self-Portrait as the Apostle Paul" (1661; Rijksmuseum); "Self-Portrait" (*c.* 1661–62; The Iveagh Bequest, Kenwood House, London); "Syndics of the Drapers' Guild" ("De Staalmeesters," 1662; Rijksmuseum); "The Jewish Bride" (*c.* 1664; Rijksmuseum); "Gerard de Lairesse" (1665; Lehman Collection, Metropolitan Museum of Art, New York City); "Jeremias de Decker" (1666; Hermitage); "Family Group" (*c.* 1666–68; Herzog Anton Ulrich Museum, Braunschweig, Ger.); "Laughing Self-Portrait" (*c.* 1668–69; Wallraf-Richartz Museum, Cologne); "Self-Portrait" (1669; National Gallery, London); "Self-Portrait" (1669; Mauritshuis).

RELIGIOUS PAINTINGS: "Stoning of St. Stephen" (1625; Museum of Fine Arts, Lyon); "Balaam's Ass and the Angel" (1626; Cognacq-Jay Museum, Paris); "Baptism of the Eunuch" (1626; St. Catherine's Convent National Museum, Utrecht, Neth.); "Presentation in the Temple" (*c.* 1627–28; Hamburg Gallery); "Christ at Emmaus" (*c.* 1628; Jacquemart-André Museum); "Capture of Samson" (*c.* 1628; Picture Gallery, Dahlem Museums); "Judas Returning the Thirty Pieces of Silver" (1629; private collection, England); "Raising of Lazarus" (*c.* 1630; Los Angeles County Museum of Art); "Jeremiah Lamenting the Destruction of Jerusalem" (1630; Rijksmuseum); "Raising of the Cross" (*c.* 1633; Alte Pinakothek, Munich); "Descent from the Cross" (*c.* 1633; Alte Pinakothek); "Christ Before Pilate" (1634; National Gallery, London); "John the Baptist Preaching" (*c.* 1634; Picture Gallery, Dahlem Museums); "Belshazzar's Feast" (*c.* 1635; National Gallery, London); "Sacrifice of Isaac" (1635; Hermitage); "Sacrifice of Isaac" (1636; Alte Pinakothek); "Holy Family with Angels" (1645; Hermitage); "Christ at Emmaus" (1648; Louvre, Paris); "Bathsheba" (1654; Louvre); "Jacob Blessing the Sons of Joseph" (1656; State Art Collections, Kassel); "Peter Denying Christ" (1660; Rijksmuseum); "Return of the Prodigal Son" (Hermitage).

HISTORY AND MYTHOLOGY: "Palamedes Before Agamemnon" (1626; Municipal Museum [De Lakenhal], Leiden, Neth.); "Andromeda" (*c.* 1630; Mauritshuis); "Pluto and Proserpina" (*c.* 1632; Picture Gallery, Dahlem Museums); "Rape of Europa" (*c.* 1632; private collection, New York City); "Rape of Ganymede" (1635; Picture Gallery, Dresden); "Danaë" (1636; Hermitage); "Aristotle Contemplating the Bust of Homer" (1653; Metropolitan Museum of Art); "Conspiracy of Julius Civilis" (*c.* 1661–62; National Museum, Stockholm); "Homer" (1661/63; Mauritshuis); "Lucretia" (1664; National Gallery of Art, Washington, D.C.); "Lucretia" (1666; Minneapolis Institute of Arts, Minnesota).

LANDSCAPES: "Landscape with Storm" (*c.* 1638; Herzog Anton Ulrich Museum, Braunschweig, Ger.); "Landscape with Good Samaritan" (1638; Czartoryski Collection, Kraków, Pol.); "Landscape with a Stone Bridge" (*c.* 1638; Rijksmuseum); "Mill" (*c.* 1650; National Gallery of Art, Washington, D.C.); "Winter Landscape" (*c.* 1650/55; State Art Collections, Kassel).

GENRE, FIGURES: "Music Lesson" (1626; Rijksmuseum); "Young Painter in the Studio" (*c.* 1629; Museum of Fine Arts, Boston); "Woman Bathing (Hendrickje Stoffels?)" (1655; National Gallery, London).

Etchings

"Self-Portrait" (1629); "Raising of Lazarus" (*c.* 1632); "Descent from the Cross" (1633); "Annunciation to the Shepherds" (1634); "Johannes Uytenbogaert" (1635); "Christ Before Pilate" (1636); "Menasseh ben Israel" (1636); "Self-Portrait with Saskia" (1636); "Self-Portrait" (1639); "Cornelis Anslo" (1641); "Skyline with Amsterdam" (*c.* 1641); "Windmill" (1641); "Three Trees" (1643); "Johannes Cornelisz. Sylvius" (1646); "Jan Asselijn" (*c.* 1647); "Ephraim Bonus" (1647); "Jan Six" (1647); "Self-Portrait by a Window" (1648); "Christ Healing the Sick (The Hundred Guilder Print)" (*c.* 1643–49); "Goldweigher's Field" (1651); "Clement de Jonghe" (1651); "Christ Preaching" (*c.* 1652); "Three Crosses" (1653ff); "Adoration of the Shepherds" (*c.* 1652/54); "Entombment" (*c.* 1654); "Descent from the Cross" (*c.* 1654); "Christ Presented to the People" (1655ff); "Abraham Francen" (*c.* 1657).

Drawings

"Self-Portrait" (*c.* 1627–28; British Museum, London); "Seated Old Man" (*c.* 1631; Print Room, Dahlem Museums); "Woman Bathing" (*c.* 1631; British Museum); "Saskia Looking out of a Window" (*c.* 1635; Boymans-van Beuningen Museum, Rotterdam); "Study After Leonardo da Vinci's 'Last Supper'" (*c.* 1635; British Museum); "Woman Carrying Child" (*c.* 1636; Pierpont Morgan Library, New York City); "Saskia in Bed" (*c.* 1635–38; State Graphics Collection, Munich); "Rear View of Woman in North Holland Costume" (*c.* 1638; Teylers Museum, Haarlem, Neth.); "Portrait of Titia van Uylenburg" (1639; National Museum, Stockholm); "Youth Pulling a Rope" (*c.* 1645;

Rijksmuseum); "Standing Male Nude" (*c.* 1646; Albertina, Vienna); "View of River IJ near Amsterdam" (*c.* 1649–50; Chatsworth Settlement); "View of Amstel River" (*c.* 1648–50; Print Room, Rijksmuseum); "View of Haarlem" (*c.* 1651; Boymans-van Beuningen Museum); "Lion" (*c.* 1650–52; Louvre); "Ruins of Old Town Hall in Amsterdam" (1652; Rembrandt House Museum, Amsterdam); "Four Orientals Beneath a Tree" (*c.* 1655; British Museum); "Sleeping Woman" (*c.* 1655; British Museum); "Woman at Window" (*c.* 1655; Louvre); "Nathan Admonishing David" (*c.* 1655–56; Metropolitan Museum of Art); "Row of Windmills" (*c.* 1655; Print Collection, State Art Museum, Copenhagen); "Anatomy Lesson of Dr. Jan Deyman" (*c.* 1656; Print Room, Rijksmuseum); "Portrait of a Man" (*c.* 1655–60; Louvre); "Female Nude on a Stool" (*c.* 1658; Art Institute of Chicago); "Christ on the Mount of Olives" (*c.* 1656–58; Hamburg Gallery); "Study for the Conspiracy of Julius Civilis" (1661; State Graphics Collection, Munich).

BIBLIOGRAPHY

Life and work: ARNOLD HOUBRAKEN, *De Groote schouburgh der nederlantsche konstschilders en schilderessen,* 2nd ed., 3 vol. (1753, reissued in 1 vol., 1976), is a long biography containing many anecdotes as well as facts that became the basis for 18th- and 19th-century criticism of Rembrandt's works. C. HOFSTEDE DE GROOT, *Die Urkunden über Rembrandt* (1575–1721), new ed. (1906), is a basic source that contains all documents concerning Rembrandt known at the time of publication, with annotations. WALTER L. STRAUSS and MARJON VAN DER MEULEN (compilers), *The Rembrandt Documents* (1979), is a collection of primary records translated into English. Other important examinations of the artist's life and work include OTTO BENESCH, *Rembrandt: Werk und Forschung,* corrected and enlarged ed., edited by EVA BENESCH (1970), a detailed analysis of Rembrandt's life and oeuvre, including a bibliography; JAKOB ROSENBERG, *Rembrandt: Life & Work,* rev. ed. (1964, reissued 1989), a survey and catalog; CHRISTOPHER WHITE, *Rembrandt and His World* (1964, reissued 1966), a study of the artist's life and environment, and *Rembrandt* (1984), a general study that draws on later research; BOB HAAK, *Rembrandt: His Life, His Work, His Time* (1969; originally published in Dutch, 1968), an illustrated biography with a discussion of the period in Dutch history, including documents concerning Rembrandt's life; CHRISTIAN TÜMPEL, *Rembrandt in Selbstzeugnissen und Bilddokumenten* (1977), a study of biographical and autobiographical documents; KENNETH CLARK, *An Introduction to Rembrandt* (1978), a biography for the general reader; GARY SCHWARTZ, *Rembrandt: His Life, His Paintings* (1985, reissued 1991; originally published in Dutch, 1984); and MARIËT WESTERMANN, *Rembrandt* (2000), a general introduction.

For studies of the artist's paintings, see C. HOFSTEDE DE GROOT, *Rembrandt,* vol. 6 of his *A Catalogue Raisonné of the Works of the Most Eminent Dutch Painters of the Seventeenth Century,* 8 vol., trans. from German (1907–27, reissued 1983); KURT BAUCH, *Rembrandt Gemälde* (1965); HORST GERSON, *Rembrandt Paintings,* trans. from Dutch, and ed. by GARY SCHWARTZ (1968, reprinted 1978), which offers a general discussion of Rembrandt's development and position in Dutch art, including a catalog of the paintings; A. BREDIUS, *Rembrandt: The Complete Edition of the Paintings,* 4th ed., rev. by HORST GERSON (1971; originally published in Dutch, 1935), a book in which the authors reject a number of works that are generally attributed to the artist and provide an understanding of many aspects of the artist's personality and oeuvre; SALLY SALVESEN (ed.), *Rembrandt: The Master & His Workshop,* 2 vol. (1991); HUBERT VON SONNENBERG *et al.*, *Rembrandt/Not Rembrandt in The Metropolitan Museum of Art: Aspects of Connoisseurship,* 2 vol. (1995); and J. BRUYN *et al.*, *A Corpus of Rembrandt Paintings* (1982–).

For the artist's etchings, see ARTHUR M. HIND, *A Catalogue of Rembrandt's Etchings, Chronologically Arranged and Completely Illustrated,* 2nd ed., 2 vol. (1923, reprinted in 1, 1967); LUDWIG MÜNZ (ed.), *Rembrandt's Etchings: Reproductions of the Whole Original Etched Work,* complete ed., 2 vol. (1952); GEORGE BIÖRKLUND, *Rembrandt's Etchings, True and False,* 2nd ed. rev. and enlarged (1968, reprinted 1988), which has a discussion of the various states of the works; CHRISTOPHER WHITE and KAREL G. BOON, *Rembrandt's Etchings: An Illustrated Critical Catalogue,* 2 vol. (1969); CHRISTOPHER WHITE, *Rembrandt as an Etcher: A Study of the Artist at Work,* 2 vol. (1969), a discussion that emphasizes the artist's technique and the connections between the etchings, drawings, and paintings; and GARY SCHWARTZ (ed.), *Rembrandt: All the Etchings Reproduced in True Size* (1977, reissued with corrections as *The Complete Etchings of Rembrandt: Reproduced in Original Size,* 1994).

For examinations of the artist's drawings, see OTTO BENESCH, *Rembrandt as a Draughtsman: An Essay* (1960), an introduction that comments upon the discovered additions to the artist's drawing oeuvre; CHRISTOPHER WHITE, *The Drawings of Rembrandt,* 2nd ed. (1966); OTTO BENESCH, *The Drawings of Rembrandt,* enlarged ed., edited by EVA BENESCH, 6 vol. (1973); BEN BROOS, *Rembrandt en tekenaars uit zijn omgeving* (1981); and HOLM BEVERS, PETER SCHATBORN, and BARBARA WELZEL, *Rembrandt, The Master & His Workshop: Drawings and Etchings* (1991).

Themes and criticism: Important critical studies include SEYMOUR SLIVE, *Rembrandt and His Critics, 1630–1730* (1953, reprinted 1988), a study of critical and biographical works on the artist; J.A. EMMENS, *Rembrandt en de regels van de kunst* (1968, reissued 1979), a work demolishing the image of Rembrandt that was created by the 17th-century classicistic critics and that continued to influence scholarship well into the 20th century, with a summary in English; JULIUS S. HELD, *Rembrandt's Aristotle, and Other Rembrandt Studies* (1969), a collection of the author's iconological studies of Rembrandt; E. HAVERKAMP-BEGEMANN, "The Present State of Rembrandt Studies," *Art Bulletin,* 53:88–104 (March 1971), an overview of modern scholarship; OTTO BENESCH, "Rembrandt," vol. 1 of his *Collected Writings,* 4 vol., ed. by EVA BENESCH (1970–73), a collection of articles; H. VAN DE WAAL, *Steps Towards Rembrandt: Collected Articles 1937–1972,* trans. from Dutch (1974;; ANTHONY BAILEY, *Responses to Rembrandt* (1994), which explores the debate over which works credited to Rembrandt are actually authentic; and ALBERT BLANKERT, *Rembrandt: A Genius and His Impact* (1997).

Important thematic studies include KENNETH CLARK, *Rembrandt and the Italian Renaissance* (1966, reissued 1968), which discusses antique and Renaissance influences on Rembrandt's style; R.W. SCHELLER, "Rembrandt en de encyclopedische verzameling," *Oud Holland,* 84(2–3):81–147 (1969), an interpretation of Rembrandt's art collecting, with a summary in English; E. HAVERKAMP-BEGEMANN, *Rembrandt: The Nightwatch* (1982); CHRISTOPHER WRIGHT, *Rembrandt: Self-Portraits* (1982), an illustrated interpretation of the self-portraits; SVETLANA ALPERS, *Rembrandt's Enterprise: The Studio and the Market* (1988, reissued 1990); BOB VAN DEN BOOGERT *et al.*, *Rembrandt's Treasures* (1999), which discusses the work by the artist in the Rembrandthuis in Amsterdam; ANN JENSEN ADAMS (ed.), *Rembrandt's Bathsheba Reading King David's Letter* (1998); CHRISTOPHER WHITE and QUENTIN BUVELOT, *Rembrandt by Himself* (1999), an examination of the self-portraits; and ERNST VAN DE WETERING, *Rembrandt: The Painter at Work* (1997, reissued 2000), an examination of the artist's technique and materials.

Much has been written about Rembrandt's influence on his pupils and contemporaries. Important studies include R.H. FUCHS, *Rembrandt in Amsterdam* (1969; originally published in Dutch, 1968), which offers essays on the artist's connections with the city and with the art of his contemporaries; WOLFGANG STECHOW, "Some Observations on Rembrandt and Lastman," *Oud Holland,* 84(2–3):148–162 (1969); E. HAVERKAMP-BEGEMANN and J.R. JUDSON, *Rembrandt After Three Hundred Years: An Exhibition of Rembrandt and His Followers* (1969), an exhibition catalog containing a discussion on Rembrandt as a teacher; ART INSTITUTE OF CHICAGO, *Rembrandt After Three Hundred Years: A Symposium, Rembrandt and His Followers* (1973), published in conjunction with the aforementioned exhibition; STEDELIJK MUSEUM DE LAKENHAL, *Geschildert tot Leyden anno 1626* (1976), an exhibition catalog of works by the artist and his contemporaries; A.B. DE VRIES, MAGDI TÓTH-UBBENS, and W. FROENTJES, *Rembrandt in the Mauritshuis: An Interdisciplinary Study,* trans. from Dutch (1978), an analysis of several works of Rembrandt in conjunction with works of his contemporaries in The Hague museum; ALBERT BLANKERT *et al.*, *Gods, Saints, & Heroes: Dutch Painting in the Age of Rembrandt* (1980), and *The Impact of a Genius: Rembrandt, His Pupils, and Followers in the Seventeenth Century* (1983); ALBERT BLANKERT, *Ferdinand Bol (1616–1680): Rembrandt's Pupil* (1982; originally published in Dutch, 1976); WERNER SUMOWSKI (ed.), *Gemälde der Rembrandt-Schüler* (1983–), a multivolume analytical catalog of paintings of the Rembrandt school; and WERNER SUMOWSKI, *Drawings of the Rembrandt School,* ed. and trans. from German by WALTER L. STRAUSS (1979–), another multivolume work.

(L.A.Si.)

Reproduction and Reproductive Systems

In a general sense reproduction is one of the most important concepts in biology: it means making a copy, a likeness, and thereby providing for the continued existence of species. Although reproduction is often considered solely in terms of the production of offspring in animals and plants, the more general meaning has far greater significance to living organisms. To appreciate this fact, the origin of life and the evolution of organisms must be considered. One of the first characteristics of life that emerged in primeval times must have been the ability of some primitive chemical system to make copies of itself.

At its lowest level, therefore, reproduction is chemical replication. As evolution progressed, cells of successively higher levels of complexity must have arisen, and it was absolutely essential that they had the ability to make likenesses of themselves. In unicellular organisms, the ability of one cell to reproduce itself means the reproduction of a new individual; in multicellular organisms, however, it means growth and regeneration. Multicellular organisms also reproduce in the strict sense of the term—that is, they make copies of themselves in the form of offspring—but they do so in a variety of ways, many involving complex organs and elaborate hormonal mechanisms.

For a depiction of some of the structures that make up the human reproductive system, shown in relation to other parts of the gross anatomy, see the colour TransVision in the *Propædia*: Part Four, Section 421.

The article is divided into the following sections:

General features of reproduction 609
- The nature of reproduction 609
  - Levels of reproduction
  - Natural selection and reproduction
- The process of fertilization 612
  - Maturation of the egg
  - Events of fertilization
  - Biochemical analysis of fertilization

Plant reproduction 615
- Plant reproductive systems 615
  - General features of asexual systems
  - General features of sexual systems
  - Bryophyte reproductive systems
  - Tracheophyte reproductive systems
  - Variations in reproductive cycles
  - Physiology of plant reproduction
- Pollination 624
  - Types: self-pollination and cross-pollination
  - Agents of pollen dispersal
- Seed and fruit 628
  - The nature of seeds and fruits
  - Form and function
  - Agents of dispersal
  - Germination

Animal reproduction 636
- Reproductive systems of invertebrates 637
  - Gonads, associated structures, and products
  - Mechanisms that aid in the union of gametes
  - Parthenogenesis
  - Provisions for the developing embryo
- Reproductive systems of vertebrates 642
  - Gonads, associated structures, and products
  - Adaptations for internal fertilization
  - Role of gonads in hormone cycles
  - Provisions for the developing embryo

The human reproductive system 650
- The male reproductive system 651
  - External genitalia
  - Structures of the sperm canal
  - Accessory organs
- The female reproductive system 653
  - External genitalia
  - Internal structures
  - The menstrual cycle
  - Menopause
- Diseases of the human reproductive system 658
  - Genetic and congenital abnormalities
  - Functional genital disorders
  - Infectious diseases spread by sexual contact
  - Other infections affecting the reproductive system
  - Structural changes of unknown causes
  - Tumours

Human reproduction from conception to birth 664
- The normal events of pregnancy 664
  - Initiation of pregnancy
  - Diagnosis of pregnancy
  - Anatomic and physiologic changes of normal pregnancy
  - Prenatal care and testing
- Abnormal changes in pregnancy 673
  - Ectopic pregnancy
  - Abortion
  - Systemic diseases and pregnancy
  - Diseases of pregnancy
- Parturition: the process of birth 680
  - The stages of labour
  - Relief of pain in labour
  - Natural childbirth
  - Operative obstetrics
  - Accidents during labour
- Puerperium, or period of involution 685
  - Lactation
  - Weaning and the cessation of lactation

Bibliography 686

GENERAL FEATURES OF REPRODUCTION

The nature of reproduction

LEVELS OF REPRODUCTION

Transmittal of genetic information

Molecular replication. The characteristics that an organism inherits are largely stored in cells as genetic information in very long molecules of deoxyribonucleic acid (DNA). In 1953 it was established that DNA molecules consist of two complementary strands, each of which can make copies of the other. The strands are like two sides of a ladder that has been twisted along its length in the shape of a double helix (spring). The rungs, which join the two sides of the ladder, are made up of two terminal bases. There are four bases in DNA: thymine, cytosine, adenine, and guanine. In the middle of each rung a base from one strand of DNA is linked by a hydrogen bond to a base of the other strand. But they can pair only in certain ways: adenine always pairs with thymine, and guanine with cytosine. This is why one strand of DNA is considered complementary to the other.

The double helices duplicate themselves by separating at one place between the two strands and becoming progressively unattached. As one strand separates from the other, each acquires new complementary bases until eventually each strand becomes a new double helix with a new com-

plementary strand to replace the original one. Because adenine always falls in place opposite thymine and guanine opposite cytosine, the process is called a template replication—one strand serves as the mold for the other. It should be added that the steps involving the duplication of DNA do not occur spontaneously; they require catalysts in the form of enzymes that promote the replication process.

Genetic code

Molecular reproduction. The sequence of bases in a DNA molecule serves as a code by which genetic information is stored. Using this code, the DNA synthesizes one strand of ribonucleic acid (RNA), a substance that is so similar structurally to DNA that it is also formed by template replication of DNA. RNA serves as a messenger for carrying the genetic code to those places in the cell where proteins are manufactured. The way in which the messenger RNA is translated into specific proteins is a remarkable and complex process. (For more detailed information concerning DNA, RNA, and the genetic code, see the articles BIOCHEMICAL COMPONENTS OF ORGANISMS: *Nucleic acids;* GENETICS AND HEREDITY: *The gene*).

The ability to synthesize enzymes and other proteins enables the organism to make any substance that existed in a previous generation. Proteins are reproduced directly; however, such other substances as carbohydrates, fats, and other organic molecules found in cells are produced by a series of enzyme-controlled chemical reactions, each enzyme being derived originally from DNA through messenger RNA. It is because all of the organic constituents made by organisms are derived ultimately from DNA that molecules in organisms are reproduced exactly by each successive generation.

Cell reproduction. The chemical constituents of cytoplasm (that part of the cell outside the nucleus) are not resynthesized from DNA every time a cell divides. This is because each of the two daughter cells formed during cell division usually inherits about half of the cellular material from the mother cell (see CELLS: *Cell division*), and is important because the presence of essential enzymes enables DNA to replicate even before it has made the enzymes necessary to do so.

Cells of higher organisms contain complex structures, and each time a cell divides the structures must be duplicated. The method of duplication varies for each structure, and in some cases the mechanism is still uncertain. One striking and important phenomenon is the formation of a new membrane. Cell membranes, although they are very thin and appear to have a simple form and structure, contain many enzymes and are sites of great metabolic activity. This applies not only to the membrane that surrounds the cell but to all the membranes within the cell. New membranes, which seem to form rapidly, are indistinguishable from old ones.

Thus, the formation of a new cell involves the further synthesis of many constituents that were present in the parent cell. This means that all of the information and materials necessary for a cell to reproduce itself must be supplied by the cellular constituents and the DNA inherited from the parent cell.

Binary fission. Of the various kinds of cell division, the most common mode is binary fission, the division of a cell into two separate and similar parts. In bacteria (prokaryotes) the chromosome (the body that contains the DNA and associated proteins) replicates and then divides in two, after which a cell wall forms across the elongated parent cell. In higher organisms (eukaryotes) there is first an elaborate duplication and then a separation of the chromosomes (mitosis), after which the cytoplasm divides in two. In the hard-walled cells of higher plants, a median plate forms and divides the mother cell into two compartments; in animal cells, which do not have a hard wall, a delicate membrane pinches the cell in two, much like the separation of two liquid drops. Budding yeast cells provide an interesting exception. In these fungi the cell wall forms a bubble that becomes engorged with cytoplasm until it is ultimately the size of the original cell. The nucleus then divides, one of the daughter nuclei passes into the bud, and ultimately the two cells separate.

Mitosis and meiosis

In some instances of binary fission, there may be an unequal cytoplasmic division with an equal division of the chromosomes. This occurs, in fact, in a large number of higher organisms during meiosis—the process by which sex cells (gametes) are formed: originally each chromosome of the cell is in a pair (diploid); during meiosis these diploid pairs of chromosomes are separated so that each sex cell has only one of each pair of chromosomes (haploid). During the two successive meiotic divisions involved in the production of eggs, a primordial diploid egg cell is converted into a haploid egg and three small haploid polar bodies (minute cells). In this instance the egg receives far more cytoplasm than the polar bodies.

Multiple fission. Some algae, some protozoans, and the true slime molds (Myxomycetes) regularly divide by multiple fission. In such cases the nucleus undergoes several mitotic divisions, producing a number of nuclei. After the nuclear divisions are complete, the cytoplasm separates, and each nucleus becomes encased in its own membrane to form an individual cell. In the Myxomycetes, the fusion of two haploid gametes or the fusion of two or more diploid zygotes (the structures that result from the union of two sex cells) results in the formation of a plasmodium—a motile, multinucleate mass of cytoplasm. The nuclei are in a syncytium, that is, there are no cell boundaries, and the nuclei flow freely in the motile plasmodium. As it feeds, the plasmodium enlarges, and the nuclei divide synchronously about once every 24 hours. The plasmodium may become very large, with millions of nuclei, but, ultimately, when conditions are right, it forms a series of small bumps, each of which becomes a small, fruiting body (a structure that bears the spores). During this process the nuclei undergo meiosis, and the final haploid nuclei are then isolated into uninucleate spores (reproductive bodies).

Development of a plasmodium

Many algae (*e.g.*, the Siphonales and related groups) are multinucleate. In most instances the nuclei are in one common cytoplasm within a large and elaborate organism surrounded by a hard cell wall. As the wall becomes extended, the nuclei, which wander freely in the central cavity, undergo repeated mitoses. Again, either during the formation of zoospores (asexual reproductive cells) or after meiosis during gamete formation, a massive progressive division occurs. The most unusual of such organisms is the marine alga *Acetabularia;* many nuclei stay clumped together in one compound nucleus in the rootlike base, which often is as much as two inches (five centimetres) away from the tip of the plant. The compound nucleus breaks up just before gamete formation, and the minute individual nuclei undergo meiosis and wander to the elaborate tip structures, where they are released as uninucleate gametes.

Syncytial organisms raise the question of whether or not cells, in the strict sense, are necessary for the development of large organisms. Syncytia are also found in animals—*e.g.*, in the early stages of development of fishes and insects—and in the voluntary muscles of man. The proposal of the 19th-century botanist Julius von Sachs is generally considered a satisfactory answer to this question; he suggested that the important matter was the existence not of a cell membrane but of a certain amount of cytoplasm surrounding a nucleus and acting as a unit of metabolism, which he called an energid. Cell reproduction, therefore, might be considered a special case of energid reproduction.

Reproduction of organisms. In single-celled organisms (*e.g.*, bacteria, protozoans, many algae, and some fungi), organismic and cell reproduction are synonymous, for the cell is the whole organism. Details of the process differ greatly from one form to the next and, if the higher ciliate protozoans are included, can be extraordinarily complex. It is possible for reproduction to be asexual, by simple division, or sexual. In sexual unicellular organisms the gametes can be produced by division (often multiple fission, as in numerous algae) or, as in yeasts, by the organism turning itself into a gamete and fusing its nucleus with that of a neighbour of the opposite sex, a process that is called conjugation. In ciliate protozoans (*e.g.*, *Paramecium*), the conjugation process involves the exchange of haploid nuclei; each partner acquires a new nuclear apparatus, half of which is genetically derived from its mate. The parent cells separate and subsequently reproduce by

binary fission. Sexuality is present even in primitive bacteria, in which parts of the chromosome of one cell can be transferred to another during mating.

Vegetative reproduction

Multicellular organisms also reproduce asexually and sexually; asexual, or vegetative, reproduction can take a great variety of forms. Many multicellular lower plants give off asexual spores, either aerial or motile and aquatic (zoospores), which may be uninucleate or multinucleate. In some cases the reproductive body is multicellular, as in the soredia of lichens and the gemmae of liverworts. Frequently, whole fragments of the vegetative part of the organism can bud off and begin a new individual, a phenomenon that is found in most plant groups. In many cases a spreading rhizoid (rootlike filament) or, in higher plants, a rhizome (underground stem) gives off new sprouts. Sometimes other parts of the plant have the capacity to form new individuals; for instance, buds of potentially new plants may form in the leaves; even some shoots that bend over and touch the ground can give rise to new plants at the point of contact (see below *Plant reproductive systems*).

Among animals, many invertebrates are equally well endowed with means of asexual reproduction. Numerous species of sponges produce gemmules, masses of cells enclosed in resistant cases, that can become new sponges. There are many examples of budding among coelenterates, the best known of which occurs in freshwater *Hydra*. In some species of flatworms, the individual worm can duplicate by pinching in two, each half then regenerating the missing half; this is a large task for the posterior portion, which lacks most of the major organs—brain, eyes, and pharynx. The highest animals that exhibit vegetative reproduction are the colonial tunicates (*e.g.*, sea squirts), which, much like plants, send out runners in the form of stolons, small parts of which form buds that develop into new individuals. Vertebrates have lost the ability to reproduce vegetatively; their only form of organismic reproduction is sexual.

In the sexual reproduction of all organisms except bacteria, there is one common feature: haploid, uninucleate gametes are produced that join in fertilization to form a diploid, uninucleate zygote. At some later stage in the life history of the organism, the chromosome number is again reduced by meiosis to form the next generation of gametes. The gametes may be equal in size (isogamy), or one may be slightly larger than the other (anisogamy); the majority of forms have a large egg and a minute sperm (oogamy). The sperm are usually motile and the egg passive, except in higher plants, in which the sperm nuclei are carried in pollen grains that attach to the stigma (a female structure) of the flower and send out germ tubes that grow down to the egg nucleus in the ovary. Some organisms, such as most flowering plants, earthworms, and tunicates, are bisexual (hermaphroditic, or monoecious)—*i.e.*, both the male and female gametes are produced by the same individual. All other organisms, including some plants (*e.g.*, holly and the ginkgo tree) and all vertebrates, are unisexual (dioecious): the male and female gametes are produced by separate individuals.

Parthenogenesis

Some sexual organisms partially revert to the asexual mode by a periodic degeneration of the sexual process. For instance, in aphids and in many higher plants the egg nucleus can develop into a new individual without fertilization, a kind of asexual reproduction that is called parthenogenesis.

Life-cycle reproduction. Although organisms are often thought of only as adults, and reproduction is considered to be the formation of a new adult resembling the adult of the previous generation, a living organism, in reality, is an organism for its entire life cycle, from fertilized egg to adult, not for just one short part of that cycle. Reproduction, in these terms, is not just a stage in the life history of an organism but the organism's entire history. It has been pointed out that only the DNA of a cell is capable of replicating itself, and even that replication process requires specific enzymes that were themselves formed from DNA. Thus, the reproduction of all living forms must be considered in relation to time; what is reproduced is a series of copies that, like the sequence of individual frames of a motion picture, change through time in an exact and orderly fashion.

A few examples serve to illustrate the great variety of life cycles in living organisms. They also illustrate how different parts of the life cycle can change, and the fact that these changes are not confined solely to adult structures. One variation is that of minimum size—that is to say, the differences in the sizes of gametes (mature sex cells) and asexual bodies. An even greater variation in life cycles, however, involves maximum size; there is an enormous difference between a single-celled organism that divides by binary fission and a giant sequoia. Size is correlated with time. A bacterium requires about 30 minutes to complete its life history and divide in two (generation time); a giant sequoia bears its first cones and fertile seeds after 60 years. Not only is the life cycle of the sequoia 10,000,000 times longer than that of the bacterium, but the large difference in size also means that the tree must be elaborate and complex. It contains different tissue types that must be carefully duplicated from generation to generation.

Life cycles of plants. Most life histories, except perhaps for the simplest and smallest organisms, consist of different epochs. A large tree has a period of seed formation that involves many cell divisions after fertilization and the laying down of a small embryo in a hard resistant shell, or seed coat. There then follows a period of dormancy, sometimes prolonged, after which the seed germinates, and the adult form slowly emerges as the shoots and roots grow at the tips and the stem thickens. In some trees the leaves of the juvenile plant have a shape that is quite different from that of the taller, more mature individuals. Thus, even the growth phase may be subdivided into epochs, the final one being the flowering or gametebearing period. Some of the parasitic fungi have much more complex life histories. The wheat rust parasite, for example, has alternate hosts. While living on wheat, it produces two kinds of spores; it produces a third kind of spore when it invades its other host, the barberry, on which it winters and undergoes the sexual part of its life cycle.

In plants, variations in the epochs of the life cycle are often centred around the times of fertilization and meiosis. After fertilization the organism has the diploid number of chromosomes (diplophase); after meiosis it is haploid (haplophase). The two events vary in time with respect to each other. In some simple algae (*e.g.*, *Chlamydomonas*), for example, most of the cycle is haploid; meiosis occurs immediately after fertilization. Yet in other algae, such as the sea lettuce (*Ulva*), two equal haploid and diploid cycles alternate. The outward morphological structures of mature *Ulva* are indistinguishable; the two cycles can be differentiated only by the size of the cell or nucleus, those of the haploid stage being half the size of those of the diploid stage.

In many of the higher algae, there is a progressive diminution of the haplophase and an increase in the importance of the diplophase, a trend that is especially noticeable in the evolution of the vascular plants (*e.g.*, ferns, conifers, and flowering plants). In mosses, the haplophase, or gametophyte, is the main part of the green plant; the diplophase, or sporophyte, usually is a sporebearing spike that grows from the top of the plant. In ferns, the haplophase is reduced to a small, inconspicuous structure (prothallus) that grows in the damp soil; the large sporebearing fern itself is entirely diploid. Finally, in higher plants the haploid tissue is confined to the ovary of the large diploid organism, a condition that is also prevalent in most animals.

Metamorphosis

Life cycles of animals. Invertebrate animals have a rich variety of life cycles, especially among those forms that undergo metamorphosis, a radical physical change. Butterflies, for instance, have a caterpillar stage (larva), a dormant chrysalis stage (pupa), and an adult stage (imago). One remarkable aspect of this development is that, during the transition from caterpillar to adult, most of the caterpillar tissue disintegrates and is used as food, thereby providing energy for the next stage of development, which begins when certain small structures (imaginal disks) in the larva start growing into the adult form. Thus, the butterfly undergoes essentially two periods of growth and

development (larva and pupa–adult) and two periods of small size (fertilized egg and imaginal disks). A somewhat similar phenomenon is found in sea urchins; the larva, which is called a pluteus, has a small, wartlike bud that grows into the adult while the pluteus tissue disintegrates. In both examples it is as if the organism has two life histories, one built on the ruins of another.

Another life-cycle pattern found among certain invertebrates illustrates the principle that major differences between organisms are not always found in the physical appearance of the adult but in differences of the whole life history. In the coelenterate *Obelia,* for example, the egg develops into a colonial hydroid consisting of a series of branching *Hydra*-like organisms called polyps. Certain of these polyps become specialized (reproductive polyps) and bud off from the colony as free-swimming jellyfish (medusae) that bear eggs and sperm. As with caterpillars and sea urchins, two distinct phases occur in the life cycle of *Obelia:* the sessile (anchored), branched polyps and the motile medusae. In some related coelenterates the medusa form has been totally lost, leaving only the polyp stage to bear eggs and sperm directly. In still other coelenterates the polyp stage has been lost, and the medusae produce other medusae directly, without the sessile stage. There are, furthermore, intermediate forms between the extremes.

NATURAL SELECTION AND REPRODUCTION

The significance of biological reproduction can be explained entirely by natural selection (see EVOLUTION, THE THEORY OF: *Natural selection*). In formulating his theory of natural selection, Charles Darwin realized that, in order for evolution to occur, not only must living organisms be able to reproduce themselves but the copies must not all be identical; that is, they must show some variation. In this way the more successful variants would make a greater contribution to subsequent generations in the number of offspring. For such selection to act continuously in successive generations, Darwin also recognized that the variations had to be inherited, although he failed to fathom the mechanism of heredity. Moreover, the amount of variation is particularly important. According to what has been called the principle of compromise, which itself has been shaped by natural selection, there must not be too little or too much variation: too little produces no change; too much scrambles the benefit of any particular combination of inherited traits.

Of the numerous mechanisms for controlling variation, all of which involve a combination of checks and balances that work together, the most successful is that found in the large majority of all plants and animals—*i.e.,* sexual reproduction. During the evolution of reproduction and variation, which are the two basic properties of organisms that not only are required for natural selection but are also subject to it, sexual reproduction has become ideally adapted to produce the right amount of variation and to allow new combinations of traits to be rapidly incorporated into an individual.

The evolution of reproduction. An examination of the way in which organisms have changed since their initial unicellular condition in primeval times shows an increase in multicellularity and therefore an increase in the size of both plants and animals. After cell reproduction evolved into multicellular growth, the multicellular organism evolved a means of reproducing itself that is best described as life-cycle reproduction. Size increase has been accompanied by many mechanical requirements that have necessitated a selection for increased efficiency; the result has been a great increase in the complexity of organisms. In terms of reproduction this means a great increase in the permutations of cell reproduction during the process of evolutionary development.

Size increase also means a longer life cycle, and with it a great diversity of patterns at different stages of the cycle. This is because each part of the life cycle is adaptive in that, through natural selection, certain characteristics have evolved for each stage that enable the organism to survive. The most extreme examples are those forms with two or more separate phases of their life cycle separated by a metamorphosis, as in caterpillars and butterflies; these phases may be shortened or extended by natural selection, as has occurred in different species of coelenterates.

Efficiency of reproduction

To reproduce efficiently in order to contribute effectively to subsequent generations is another factor that has evolved through natural selection. For instance, an organism can produce vast quantities of eggs of which, possibly by neglect, only a small percent will survive. On the other hand, an organism can produce very few or perhaps one egg, which, as it develops, will be cared for, thereby greatly increasing its chances for survival. These are two strategies of reproduction; each has its advantages and disadvantages. Many other considerations of the natural history and structure of the organism determine, through natural selection, the strategy that is best for a particular species; one of these is that any species must not produce too few offspring (for it will become extinct) or too many (for it may also become extinct by overpopulation and disease). The numbers of some organisms fluctuate cyclically but always remain between upper and lower limits. The question of how, through natural selection, numbers of individuals are controlled is a matter of great interest; clearly, it involves factors that influence the rate of reproduction.

The evolution of variation control. Because inherited variation is largely handled by genes in the chromosomes, organisms that reproduce sexually require a single-cell stage in their life cycle, during which the haploid gamete of each parent can combine to form the diploid zygote. This is also often true in organisms that reproduce asexually, but in this case the asexual reproductive bodies (*e.g.,* spores) are small and hence are effectively dispersed.

The amount of variation is controlled in a large number of ways, all of which involve a carefully balanced set of factors. These factors include whether the organism reproduces asexually or sexually: the mutation (gene change) rate; the number of chromosomes; the amount of exchange of parts of chromosomes (crossing over); the size of the individual (which correlates with complexity and generation time); the size of the population; the degree of inbreeding versus outbreeding; and the relative amounts and position of haploidy and diploidy in the life cycle. It is clear, therefore, that the mode of reproduction influences the amount of variation and vice versa; the two together permit natural selection to operate, and selection in turn modifies the mechanisms of reproduction and variation.

(J.T.Bo.)

The process of fertilization

Fertilization, the completion of the cycle of sexual reproduction, may be formally defined as the union of a spermatozoal nucleus, of paternal origin, with an egg nucleus, of maternal origin, to form the primary nucleus of an embryo. In all organisms the essence of fertilization is, in fact, the fusion of the hereditary material of two different sex cells, or gametes, each of which carries half the number of chromosomes typical of the species. The most primitive form of fertilization, found in micro-organisms and protozoans, consists of an exchange of genetic material between two cells.

The first significant event in fertilization is the fusion of the membranes of the two gametes resulting in the formation of a channel that allows the passage of material from one cell to the other. Fertilization in advanced plants is preceded by pollination, during which pollen is transferred to, and establishes contact with, the female gamete or macrospore (see below *Pollination).* Fusion in advanced animals is usually followed by penetration of the egg by a single spermatozoon. The result of fertilization is a cell (zygote) capable of undergoing cell division to form a new individual.

Formation of the zygote

The fusion of two gametes initiates several reactions in the egg. One of these causes a change in the egg membrane(s), so that the attachment of and penetration by more than one spermatozoon cannot occur. In species in which more than one spermatozoon normally enters an egg (polyspermy), only one spermatozoal nucleus actually merges with the egg nucleus. The most important result of fertilization is egg activation, which allows the egg to

undergo cell division. Activation, however, does not necessarily require the intervention of a spermatozoon; during parthenogenesis, in which fertilization does not occur, activation of an egg may be accomplished through the intervention of physical and chemical agents. Invertebrates such as aphids, bees, and rotifers normally reproduce by parthenogenesis.

In plants certain chemicals produced by the egg may attract spermatozoa. In animals, with the possible exception of some coelenterates, it appears likely that contact between eggs and spermatozoa depends on random collisions. On the other hand, the gelatinous coats that surround the eggs of many animals exert a trapping action on spermatozoa, thus increasing the chances for successful sperm-egg interaction.

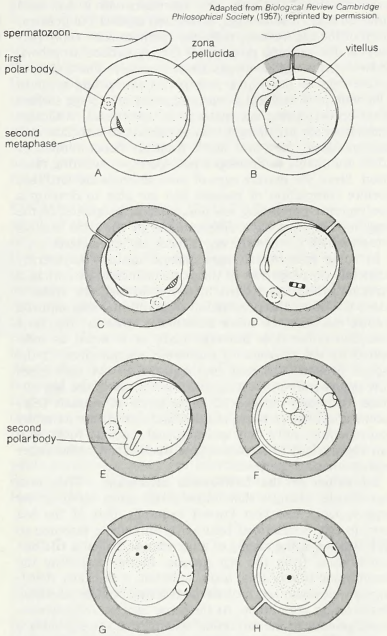


Figure 1: *Steps in fertilization in the rat egg.*
(A–C) Entrance of spermatozoon; shading denotes zona reaction. (D, E) Completion of meiotic division. (F) Pronuclear development. (G) Reappearance of chromosome group. (H) First cleavage metaphase.

The eggs of marine invertebrates, especially echinoderms, are classical objects for the study of fertilization. These transparent eggs are valuable for studies observing living cells and for biochemical and molecular investigations because the time of fertilization can be accurately fixed, the development of many eggs occurs at about the same rate under suitable conditions, and large quantities of the eggs are obtainable. The eggs of some teleosts and amphibians also have been used with favourable results, and techniques for fertilization of mammalian eggs in the laboratory may allow their use even though only small numbers are available.

MATURATION OF THE EGG

Maturation is the final step in the production of functional eggs (oogenesis) that can associate with a spermatozoon and develop a reaction that prevents the entry of more than one spermatozoon; in addition, the cytoplasm of a mature egg can support the changes that lead to fusion of spermatozoal and egg nuclei and initiate embryonic development.

Egg surface. Certain components of an egg's surface, especially the cortical granules, are associated with a mature condition. Cortical granules of sea urchin eggs, aligned beneath the plasma membrane (thin, soft, pliable layer) of mature eggs, have a diameter of 0.8–1.0 micron (0.0008–0.001 millimetre) and are surrounded by a membrane similar in structure to the plasma membrane surrounding the egg. Cortical granules are formed in a cell component known as a Golgi complex, from which they migrate to the surface of the maturing egg.

The surface of a sea urchin egg has the ability to affect the passage of light unequally in different directions; this property, called birefringence, is an indication that the molecules comprising the surface layers are arranged in a definite way. Since birefringence appears as an egg matures, it is likely that the properties of a mature egg membrane are associated with specific molecular arrangements. A mature egg is able to support the formation of a zygote nucleus; *i.e.,* the result of fusion of spermatozoal and egg nuclei. In most eggs the process of reduction of chromosomal number (meiosis) is not completed prior to fertilization. In such cases the fertilizing spermatozoon remains beneath the egg surface until meiosis in the egg has been completed, after which changes and movements that lead to fusion and the formation of a zygote occur.

Egg coats. The surfaces of most animal eggs are surrounded by envelopes, which may be soft, gelatinous coats (as in echinoderms and some amphibians) or thick membranes (as in fishes, insects, and mammals). In order to reach the egg surface, therefore, spermatozoa must penetrate these envelopes; indeed, spermatozoa contain enzymes (organic catalysts) that break them down. In some cases (*e.g.,* fishes and insects) there is a channel, or micropyle, in the envelope, through which a spermatozoon can reach the egg.

The jelly coats of echinoderm and amphibian eggs consist of complex carbohydrates called sulfated mucopolysaccharides; it is not yet known if they have a species-specific composition. The envelope of a mammalian egg is more complex. The egg is surrounded by a thick coat composed of a carbohydrate protein complex called zona pellucida. The zona is surrounded by an outer envelope, the corona radiata, which is many cell layers thick and formed by follicle cells adhering to the oocyte before it leaves the ovarian follicle.

Although it once was postulated that the jelly coat of an echinoderm egg contains a substance (fertilizin) thought to have an important role not only in the establishment of sperm-egg interaction but also in egg activation, fertilizin now has been shown identical with jelly-coat material, rather than a substance continuously secreted from it. Yet there is evidence that the egg envelopes do play a role in fertilization; *i.e.,* contact with the egg coat elicits the acrosome reaction (described below) in spermatozoa.

EVENTS OF FERTILIZATION

Sperm–egg association. The acrosome reaction of spermatozoa is a prerequisite for the association between a spermatozoon and an egg (Figure 2), which occurs through fusion of their plasma membranes. After a spermatozoon comes in contact with an egg (2A), the acrosome, which is a prominence at the anterior tip of the spermatozoa, undergoes a series of well-defined structural changes. A structure within the acrosome, called the acrosomal vesicle (labelled a), bursts, and the plasma membrane (labelled s) surrounding the spermatozoon fuses at the acrosomal tip with the membrane surrounding the acrosomal vesicle to form an opening (2B). As the opening is formed, the acrosomal granule (labelled g), which is enclosed within the acrosomal vesicle, disappears. It is thought that dissolution of the granule releases a substance called a lysin, which breaks down the egg envelopes, allowing passage of the spermatozoon to the egg. The acrosomal membrane region opposite the opening adheres to the nuclear envelope of the spermatozoon and forms a shallow outpocketing, which rapidly elongates into a thin tube, the acrosomal

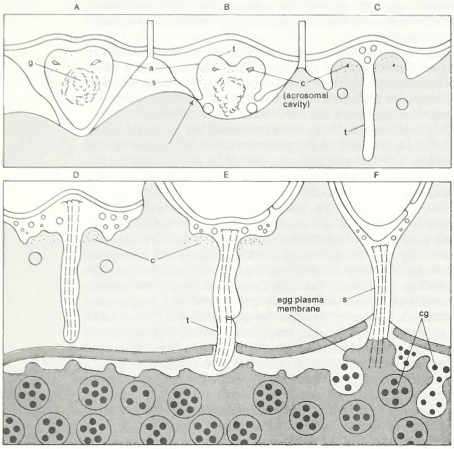


Figure 2: A diagrammatic representation of successive states of the sperm penetration in the egg of *Saccoglossus*.

From Colwin and Colwin, *Journal of Cell Biology*, vol. 19 (December 1963)

tubule (labelled t) that extends to the egg surface and fuses with the egg plasma membrane (2C, 2D). The tubule thus formed establishes continuity between the egg and the spermatozoon and provides a way for the spermatozoal nucleus to reach the interior of the egg. Other spermatozoal structures that may be carried within the egg include the midpiece and part of the tail; the spermatozoal plasma membrane and the acrosomal membrane, however, do not reach the interior of the egg. In fact, whole spermatozoa injected into unfertilized eggs cannot elicit the activation reaction or merge with the egg nucleus. As the spermatozoal nucleus is drawn within the egg, the spermatozoal plasma membrane breaks down; at the end of the process, the continuity of the egg plasma membrane is re-established. This description of the process of sperm-egg association, first documented for the acorn worm *Saccoglossus* (phylum Enteropneusta), generally applies to most eggs studied thus far.

During their passage through the female genital tract of mammals, spermatozoa undergo physiological change, called capacitation, which is a prerequisite for their participation in fertilization; they are able to undergo the acrosome reaction, traverse the egg envelopes, and reach the interior of the egg. Dispersal of cells in the outer egg envelope (corona radiata) is caused by the action of an enzyme (hyaluronidase) that breaks down a substance (hyaluronic acid) binding corona radiata cells together. The enzyme may be contained in the acrosome and released as a result of the acrosome reaction, during passage of the spermatozoon through the corona radiata. The reaction is well advanced by the time a spermatozoon contacts the thick coat surrounding the egg itself (zona pellucida). The pathway of a spermatozoon through the zona pellucida appears to be an oblique slit.

Association of a mammalian spermatozoon with the egg surface occurs along the lateral surface of the spermatozoon, rather than at the tip as in other animals, so that the spermatozoon lies flat on the egg surface; several points of fusion occur between the plasma membranes of the two gametes (*i.e.*, the breakdown of membranes occurs by formation of numerous small vesicles).

Specificity of sperm–egg interaction. Although fertilization is strictly species-specific, very little is known about the molecular basis of such specificity. The egg coats may have a role. Among the echinoderms solutions of the jelly coat clump, or agglutinate, only spermatozoa of their own species. In both echinoderms and amphibians, however, slight damage to an egg surface makes fertilization possible with spermatozoa of different species (heterologous fertilization); this procedure has been used to obtain certain hybrid larvae.

The eggs of ascidians, or sea squirts, members of the chordate subphylum Tunicata, are covered with a thick membrane called a chorion; the space between the chorion and the egg is filled with cells called test cells. The gametes of ascidians, which have both male and female reproductive organs in one animal, mature at the same time; yet self-fertilization does not occur. If the chorion and the test cells are removed, however, not only is fertilization with spermatozoa of different species possible, but self-fertilization also can occur.

Prevention of polyspermy. Most animal eggs are monospermic; *i.e.*, only one spermatozoon is admitted into an egg. In some eggs, protection against the penetration of the egg by more than one spermatozoon (polyspermy) is due to some property of the egg surface; in others, however, the egg envelopes are responsible. The ability of some eggs to develop a polyspermy-preventing reaction depends on a molecular rearrangement of the egg surface that occurs during egg maturation (oogenesis). Although immature sea urchin eggs have the ability to associate with spermatozoa, they also allow multiple penetration; *i.e.*, they are unable to develop a polyspermy-preventing reaction. Since the mature eggs of most animals are fertilized before completion of meiosis and are able to develop a polyspermy-preventing reaction, specific properties of the egg surface must have differentiated by the time meiosis stops, which is when the egg is ready to be fertilized.

Roles of the egg surface and coats

In some mammalian eggs defense against polyspermy depends on properties of the zona pellucida; *i.e.*, when a spermatozoon has started to move through the zona, it does not allow the penetration of additional spermatozoa (zona reaction). In other mammals, however, the zona reaction either does not take place or is weak, as indicated by the presence of numerous spermatozoa in the space between the zona and egg surface. In such cases the polyspermy-preventing reaction resides in the egg surface. Although the eggs of some kinds of animals (*e.g.*, some amphibians, birds, reptiles, and sharks) are naturally polyspermic, only one spermatozoal nucleus fuses with an egg nucleus to form a zygote nucleus; all of the other spermatozoa degenerate.

Formation of the fertilization membrane. The most spectacular changes that follow fertilization occur at the egg surface. The best known example, that of the sea urchin egg, is described below. An immediate response to fertilization is the raising of a membrane, called a vitelline membrane, from the egg surface. In the beginning the membrane is very thin; soon, however, it thickens, develops a well-organized molecular structure, and is called the fertilization membrane. At the same time an extensive rearrangement of the molecular structure of the egg surface occurs. The events leading to formation of the fertilization membrane require about one minute.

At the point on the outer surface of the sea urchin egg at which a spermatozoan attaches, the thin vitelline membrane becomes detached. As a result the membranes of the cortical granules (cg, see Figure 2F) come into contact with the inner aspect of the egg's plasma membrane and fuse with it, the granules open, and their contents are extruded into the perivitelline space; *i.e.*, the space between the egg surface and the raised vitelline membrane. Part of the contents of the granules merge with the vitelline membrane to form the fertilization membrane; if fusion of the contents of the cortical granules with the vitelline membrane is prevented, the membrane remains thin and soft. Another material that also derives from the cortical granules covers the surface of the egg to form a transparent layer, called the hyaline layer, which plays an important role in holding together the cells (blastomeres) formed during division, or cleavage, of the egg. The plasma membrane surrounding a fertilized egg, therefore, is a mosaic structure containing patches of the original plasma membrane of the unfertilized egg and areas derived from membranes of the cortical granules. The events leading to the formation of the fertilization membrane are accompanied by a change of the

Effects on membrane properties

electric charge across the plasma membrane, referred to as the fertilization potential, and a concurrent outflow of potassium ions (charged particles); both of these phenomena are similar to those that occur in a stimulated nerve fibre. Another effect of fertilization on the plasma membrane of the egg is a several-fold increase in its permeability to various molecules; this change may be the result of the activation of some surface-located membrane transport mechanism.

Formation of the zygote nucleus. After its entry into the egg cytoplasm, the spermatozoal nucleus, now called the male pronucleus, begins to swell, and its chromosomal material disperses and becomes similar in appearance to that of the female pronucleus. Although the membranous envelope surrounding the male pronucleus rapidly disintegrates in the egg, a new envelope promptly forms around it. The male pronucleus, which rotates 180° and moves towards the egg nucleus, initially is accompanied by two structures (centrioles) that function in cell division. After the male and female pronuclei have come into contact, the spermatozoal centrioles give rise to the first cleavage spindle, which precedes division of the fertilized egg. In some cases fusion of the two pronuclei may occur by a process of membrane fusion; in this process, two adjoining membranes fuse at the point of contact to give rise to the continuous nuclear envelope that surrounds the zygote nucleus.

BIOCHEMICAL ANALYSIS OF FERTILIZATION

Many of the early studies on biochemical changes occurring during fertilization were concerned with the respiratory metabolism of the egg. The results, however, were deceiving; the sea urchin egg, for example, showed an increased rate of oxygen consumption as an immediate response to either fertilization or parthenogenetic activation, in apparent support of the idea that the essence of fertilization is the removal of a respiratory or metabolic block in the unfertilized egg. Extensive comparative studies have shown that the increased rate of oxygen consumption in fertilized sea urchin eggs is not a general rule; indeed, the rate of oxygen consumption of most animal eggs does not change at the time of fertilization and may even temporarily decrease.

At the time of fertilization the egg contains the components required to carry out protein synthesis, and hence development, through an early embryonic stage called the blastula. Most immediate post-fertilization protein synthesis is directed by molecules of ribonucleic acid, known as messenger RNA, that were formed during oogenesis and stored in the egg. In addition, protein synthesis up to the blastula stage (up to a much earlier stage in the mammalian embryo) is directed by the cell components called ribosomes, which are present in the unfertilized egg; new ribosomes, as well as molecules of another type of RNA involved in protein synthesis, and called transfer RNA, are synthesized at a later stage in embryonic development (gastrulation). Eggs fertilized and allowed to develop in the presence of the antibiotic actinomycin, which suppresses RNA synthesis, not only reach the blastula stage but their rate of protein synthesis is the same as that in untreated embryos.

Unfertilized sea urchin eggs, as well as those of other marine animals studied thus far, have a very low rate of protein synthesis, suggesting that something in the unfertilized egg inhibits its protein synthesizing machinery. Since the rate of protein synthesis increases immediately following fertilization, it may depend on some change in, or removal of, an inhibitor. In the sea urchin egg, for example, the low efficiency of the protein synthesizing apparatus apparently depends on certain properties of the ribosomes. Most of the ribosomes found in an unfertilized sea urchin egg are single ribosomes (so-called monosomes); soon after fertilization, however, the single ribosomes interact with messenger RNA molecules thus giving rise to the polyribosomes, which are the active units in protein synthesis. This process also occurs in eggs of a few other marine animals that have been studied. The protein-synthesizing inefficiency of unfertilized sea-urchin-egg ribosomes is caused by an inhibitor that is associated with them and interferes with the binding of messenger RNA molecules to the ribosomes; the inhibitor is removed almost immediately following fertilization, perhaps by enzymatic breakdown. Protein synthesis

It thus appears that at least in the sea urchin egg the overall rate of protein synthesis is controlled at the ribosome level and that the first step in the activation of protein synthesis following fertilization is the "turning on" of the ribosomes.

In vertebrates such as amphibians, activation of protein synthesis takes place at the onset of egg maturation, apparently initiated by the action of a hormone, progesterone. The effect of progesterone is not mediated by the nucleus but is a direct effect on the cytoplasm.

(A.Mo.)

PLANT REPRODUCTION

Plant reproductive systems

In plants, as in animals, the end result of reproduction is the continuation of a given species, and the ability to reproduce is, therefore, rather conservative, or given to only moderate change, during evolution. Changes have occurred, however, and the pattern is demonstrable through a survey of plant groups. Reproduction is basically either asexual or sexual. Asexual reproduction in plants involves a variety of widely disparate methods for creating new plants identical in every respect to the parent plant. Sexual reproduction, on the other hand, depends on a complex series of basic cellular events, involving chromosomes and their genes, that take place within an elaborate sexual apparatus evolved precisely for the creation of new plants in some respects different from the two parents that played a role in their production. (For an account of the common details of asexual and sexual reproduction and the evolutionary significance of the two methods see above *The nature of reproduction*.)

In order to describe the modification of reproductive systems, plant groups must be identified. One convenient classification of organisms sets plants apart from lower forms such as bacteria, algae, fungi, and protozoans. Under such an arrangement, the plants, as separated, comprise two great divisions (or phyla)—the Bryophyta (mosses and liverworts) and the Tracheophyta (vascular plants). The vascular plants include four subdivisions: the three entirely seedless groups are the Psilopsida, Lycopsida, and Sphenopsida; the fourth group, the Pteropsida, consists of the ferns (seedless) and the seed plants (gymnosperms and angiosperms).

A comparative treatment of the two patterns of reproductive systems will introduce the terms required for an understanding of the survey of those systems as they appear in selected plant groups.

GENERAL FEATURES OF ASEXUAL SYSTEMS

Asexual reproduction involves no union of cells or nuclei of cells and, therefore, no mingling of genetic traits, since the nucleus contains the genetic material (chromosomes) of the cell. Only those systems of asexual reproduction that are not really modifications of sexual reproduction are considered below. They fall into two basic types: systems that utilize almost any fragment or part of a plant body; and systems that depend upon specialized structures that have evolved as reproductive agents. Two basic asexual processes

Reproduction by fragments. In many plant groups, fragmentation of the plant body, followed by regeneration and development of the fragments into whole new organisms, serves as a reproductive system. Fragments of the plant bodies of liverworts and mosses regenerate to form new plants. In nature and in laboratory and greenhouse culture, liverworts fragment as a result of growth; the growing

fragments separate by decay at the region of attachment to the parent. During prolonged drought, the mature portions of liverworts often die, but their tips resume growth and produce a series of new plants from the original parent plant.

In mosses, small fragments of the stems and leaves (even single cells of the latter) can, with sufficient moisture and under proper conditions, regenerate and ultimately develop into new plants.

It is common horticultural practice to propagate desirable varieties of garden plants by means of plant fragments, or cuttings. These may be severed leaves or portions of roots or stems, which are stimulated to develop roots and produce leafy shoots. Naturally fallen branches of willows (*Salix*) and poplars (*Populus*) root under suitable conditions in nature and eventually develop into trees. Other horticultural practices that exemplify asexual reproduction include budding (the removal of buds of one plant and their implantation on another) and grafting (the implantation of small branches of one individual on another).

Reproduction by special asexual structures. Throughout the plant kingdom, specially differentiated or modified cells, groups of cells, or organs have, during the course of evolution, come to function as organs of asexual reproduction. These structures are asexual in that the individual reproductive agent develops into a new individual without the union of sex cells (gametes). A number of examples of special asexual agents of reproduction from several plant groups are cited in this section.

Spores and vegetative organs

Airborne spores characterize most nonflowering land plants, such as mosses, liverworts, and ferns. Although the spores arise as products of meiosis, a cellular event in which the number of chromosomes in the nucleus is halved, such spores are asexual in the sense that they may grow directly into new individuals, without prior sexual union.

Among liverworts, mosses, lycopods, ferns, and seed plants, few to many celled, specially organized buds, or gemmae, also serve as agents of asexual reproduction.

The vegetative, or somatic, organs of plants may, in their entirety, be modified to serve as organs of reproduction. In this category belong such flowering-plant structures as stolons, rhizomes, tubers, corms, and bulbs, as well as the tubers of liverworts, ferns, and horsetails, the dormant buds of certain moss stages, and the leaves of many succulents (Figure 3). Stolons are elongated runners, or horizontal stems, such as those of the strawberry, which root and form new plantlets when they make proper contact with a moist soil surface. Rhizomes, as seen in iris, are fleshy, elongated, horizontal stems that grow within or upon the soil. The branching of rhizomes results in multiplication of the plant. The enlarged, fleshy tips of subterranean rhizomes or stolons are known as tubers, examples of which are potatoes. Tubers are fleshy storage stems, the buds ("eyes") of which, under proper conditions, can develop into new individuals. Erect, vertical, fleshy, subterranean stems, which are known as corms, are exemplified by crocuses and gladioli. These organs tide the plants over periods of dormancy and may develop secondary cormlets, which give rise to new plantlets. Unlike the corm, only a small portion of the bulb, as in lilies and the onion, represents stem tissue. The latter is surrounded by the fleshy, food-storage bases of earlier formed leaves. After a period of dormancy, bulbs develop into new individuals. Large bulbs produce secondary bulbs through development of buds, resulting in an increase in number of individuals.

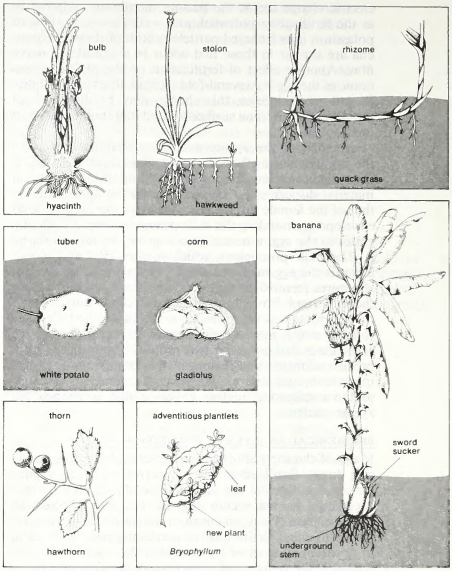


Figure 3: Structures serving asexual reproduction.

From (rhizome, tuber, corm, thorn, adventitious plantlets, banana) *Biological Science: An Inquiry into Life*, 2nd ed. (1968), Harcourt Brace Jovanovich, Inc., New York, by permission of the Biological Sciences Curriculum Study. (bulb) Gilbert M. Smith, *et al., A Textbook of General Botany* (S. 1953), The Macmillan Company. (stolon) *Botany*, 3rd ed. by Carl L. Wilson and Walter E. Loomis, Copyright 1952, © 1957, 1962 by Holt, Rinehart and Winston, Inc., reproduced by permission of Holt, Rinehart and Winston, Inc.

GENERAL FEATURES OF SEXUAL SYSTEMS

In most plant groups both sexual and asexual methods of reproduction occur. Some species, however, seem secondarily to have lost the capacity for sexual reproduction. Such cases are described below (see *Variations in reproductive cycles*).

The cellular basis. Sexual reproduction at the cellular level generally involves the following phenomena: the union of sex cells and their nuclei, with concomitant association of their chromosomes, which contain the genes, and the nuclear division called meiosis. The sex cells are called gametes, and the product of their union is a zygote. All gametes are normally haploid (having a single set of chromosomes) and all zygotes, diploid (having a double set of chromosomes, one set from each parent). Gametes may be motile, by means of whiplike hairs (flagella) or of flowing cytoplasm (amoeboid motion). In their union, gametes may be morphologically indistinguishable (*i.e.*, isogamous) or they may be distinguishable only on the criterion of size (*i.e.*, heterogamous). The larger gamete, or egg, is nonmotile; the smaller gamete, or sperm, is motile. The last type of gametic difference, egg and sperm, is often designated as oogamy. In oogamous reproduction, the union of sperm and egg is called fertilization. Isogamy, heterogamy, and oogamy are often considered to represent an increasingly specialized evolutionary series (Figure 4).

From P.B. Weisz, *The Science of Biology*, 3rd ed. (1967), used with permission of McGraw-Hill Book Company

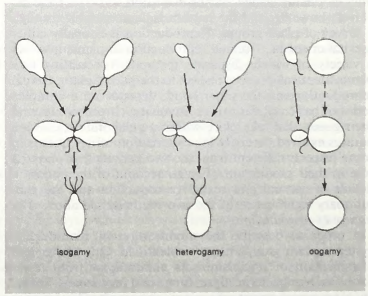


Figure 4: Patterns of fertilization based on gamete types. In isogamy and heterogamy all gametes are motile; in oogamy only the smaller is motile, and moves to the larger.

In the plants included in this article—bryophytes (mosses and liverworts) and tracheophytes (vascular plants)—sexual reproduction is of the oogamous type, or a modification thereof, in which the sex cells, or gametes, are of two types, a larger nonmotile egg and a smaller motile sperm. These gametes are often produced in special containers called gametangia, which are multicellular. In cases in which special gametangia are lacking, every cell produces a gamete. In oogamy, the male gametangia are called antheridia and the female oogonia or archegonia. A female gametangium with a sterile cellular jacket is called an archegonium although, like an oogonium, it produces eggs. In most of the plants dealt with in this article the eggs are produced in archegonia and the sperms in antheridia with surface layers of sterile cells.

The plant basis. Individual plants may be either bisexual (hermaphroditic), in which male and female gametes are produced by the same organism, or unisexual, producing either male or female gametes but not both. A bisexual individual, however, is not necessarily capable of fertilizing its own eggs. In certain ferns, for example, male gametes of one individual are not compatible with the female gametes of the same individual, so that cross-fertilization (with another individual of the species) is obligatory. This situation, of course, is similar in adaptive significance to cross-pollination (which leads to cross-fertilization) among seed plants.

Among the liverworts, mosses, and vascular plants, the life cycle involves two different phases, often called generations, although only one plant generation is, in fact, involved in one complete cycle. This type of life cycle is often said to illustrate the "alternation of generations" in which a haploid individual (*i.e.*, with one set of chromosomes), or tissue, called a gametophyte, at maturity produces gametes that unite in pairs to form diploid (*i.e.*, containing two sets of chromosomes) zygotes. The latter develop directly into individuals, or tissues, called sporophytes, in which the nuclei of certain fertile cells, called spore mother cells, or sporocytes, give rise to haploid spores (sometimes called meiospores). These spores are lightweight and are borne by air currents; they germinate to form the haploid, sexual, gamete-producing phase, usually designated the gametophyte.

The "alternation of generations"

There are several variations in the above described life cycle. The haploid gametophyte and sporophyte may be free-living, independent plants (*e.g.*, certain algae and yeasts), in which case the life cycle is said to be diplobiontic; or the sporophyte may be physically attached to the gametophyte, as it is in liverworts and mosses. By contrast, the gametophytic phases develop as parasites on the sporophytes of the seed plants, as in certain algae. In further variation, the alternating phases may be similar morphologically except for the type of reproductive cells, gametes or spores they produce (isomorphic life cycle); or they may be strikingly dissimilar, as in some algae, mosses, ferns, and seed plants (heteromorphic life cycle). Only heteromorphic life cycles occur in liverworts, mosses, vascular plants, and certain fungi.

The differences between the gametophyte and sporophyte are often great, especially those of the diplobiontic types, so that the alternates seem to be two different, unrelated individuals rather than different manifestations of the same organism.

BRYOPHYTE REPRODUCTIVE SYSTEMS

Liverworts and hornworts. The plant bodies of liverworts and hornworts represent the gametophytic (sexual) phase of the life cycle, which is dominant in these plants. In the liverworts, the sporophyte is borne upon or within the gametophyte but is transitory. Liverwort and hornwort plants, depending on the species, may be bisexual or unisexual, and the sex organs may be distributed on the surface (*Riccia, Ricciocarpus, Sphaerocarpos, Pellia*) or localized in groups and borne on special branches (antheriodiophores and archegoniophores) as in *Marchantia;* the sperms are biflagellate (Figure 5).

Release of the mature sperm and the process of fertilization require moisture in the form of heavy dew or raindrops. In all but a few genera (*Riccia, Ricciocarpus*),

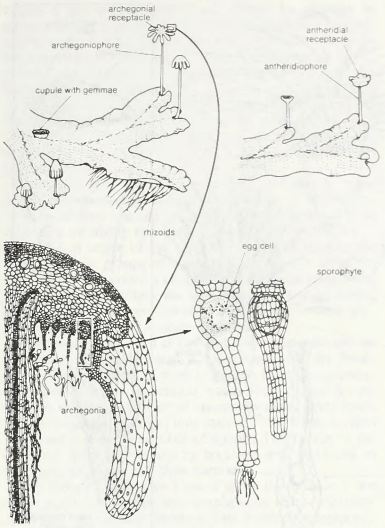


Figure 5: Archegonia and antheridia in the liverwort *Marchantia*.

From (top) *Botany*, 3rd ed. by Carl L. Wilson and Walter E. Loomis: Copyright 1952, © 1957, 1962 by Holt, Rinehart and Winston, Inc., reproduced by permission of Holt, Rinehart and Winston, Inc.; (bottom) T.E. Weir *et al.*, *Botany* (1967), John Wiley & Sons, Inc.

the developing sporophytes are actively photosynthetic—*i.e.*, capable of utilizing light energy to form organic substances. They are, however, dependent on gametophytic tissues for water (and the inorganic salts dissolved in it) and probably derive and utilize in their nutrition some organic substances manufactured by the gametophytes. Liverwort spores are meiospores; *i.e.*, they arise by meiosis from cells called sporocytes.

The sporophytes may consist almost completely of fertile (sporogenous) tissues (*Riccia, Oxymitra*), or they may contain sterile cells (nurse cells or elaters) among the developing spores. In *Marchantia* and *Porella*, a sterile foot and seta, or stalk, are present; the foot anchors the spore-bearing capsule (sporangium) to the gametophyte and also probably serves an absorptive function. The seta connects the foot and capsule. The elongation of the seta raises the capsule from its protective envelopes, thus, placing it in a favourable position for spore dispersal. The capsules of liverworts may shed their spores only by decay of the capsule wall and gametophytic tissues (*Riccia, Oxymitra*), or they may open irregularly or into two or four segments.

Spore germination in some species may occur immediately after deposition if the spores are in a favourable environment; or, as in other species, the spores may require a period of dormancy before germination.

Mosses. In mosses, as in liverworts and hornworts, the leafy shoots belong to the gametophytic phase and produce sex organs when they mature (Figure 6). The leafy shoots (often called gametophores, because they bear the sex organs) arise from a preliminary phase called the protonema, the direct product of spore germination. Filamentous, straplike, or membranous, it grows along the soil surface. A protonema of a moss may proliferate, apparently indefinitely, under favourable conditions and thus increase the population of leafy shoots that arise as buds. Under adverse conditions, certain buds and branches of the protonema may thicken their walls and thus serve to tide the species over an unfavourable growing period.

The gametophores of mosses

The antheridia and archegonia may be borne at the tips (apices) of the main shoots or on special, lateral branchlets. Both bisexual and unisexual leafy shoots occur, de-

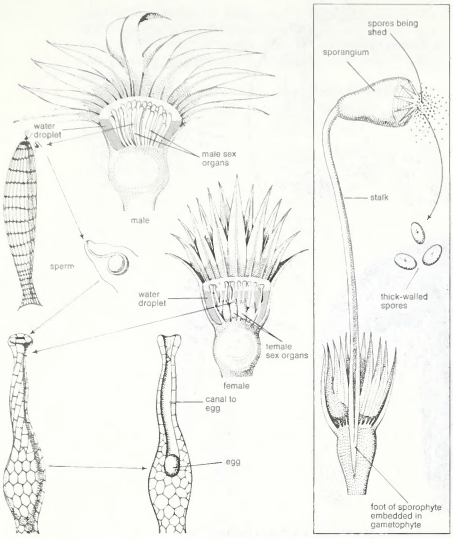


Figure 6: *The two phases of a moss plant.*
(Left) Sectioned tips of male and female moss gametophyte plants. (Right) Mature sporophyte releasing spores that will develop into gametophyte.

From *Biological Science: An Inquiry into Life*, 2nd ed. (1968); Harcourt Brace Jovanovich, Inc., New York; by permission of the Biological Sciences Curriculum Study

pending on the species. In a number of mosses (*Mnium, Polytrichum, Funaria*), the sexually mature shoots become recognizable through the production of special, prominent leaves that form an apical cup around the sex organs. If brightly coloured, the cup is often flowerlike. In species with bisexual leafy gametophores, the archegonia and antheridia may be present on the same apex (as can be seen, for example, in *Bryum*) or at the apices of separate branches as is exemplified in the moss *Funaria*.

The archegonia and antheridia of mosses are large enough in many species to be just barely visible to the unaided eye. The jacket cells of the antheridia are often coloured bright orange or rust; their sperm are biflagellate. As in liverworts and hornworts, rains and even heavy dews evoke the liberation of sperm and the opening of the mature archegonia so that fertilization may be accomplished.

The moss sporophyte, which is attached to the gametophyte, photosynthesizes during much of its development and is more or less self-supporting. It is, to a certain degree, dependent upon the gametophyte for nutrients such as water and mineral salts and, in some cases, even for elaborated foods.

After elongation of the moss sporophyte has ceased, the distal portion (farthest away) enlarges to form the capsule (sporangium), or spore-bearing region. The spores (meiospores), which arise by meiosis, are shed from the capsules gradually through a variety of mechanisms. After the operculum (cover) of the capsule has been shed, its mouth is usually partially closed by the peristome (teeth) and sometimes by associated structures. These teeth absorb moisture, and their resultant swelling and contraction open spaces through which the spores are shed.

TRACHEOPHYTE REPRODUCTIVE SYSTEMS

Spore plants. In liverworts, hornworts, and mosses, the dominant phase in the life cycle is the sexual gametophyte. In the tracheophytes (vascular cryptogams and seed plants), on the other hand, the sporophyte is the dominant phase in the life cycle. The gametophytes of the vascular cryptogams mature after the spores that initiated them have been shed from the parent plant, so that the gametophytes are free-living. In the seed plants the gametophytes mature as parasites on the sporophytes.

Psilopsids. The trilobed sporangia of the whisk fern *Psilotum*—not a true fern but a psilopsid—are borne terminally on short lateral branches. During development, some of the potentially spore-bearing tissue is used as nutrient by the sporocytes as they complete the meiotic divisions that result in colourless kidney-shaped spores. The latter, which are shed as the sporangia open along three lines, germinate and slowly develop into cylindrical, sparingly branched gametophytes, about 0.5 to two millimetres (0.02 to 0.08 inch) in diameter and several millimetres long. They presumably derive nourishment from decaying matter and occur in humus-rich soil, in rock fissures, or among roots on the trunks of tree ferns. The cells of the gametophyte contain fungal structures (hyphae) that probably are involved in some type of nutritional relation with the gametophyte.

The gametophytes are bisexual and the sperms multiflagellate. The embryonic sporophyte is not easily distinguished from the gametophyte that bears it. At first anchored to the gametophyte by an absorptive foot, the sporophyte ultimately becomes separated from both the foot and the gametophyte.

Lycopsids. In the genus *Lycopodium*, the sporangia are closely associated with the leaves. In some species (*L. lucidulum*), the sporangium-bearing leaves (sporophylls) occur in zones among the vegetative portions of the stems. In most, however, the sporophylls occur in specialized compressed stems, called cones, or strobili (Figure 7). Each sporophyll is associated with one yellow to orange, kidney-shaped sporangium.

From *Biological Science: An Inquiry into Life*, 2nd ed. (1968); Harcourt Brace Jovanovich, Inc., New York; by permission of the Biological Sciences Curriculum Study

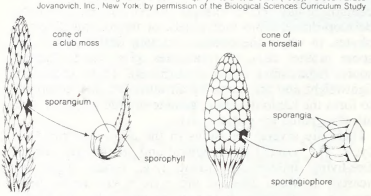


Figure 7: Sporangial protection of a club moss and a horsetail.

In several species the spores develop rapidly on the soil surface into ovoid-cylindrical gametophytes about two to three millimetres (0.08 to 0.12 inch) long, with green lobes and colourless bases; they usually contain a fungus. In other species, development of a colourless gametophyte is slow, so that at maturation, which may require up to eight years, the fleshy gametophyte will have become buried in successive layers of humus. These subterranean gametophytes, which contain fungi, are long-lived and are larger (up to two centimetres) than the surface types.

The gametophytes of *Lycopodium* are bisexual, although the antheridia and archegonia may develop into separate groups. The sperms are biflagellate and apparently more than one egg of the same gametophyte may be fertilized.

The zygote divides at a right angle to the long axis of the archegonium. The inner cell gives rise to the embryo, which thus is oriented as if it will develop within the gametophyte; it turns 180° during later development, however, and the axis grows vertically outward from the gametophyte.

The heterosporous condition

In contrast to *Lycopodium*, all *Selaginella* sporophytes have sporophylls localized in strobili, and all species of *Selaginella* are heterosporous: that is, they produce spores of two sizes, the larger designated as megaspores and the smaller as microspores. The megaspores develop into female gametophytes and the microspores into male gametophytes. Accordingly, strobili bear megasporophylls that contain megasporangia, which will produce megaspores, and microsporophylls that contain microsporangia, which will yield microspores. Although the evolutionary origin

of two kinds of spores (dimorphism) is unknown, the development of megaspores in living plants suggests that differences in nutrition in the two kinds of sporangia are significant. In a microsporangium, most of the microsporocytes undergo meiosis, forming four spores each; by contrast, all but one or, occasionally, several of the sporocytes in the megasporangium do not complete development. As a result, only four megaspores usually mature in such a sporangium, enlarging as they become gorged with the nutrients made available by disintegration of the other cells. The megaspores, accordingly, are much larger than microspores, although both contain stored food. Both types of spores are thick walled, and both have prominent three-part (triradiate) ridges.

Unlike the homosporous spores of most liverworts, hornworts, mosses, ferns, and *Lycopodium,* the spores of *Selaginella* begin to develop into gametophytes before they have been shed from their sporangia and attain maturity on a suitable, moist substrate.

The microscopic male gametophyte is composed essentially of a single antheridium, which produces biflagellate sperm. The female gametophyte, which protrudes after the megaspore wall cracks open in the region of the triradiate ridge, consists of vegetative cells, has several archegonia at maturity, and usually has three groups of rhizoids. Both male and female gametophytes lack the chlorophyll (green pigment) necessary for photosynthesis; they utilize nutrients stored in the spores.

After fertilization, one zygote of each female gametophyte develops into an embryonic sporophyte. There is considerable variation in details of development among the species of *Selaginella.* In some, the spores may develop mature gametophytes before they are shed from their sporangia, and fertilization may occur, so that female gametophytes with embryos may be found in the strobili (compressed stems, or cones). The megaspores of *Selaginella,* containing female gametophytes with still-attached juvenile sporophytes, have the superficial appearance of germinating seeds, from which, however, they differ in many significant respects.

Isoetes, like *Selaginella,* is monoecious and heterosporous. Most of the leaves are fertile; some bear one large megasporangium each, and others support a single microsporangium on the inner surface of a spoonlike leaf base. The microsporangia can produce enormous numbers of microspores—as many as 1,000,000—and the megasporangia give rise to 50 to 300 megaspores. The spores are liberated as the older sporophylls decay. Unlike those of *Selaginella,* the spores of *Isoetes* do not germinate until they have been shed from their sporangia. The unisexual gametophytes are much like those of *Selaginella,* but the sperm are multiflagellate. The embryonic sporophyte is nourished by food stored in the megaspore and transported through a massive foot.

Sphenopsids. The perennial sporophytes of horsetails (*Equisetum* species) produce strobili once during every growing season. They may be borne at the tips of green shoots (*E. hyemale, E. kansanum*); at the tips of non-green shoots that become green after the spores have been shed (*E. fluviatile, E. sylvaticum*); or on special nongreen branches that wither and die after the spores have been shed (*E. arvense, E. talmateia*). The appendages of the strobilus are often called sporangiophores and have been considered to be both stem branches and of leafy origin; in the latter case, they are called sporophylls (Figure 7). Each sporangiophore bears a number of fingerlike sporangia, which produce large numbers of thin-walled, green spores. The outermost wall layer of the spore breaks down into four appendages, which, by their sensitivity to moisture, coil and uncoil, thereby disseminating the spores.

The spores of *Equisetum* germinate rapidly and grow into green, pincushion-like gametophytes anchored to the surface by rhizoids. Apparently, two types of gametophytes are produced from the homosporous spores; some mature slowly, are smaller than others, and always produce antheridia, never archegonia. Others are larger and hermaphroditic, producing archegonia at first and, later, antheridia. The ratios of male to hermaphroditic gametophytes vary among species but are relatively uniform within a species. The ratios are altered by changes in environmental conditions; for example, at certain temperatures (*e.g.,* 32° C, or 90° F) only male gametophytes develop from the spores of five species; whereas at 15° C (59° F) approximately 50 percent are male and 50 percent hermaphroditic gametophytes.

Environmental influences on sex ratios

Self-fertilization of hermaphroditic gametophytes can occur, and several sporophytes may be produced on one gametophyte. The embryo consists of an absorptive foot, a primary root or radicle, and a shoot with whorled appendages.

Ferns. As they mature, many fern sporophytes begin to produce spores in clusters of sporangia on the undersurfaces of their vegetative "leaves." Others produce their sporangia on highly modified leaves or portions thereof.

The site of origin of the sporangia is the receptacle; the latter, with its groups of sporangia, is called a sorus. In many ferns each sorus is covered with a special outgrowth, the indusium; in others, the sporangia are covered during development by the margin of the leaf. In a few ferns (*e.g., Polypodium*), the sori remain uncovered.

In primitive ferns, such as *Ophioglossum* and *Botrychium,* the spores are borne upon a specialized axis, the fertile spike. The sporangia of such primitive ferns are massive, with several layers of cellular walls, and produce an indefinite but large number of spores. In most other ferns, the sporangia are smaller, long stalked, with single-layered walls and a definite number of spores. The spores of the latter are shed explosively by breakage and shrinking as the sporangia open and then slam shut.

Homospory and heterospory in ferns

Most ferns produce one kind of spore (homospory), but a few genera of aquatic and amphibious ferns (*Marsilea, Salvinia,* and *Azolla*) produce two kinds (heterospory), small microspores and much larger megaspores. In either case, after being shed from the parent sporophyte, the spores that have suitable environmental conditions germinate and develop into the gametophytic phase. The ribbonlike, filamentous or heart-shaped gametophytes of most ferns contain chlorophyll, are anchored to some surface—moist soil, moist rocks, or tree bark—by unicellular root-like rhizoids, and rarely exceed one-half inch (13 millimetres) in diameter. In a few ferns (*Ophioglossum, Botrychium,* and certain species of *Schizaea*), the gametophytes are subterranean, lack chlorophyll, are cylindrical or tuberous, and contain the filamentous structures (hyphae) of an associated fungus.

Fern gametophytes, often called prothalli (singular, prothallus) are one cell layer thick except in the centre. Most fern prothalli are bisexual—*i.e.,* have both male (antheridia) and female (archegonia) sex organs, which develop usually on the undersurface of the prothallus.

Although the eggs of several archegonia may be fertilized, only one zygote usually develops into a juvenile sporophyte. The latter consists of an absorbing foot; a primary root, or radicle, which promptly penetrates the surface; a prominent first leaf; and a rudimentary, slowgrowing stem. As the juvenile sporophyte becomes established, the parental gametophyte dies. The series of leaves formed from the stem of the juvenile sporophyte gradually attain the form and vein pattern that characterize the mature sporophyte.

In most ferns, the antheridia appear before the archegonia and continue to develop as the latter mature; furthermore, the archegonial necks curve toward the mature antheridia so that fertilization can readily occur. Both gametes may be derived from one individual, or from different individuals. In the bracken fern (*Pteridium aquilinum*), although the gametophytes are bisexual, self-incompatability factors reduce self-fertilization. In *Onoclea sensiblis,* the gametophytes are unisexual in early development, thus favouring cross-fertilization; but, later, the gametophytes become bisexual so that, if cross-fertilization fails, the species can still be maintained.

Seed plants. In the two great groups of seed plants, gymnosperms and angiosperms, the sporophyte is the dominant phase in the life cycle, as it is also in the vascular cryptogams; the gametophytes are microscopic parasites on the sporophytes (Figure 8).

In the gymnosperms, the seeds occur individually, ex-

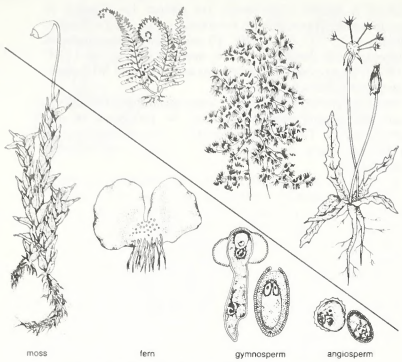


Figure 8: Change in relative size of the gametophyte (below the line) and sporophyte (above the line) in the course of plant evolution. In the moss, the sporophyte is entirely dependent upon the gametophyte.

Reprinted from *The Spectrum of Life* by Harold A. Moore and John R. Carlock. Copyright © 1970, Harper & Row, Publishers, Inc.

posed at the ends of stalks, sometimes in whorls on an axis, or on the scales of a cone, or megastrobilus. In angiosperms, or flowering plants, by contrast, the seeds are enclosed during development in a structure variously termed a pistil or carpel, which is sometimes considered to represent an enfolded megasporophyll.

Events in the reproductive process

A number of parts of the reproductive process are common to both angiosperms and gymnosperms: (1) they produce seeds at maturity; (2) the megasporangium, unlike that of heterosporous seedless plants, is covered by one or two cellular layers called integuments and is termed an ovule; (3) there is a minute passageway, or micropyle, through the integuments; (4) the ovule matures as a seed; (5) only one megasporocyte is present and undergoes meiosis in the megasporangium to produce four megaspores, only one of which usually is functional; (6) the megaspore is never discharged from its megasporangium and ovule; (7) one female gametophyte is produced within each megasporangium and ovule; (8) the microspores begin their development into male gametophytes while still enclosed in the microsporangia; (9) as they mature, the male gametophytes, which are contained within the microspore wall and are termed pollen grains, develop a tube that conveys sperm to the egg cell; (10) union of sperm and egg and development of an embryonic sporophyte from the zygote occur within the female gametophyte (sometimes called the "embryo sac"), which is covered by the remains of the megasporangium and by integuments; (11) as the embryo develops, the ovule matures as a seed.

In contrast to this impressive list of similarities are important differences, which, in addition to seed position, serve to distinguish angiosperms from gymnosperms. The reproductive cycle in most angiosperms is completed quicker than that in gymnosperms, and the gametophytes are smaller and simpler and, unlike those of most gymnosperms, lack archegonia. The pollen in angiosperms is transferred to the surface of the megasporophyll, whereas in gymnosperms it is brought to the micropyle of the ovule itself. Two sperm are involved in the sexual union in angiosperms: one unites with the egg to form a zygote; the other unites with two nuclei of the female gametophyte to form the primary endosperm nucleus. The latter divides to form a postfertilization storage tissue, which serves as a food source for the embryo; the embryo of gymnosperms is nourished by the somatic (nonreproductive) tissues of the female gametophyte. The angiosperm ovule increases to mature seed size after fertilization, whereas in gymnosperms, this enlargement occurs prior to fertilization.

The general features of the reproduction of seed plants having now been summarized, certain special aspects of the reproduction in representative seed plants are described below.

Gymnosperms. The cycads are slow-growing dioecious gymnosperms, the microsporangia (potential pollen) and megasporangia (potential ovules) occurring on different individual sporophytes. In all cycads except the genus *Cycas*, the ovules are borne on megasporophylls in megastrobili; in *Cycas* the ovules develop on individual, leaflike megasporophylls in what is regarded as a primitive arrangement. The microspores of all cycads develop in microstrobili.

The microspores reach the three-celled stage of development of the male gametophyte before they are shed as pollen grains from the microsporangia. At this time, elongation of the megastrobilus separates the megasporophylls, and the wind-borne pollen grains have access to the micropyles of the ovules. At the time of pollination each ovule exudes a mucilaginous droplet, the pollination droplet, through the micropyle; some of the pollen grains become engulfed in this droplet and are drawn into the ovule.

The interval between pollination and fertilization is several months in cycads. The sperm are multiflagellate and are among the largest (about 300 microns, or 0.01 inch) in the plant kingdom. Each pollen tube may contain from two to 22 sperm, depending on the genus. The pollen tubes, which develop from the pollen grains, work their way through the megasporangium of the ovule to the archegonia of the female gametophyte. Fertilization of the eggs of the several archegonia is followed by the early development of several embryos (polyembryony), only one of which survives in the mature seeds. Cycad embryos produce two seed leaves, or cotyledons. The seeds are brightly coloured (yellow or scarlet) and covered by an outer fleshy layer and a stony layer of the integument. The seeds of some cycads (*e.g., Cycas*) may germinate in the megastrobilus without a period of dormancy.

System of the ginkgo

The maidenhair tree, or ginkgo (*Ginkgo biloba*), is sometimes classified with the conifers (see below) or separately in a group of which it is the sole living representative. The mature ginkgo (sporophyte) produces microstrobili and ovules each spring as the buds unfold. They occur on the spur shoots among the bases of the young leaves. The ginkgo, like the cycads, is strictly dioecious, so that some trees produce ovules and others produce pollen. The ovules occur in pairs at the tips of stalks that emerge among the leaf bases.

Ginkgo pollen, like that of pines, is four-celled at the time of pollination (spring season), which is accomplished by wind. Development of male and female gametophytes is similar to that in cycads, and the sperm are also multiflagellate. The female gametophyte, within the ovule of *Ginkgo*, is unique among seed plants in containing chlorophyll. The ovules enlarge tremendously after pollination, and as the seeds mature the integument differentiates into several coats, of which a stony layer and outer fleshy layer are most prominent. The latter becomes mottled, purplish green and foul smelling. Its tissues may cause nausea or skin eruptions in man. The inner tissues of the seed (the embryo and the female gametophyte) are palatable and prized among some peoples. Fertilization often occurs after the ovules have fallen from the trees, three or four months after pollination. The ginkgo embryo has two cotyledons.

The sporophytes of most of the species of living conifers, like those of the ginkgo, are woody trees at maturity. They usually grow for a number of years beyond the seedling stage before they mature and produce seeds.

The sporophyte of a typical conifer, such as a pine, may become a large tree. Unlike the cycads and ginkgo, a pine is monoecious, both microstrobili and megastrobili occurring on the same tree. At the beginning of each growing season, the microstrobili enlarge and emerge from their bud scales; they are borne at the base of the terminal bud, which is destined to develop into the current season's growth. The megastrobili, by contrast, arise singly or in a whorl near the apex of the current season's growth.

The microstrobili are called simple strobili because the microsporangia are borne in pairs on the appendages (microsporophylls) that emerge from the axis of the strobilus. The megastrobili, however, are compound, for the ovules

are borne in pairs upon the upper (adaxial) surface of scales, which, in turn, are borne on bracts attached to the megastrobilus.

The pollen of pine, four-celled when shed, is characterized by two lateral, air-filled "wings," enlarged cavities between two layers of the pollen-grain wall. The pollen is produced in large amounts and may be transported great distances by air currents. During the time of pollination, the ovuliferous scales on the megastrobili separate slightly, and pollen can be trapped in the pollination droplet of the micropyles of the ovules. Pollen grains that make contact with a droplet are transferred by its subsequent contraction through the micropyle and to the surface of a small depression (pollen chamber) at the tip of the megasporangium.

As the pollen grain germinates, forming a tube that works its way through the megasporangium, it arrives at the female gametophyte as the latter matures its several archegonia. The pollen tubes discharge their sperm nuclei into the archegonia, and fertilization is accomplished. As in the cycads and ginkgo, the zygotes of several archegonia may initiate embryogeny. Furthermore, in pine and certain other conifers, the young embryos may form several embryos. At maturity of the seed, however, only one embryo is normally present, embedded in the remains of the female gametophyte and megasporangium, all surrounded by the seed coat (the former integument).

The reproductive process in pine occupies two full growing seasons: ovules pollinated in the spring of a given year do not mature as seeds until the late summer of the next year. The interval between pollination and fertilization is about 14 months.

Reproductive variations in gymnosperms

Among the numerous other gymnosperm species are many different reproductive processes. Some gymnosperms, for example, are dioecious, with microstrobili and megastrobili being borne on separate plants, as in junipers (*Juniperus*), plum yews (*Cephalotaxus*), yews (*Taxus*), and podocarps (*Podocarpus*). Furthermore, in larches (*Larix*) and other groups the pollen grains lack wings. The pollen grains in larches become attached at pollination to a special receptive enlargement of the integument. In podocarps, the megasporangium bulges through the micropyle at pollination and receives the pollen directly. The interval between pollination and fertilization may be as short as four to five weeks in firs (*Abies*). The number of ovules formed on the ovuliferous scale varies, as does the number of microsporangia on the microsporophyll. There may be only one ovule in a megastrobilus, as in some junipers, and the megastrobili may become fleshy, also in junipers. In yews the solitary ovules are terminal on dwarf shoots; each ovule is surrounded by a cuplike structure called an aril, which becomes fleshy and brightly coloured as the seed matures. The number of sperm produced in each male gametophyte varies also—from two in pine to 20 in some cypresses (*Cupressus*).

The genera *Ephedra, Gnetum,* and *Welwitschia,* which are often grouped together in one category (Gnetales, or Gnetophyta), differ among themselves and from other gymnosperms with respect to several details of reproduction. The microsporangia and ovules of both *Ephedra* and *Welwitschia* are produced in compound strobili; those of *Gnetum* are borne in a series of whorls on elongated axes sometimes called misleadingly "inflorescences." The ovules of these genera, unlike those of other gymnosperms, have two integuments instead of one, as in angiospermous ovules. Archegonia are present in the female gametophytes of *Ephedra,* but only eggs occur in those of *Gnetum* and *Welwitschia.* The sperm, like those of the conifers, lack flagella.

Angiosperms. Although the angiosperms are known as flowering plants, they are difficult to distinguish from gymnosperms solely on the basis of bearing flowers, for, like the strobilus, a flower is a compressed stem, with crowded, spore-bearing appendages. The occurrence of coloured petals and attractive scents is not essential and is by no means characteristic of all flowers. The most important distinguishing feature separating flowering plants from gymnosperms is that the ovules of flowering plants

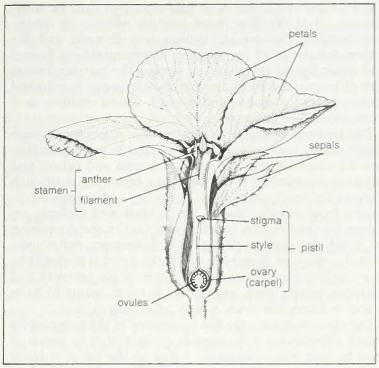


Figure 9: Gross morphology of the flower.

Reprinted from *The Spectrum of Life* by Harold A. Moore and John R. Carlock. Copyright © 1970, Harper & Row, Publishers, Inc.

are produced within enclosed containers called carpels (Figure 9).

Flowers may occur singly at the ends of stems (*e.g.,* tulip, poppy, rose), or they may be grouped in various clusters or inflorescences (gladiolus, sunflower, delphinium, and yucca).

An individual flower may be complete, in that a given floral receptacle produces sepals (often greenish and leaflike), petals (often white or coloured other than green), stamens, and a pistil (or pistils). The sepals are collectively known as the calyx, and the petals as the corolla; the calyx and corolla comprise the perianth. If sepals or petals are lacking, the flower is said to be incomplete. Although incomplete, a flower that has both stamens and a pistil is said to be perfect; lacking either of these parts, it is imperfect.

In practice, groups of solitary flowers are not easily distinguished from inflorescences; the latter seemingly evolved from a system of branches, each with a terminal, solitary flower. The inflorescence may be few flowered or have up to 6,000,000 flowers, as in certain palms. Inflorescences vary also in their position, being terminal, axillary, or intercalary. Terminal inflorescences are at the tips of the major or dominant branches; axillary ones are at the tips of axillary, or side, branches. In intercalary inflorescences, the stem continues beyond the inflorescence, which may result in alternating fertile and sterile areas of the axis.

Types of floral clusters

Inflorescences can be distinguished by their growth patterns as determinate or indeterminate. In determinate inflorescences the first formed flower at the tip of the dominant stem matures first, and younger flowers develop on lower lateral branches; the cyme of the forget-me-not (*Myosotis*) is a typical example. In indeterminate inflorescences the growing region of the axis functions for extended periods so that as the older flowers mature and set fruit near the base of the inflorescence axis, younger buds develop and continue to expand into flowers at the apex. This is exemplified in the spikes of yucca and the racemes of delphinium, in which the youngest flowers are farthest away from the root. Other types of indeterminate inflorescences include umbels and capitula, or heads. The youngest flower is terminal or central in umbels and in heads.

The head is the type of inflorescence that characterizes the composite, or aster family. It may be few to many flowered and usually has at its base one or more series of leaflike bracts. The small individual flowers arise in spiral order on the receptacle, the youngest being at the centre. The basal calyx of each flower, known as a pappus, is bristle-like, scaly, or feathery and borne at the top of the ovary. The corolla, formed of the petals, may be: (1) tubular, with five petal lobes, sometimes split open; (2) ligulate, or tonguelike, with a very short basal tube; or

(3) bilabiate, with the tube split into two tips. In some genera, all of the flowers are ligulate, whereas in others, the marginal flowers are ligulate (ray flowers) and the others tubular or all are tubular. The marginal ray flowers are either female (pistillate) or sterile. The tubular flowers are characterized by male and female parts: five united pollen-bearing stamens and a pistil, which matures as a one-seeded fruit (achene).

The position of the floral organs with reference to each other and to the tip of the floral receptacle varies in different flowers; in some, the perianth (sepals and petals) and stamens are attached to the receptacle below the pistil; such flowers are hypogynous (*e.g.,* buttercup and magnolia). In others (rose, cherry, peach), the perianth and stamens are borne on the rim of a concave structure in the depression of which the pistil is borne; such flowers are perigynous. Finally, there are flowers in which the ovary is enclosed by a tissue composed of the fused bases of the perianth and stamens (apple, pear, aster); the blossom seems to arise upon or above the ovary and is called epigynous.

The stamen, seemingly the equivalent of the gymnospermous microsporophyll, consists of an anther (a group of two to four microsporangia) borne at the tip of a blade stalk, or filament. The pistil, most often composed of an enlarged basal ovary, a columnar style, and distal stigma, is the ovule-producing organ of the flower. It is often considered to have evolved from enfolded megasporophyll or some other ovuliferous structure with enclosed ovules (angiospermy); alternatively, it is thought to have arisen from the cuplike bracts of extinct seed-bearing plants on which the leafy bracts grew together and thus enclosed the ovules.

There may be one or more pistils on the floral receptacle, depending on the species. Furthermore, pistils may be simple (composed of one ovule-bearing unit, megasporophyll, or carpel) or compound (composed of more than one carpel). Compound pistils are thought to have arisen as a result of crowding of simple pistils on the floral axis; for example, variation in the degree of fusion may be observed in members of the saxifrage family. The ovary—which matures as the fruit—usually reveals by the number of ovule-containing chambers (locules) the number of carpels it contains. The stigma is a specially adapted portion of the pistil modified for the reception of pollen. It may be feathery and branched or elongated, as in such wind-pollinated flowers as those of the grasses, or it may be compact and with a sticky surface. The ovary may contain one ovule (*e.g.,* buckwheat, avocado), a few ovules (*e.g.,* grape, bean) or a large number of ovules (tobacco, begonia, snapdragon).

In some angiosperms (*e.g.,* corn, hickory, walnut, pecan, oak) both types of imperfect flower are borne on the same plant, which is, therefore, called monoecious. By contrast, staminate flowers may occur on one plant and pistillate flowers on another, as in willows, poplars, and mulberries, which are said to be dioecious. In common parlance (and unfortunately in some botanical textbooks) staminate flowers and plants that bear them are often designated "male," and pistillate flowers and the plants that bear them are called "female." This may be traced back at least as far as to the time of Linnaeus (1753), who interpreted stamens and pistils as sex organs. Comparative morphology indicates clearly, however, that stamens and pistils are the spore-bearing structures of the sporophyte and not actually the gamete-bearing organs of the gametophyte. The terms "male" and "female," applied to angiosperm plants and their flowers, is often condoned because the gametophytic phase is so condensed in angiosperms. The designations suggest to the uninitiated, however, that pollen grains and sperm, on the one hand, and eggs and ovules, on the other, are identical, which is not the case.

Floral organization

Among the vast number of species of angiosperms there is considerable variation in floral organization. The perianth may be absent or present; clearly differentiated as calyx and corolla (*e.g.,* pea); or the perianth segments may be similar (magnolia, tulip tree). The number of stamens and pistils may be large and separately attached to the receptacle in a spiral pattern (buttercup), or the numbers may be reduced and the attachment cyclic or whorled (lily). The stamens may be fused together by their anthers (daisy) or their filaments (peas, beans). The filaments may be petallike (water lilies) or stalklike. Opening of the anther may be by longitudinal or transverse fissures or by terminal pores.

The reproductive cycle in angiosperms can be traced from before the shedding of pollen (Figure 10). The microspores begin their development of male gametophytes, which involves formation of a small generative cell and a tube cell. The generative cell may divide to form two sperm before the pollen grain (developing male gametophyte) is shed, or while the pollen tube is growing during germination. The pollen grains of angiosperms have variously, and often elaborately, ornamented walls characteristic of the species.

From (top) P.B. Weisz, *The Science of Biology,* 4th ed. (1971), used with permission of McGraw-Hill Book Company. (bottom) P.B. Weisz, *The Science of Biology,* 3rd ed. (1967), used with permission of McGraw-Hill Book Company

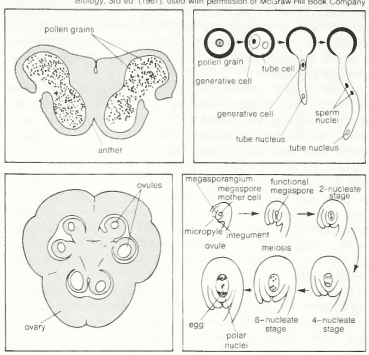


Figure 10: *Gametophyte development in a flowering plant (lily).* (Top left) Cross section through the anther of a lily. (Top right) Microgametophyte from the microspore. (Bottom left) Cross section through the ovary of a lily. (Bottom right) megagametophyte from the megaspore.

Pollination in angiosperms is the transfer of the pollen grains from the anther of a stamen to the stigma of a pistil.

The pistil of a flower may receive pollen from the stamens of the same flower, in self-pollination (*e.g.,* peas and tomatoes). In many other flowers, however, pollen from one or more flowers is transferred to the stigmas of other flowers. A number of specialized relationships have evolved between floral organization and animal pollinators such as insects. (See below *Pollination* for a complete treatment of the processes and mechanisms of pollination in plants.)

In the majority of angiosperms one megasporocyte develops in the megasporangium (often called the nucellus) of the ovule, and a tetrad of megaspores is formed as a result of meiosis. Three megaspores (nearest the micropyle) degenerate; only one enlarges, and then undergoes divisions to form the eight-nucleate, seven-celled female gametophyte ("embryo sac"). Of the three cells of this gametophyte near the micropyle, one functions as an egg. As the pollen tube discharges its contents into the female gametophyte, the egg nucleus is fertilized by one of the sperm, the other unites with the two nuclei (polar nuclei) within the large central cell of the female gametophyte. The resultant nucleus, which has three sets of chromosomes, is the primary endosperm nucleus. This process, double fertilization, occurs only in angiosperms.

Pattern of development

Both pollination and fertilization stimulate cell division in the ovary, ovules, and zygotes, all of which enter upon a period of rapid enlargement. In most angiosperms, the primary endosperm nucleus divides to form endosperm tissue, the cells of which become filled with stored food, such as starches, oils, and proteins. As the rate of embryonic development decreases, the seeds of most angiosperms enter a period of dormancy, accompanied by dehydration and hardening of the integuments, which form seed coats.

At this period, the enlarged ovary (and sometimes adjacent structures) matures as fruit. (Further discussion of the form, function, and development of seeds and fruits is presented below, in the section *Seed and fruit.*)

Angiosperm seeds may germinate as soon as they reach maturity, or they may undergo various kinds of dormancy. In some cases (*e.g.,* coconut) the embryo is rudimentary and undifferentiated when the seed is shed, so that a period of preparation, or after-ripening, is required. In other types of dormancy, germination is retarded by the hardness and impermeability of the seed coat or by special requirements of light, temperature, and moisture.

Some representative variations occur in the reproductive process of angiosperms. In violets (*Viola*), in addition to the ordinary flowers produced first during the usual flowering season, less conspicuous flowers later develop; called cleistogamous flowers, they do not open but are self-pollinated, thus insuring augmentation of the population during a period less favourable for the usual blossoms.

The pollen grains of most angiosperms separate from each other, but in some cases (*e.g., Rhododendron*) they remain attached in original groups of four, called tetrads. The very tiny pollen grains of orchids, certain mimosas, and milkweeds are clustered in waxy masses called pollinia (singular pollinium).

A number of variations in pattern of development of the female gametophyte occur in various angiosperms; for example, in certain species of evening primrose (*Oenothera*), the female gametophyte contains only four nuclei, whereas, in *Peperomia,* as many as 16 may be present. In lily, all four megaspore nuclei are involved in the formation of the female gametophyte.

Pollen may germinate immediately after contact with a stigma (sugar cane), within five minutes (corn), in two hours (beet), or after one or two days. The pollen grains of most plants produce only one pollen tube, but ten or more pollen tubes have been observed to develop from one pollen grain in plants of the mallow family. The pollen tubes usually enter through the micropyle (porogamy), but they may also enter through the base of the ovule (chalazogamy).

The interval between pollination and fertilization varies. It may be as long as 12 to 14 months in certain species of oak, five to seven months in witch hazel; two to 20 weeks among the orchids; three to four hours in lettuce and as little as 15 to 45 minutes in dandelions.

The postfertilization endosperm fails to develop in orchid seeds but is present at least during early embryogeny in most others. The endosperm may arise by nuclear divisions and become cellular as nuclear divisions terminate, or its development may involve both nuclear and cell divisions from the beginning. In a number of cases (*e.g.,* legumes) the embryo consumes the endosperm during its development, resulting in mature seeds with massive embryos and no endosperm. Most angiosperm embryos have two seed leaves (are dicotyledonous), some have one lateral cotyledon (are monocotyledonous), and a few (*e.g., Degeneria*) have three to four cotyledons.

In seed germination, the cotyledons may remain below the soil surface within the seed (hypogean germination) and may function in digesting and absorbing endosperm (corn), may serve as sources of stored food themselves (pea), or may rise above the soil surface (epigean germination) by elongation of the hypocotyl, the embryonic axis between the root, the growing stem, or epicotyl. Cotyledons that emerge above the soil may wither and drop off as their food is used (*e.g.,* bean), or they may persist and function as photosynthetic leaves (*e.g.,* castor bean).

An even greater range of variation occurs in angiospermous fruits. The fruit may arise from one pistil (simple or compound) of one flower (*i.e.,* the simple fruits of pea and peach), from several pistils of one flower (*i.e.,* the aggregate fruits of strawberry and raspberry), or from the pistils of several flowers (*i.e.,* the multiple fruits of pineapple, mulberry, and corn). Simple fruits may be dry (legumes) or fleshy (peach, apple, tomato) at maturity. Dry fruits may open (dehisce; many legumes) or remain closed about the seed (be indehiscent; grasses and sunflower).

The manner of ovular attachment is known as placentation. The ovary may contain one to many ovules, which may be attached to the ovary wall (parietal placentation) or to the central axis (axial, or free-central placentation). Despite these and other variations in the morphology of flower parts, the reproductive process is, with minor diversities, remarkably uniform.

VARIATIONS IN REPRODUCTIVE CYCLES

The life cycles and reproductive processes described above characterize the vast majority of their respective plant groups.

Among the liverworts it has been demonstrated that small fragments of the stalk of the sporophyte are capable of regenerating diploid gametophytes. In the mosses, both haploid and diploid apospory have been experimentally evoked. As in the liverworts, injury and regeneration of fragments of the sporophytic seta result in diploid gametophytes. By contrast, fragments of moss leaves, stems, and rhizoids (and even the sterile tissues of the sex organs) can regenerate haploid gametophytes.

In certain strains of mosses, the gametophyte can give rise to clusters of presumably haploid sporophytes without the functioning of gametes; such apogamous formation of sporophytes may also be chemically induced (by application of a solution containing a specific amount of chloral hydrate to both the protonema and leafy shoots).

Among the vascular plants both natural and induced apogamy and apospory are known. In certain ferns, gametophytes may develop at the leaf margins or in sori from transformed sporangia. Certain other ferns reproduce apogamously in nature; thus, for example, in the holly fern (*Crytomium falcatum*), the gametophytes give rise directly to sporophytes by nuclear and cell division on vegetative cells of the gametophyte. In almost every group, however, variations of the usual reproductive process occur. These may involve substitution of asexual reproduction for sexual or the direct production of plants by cells other than the usual ones (apomixis). Apomictic phenomena—which are in the strictest sense asexual—include apospory, in which the gametophyte phase is produced without the need of spores, and apogamy, in which the sporophyte phase is produced without the need of gametes, or sex cells.

Apogamy, apospory, and apomixis

Apogamy may be induced in normally sexual ferns by withholding water from the gametophytes, which prevents the liberation and functioning of sperm. Similarly, when gametophytes are grown in inorganic culture media supplemented by a variety of sugars, they produce sporophytes apogamously. Colourless roots removed from the bracken fern (*Pteridium aquilinum*) have been induced to develop diploid gametophytes aposporously, as have the injured juvenile leaves of a number of ferns.

Apomictic phenomena occur also among many angiosperms. In some species, haploid sporophytes may develop either from the unfertilized egg or from some other cell of the gametophyte. Such apogamy occurs, for example, after stimulation of one species with the pollen of a related one (*e.g., Solanum nigrum* by the pollen of *S. luteum*). Apogamy involving an unfertilized egg (a phenomenon termed parthenogenesis) occurs in certain orchids. Male parthenogenesis, or the production of a sporophyte from a sperm, has been detected in tobacco hybrids. Finally, a form of haploid apogamy is known in which a cell of the female gametophyte other than an egg may develop into an embryo.

In certain species of hawkweed, the embryo develops from a certain cell of the ovule or the megasporangium. In others, the female gametophyte is diploid through an impairment of the meiotic process; in this case, the egg (diploid parthenogenesis) or one of the related cells may form an embryo. In citrus trees a number of embryos (polyembryony) arise from diploid cells of the megasporangium or integuments.

PHYSIOLOGY OF PLANT REPRODUCTION

The maturation of sporophytes and gametophytes, as manifested by their ability to produce spores and gametes respectively, involves both internal and environmental factors. With respect to the former, the organism must have completed a certain minimum period of vegetative devel-

opment before environmental factors are able to stimulate formation of spores and gametes.

Effects of light and temperature

Among environmental factors affecting reproduction, the duration, intensity, and quality of light, as well as temperature, have primary roles; for example, the liverwort *Marchantia polymorpha* continues in the vegetative state indefinitely under daily fluorescent illumination of 16 hours. Control plants exposed to daily incandescent lighting of 16 hours become sexually mature after 30 days. Addition of the sugar sucrose hastens the development of the sexual phase, an indication that chemical factors also play a role. In hornworts (*e.g., Anthoceros* and *Phaeoceros*), antheridia develop under daily light periods (photoperiods) of four to 12 hours; none develop when the plants are illuminated for longer periods.

Temperature also affects sexual maturation. In mosses, for example, initiation of sex organs in bacteria-free laboratory cultures of *Funaria* occurs at 10° C (50° F) when cultures are illuminated six, 12, or 20 hours daily. On the other hand, the moss *Polytrichum* is seemingly not affected by duration of light but forms sex organs best at 21° C (70° F).

Among vascular cryptogams little is known about the various environmental factors that affect development of sex organs in nature, and, except for certain ferns and horsetails, experimental studies of rigidly controlled laboratory cultures are lacking. It has been shown that laboratory cultures of gametophytes of certain horsetails derived from single spores produce sex organs 40 to 60 days after the spores germinate. In horsetails generally, conditions favouring vegetative growth evoke a preponderance of initially female gametophytes; less favourable conditions induce the production of male gametophytes.

Considerable information is available about the physiology of reproduction in ferns, especially with respect to the gametophyte generation. Spore germination occurs if adequate moisture is present and in temperatures between 15° and 35° C (59° and 95° F). The spores germinate and the young gametophytes do best under neutral or slightly acid conditions. Most fern spores require light for germination. Some fern spores remain viable for as long as 20 years.

Spores germinating in darkness or in the absence of blue wavelengths of light remain in the protonemal, or filamentous, condition. Growth to the mature gametophyte requires light of blue wavelengths in most species.

An interesting aspect of fern reproductive physiology was the discovery that antheridial production is under hormonal control. Several types of antheridium-evoking substances (hormones) have been recognized, and there is some evidence that they are specific. The sperm of ferns are attracted chemically to the vicinity of the maturing archegonia, where they become trapped in extruded mucilage. Although only one sporophyte usually develops on a gametophyte, as many as five may be induced to form if the gametophyte is repeatedly exposed to sperm.

The preceding paragraphs, dealing with vascular cryptogams, have emphasized the influence of environmental factors on the expression of sexual reproduction; the remaining account deals with the angiosperms. Because the gametophytic phase of angiosperms is abbreviated and because it is parasitic on the sporophyte, the maturation of the sporophyte, as manifested by flowering, has received more intensive study in angiosperms than has that of the gametophyte.

Durations of life cycles

Angiosperms that complete their life cycles within one growing season (in the temperate zone) are known as annuals. Perhaps the shortest angiospermous life cycle known is that of *Plantago insularis*, a southern California species, that can grow from seed to the production of its own seed in four to six weeks under domestication; the cycle of most annuals, however, is longer. A number of angiosperms are biennial in the temperate zone: they grow vegetatively for one season, but their flowering and seed production are delayed until a second growing season, after which the plants die (*e.g.,* beets, carrots). Still other angiosperms are perennial: they continue growing, flowering, and producing seeds for a number of growing seasons (*e.g.,* irises, roses, oaks).

It has been demonstrated that duration and wavelength of light are of paramount importance in controlling the flowering of angiosperms. Based on extensive experimental studies, flowering plants have been classified as "long-day," "short-day," or "day-neutral" with respect to their requirements of light for flowering.

Temperature also plays an important role in flowering. Thus, members of the cabbage family can be grown without flowering when high enough temperatures are maintained. A number of perennial and biennial plants (bulbous plants, beets) require a prior period of low temperature before they flower. In addition to internal (*i.e.,* genetic factors), therefore, temperature and light are of paramount importance for many plants in evoking maturation.

Several other physiological aspects of reproduction in angiosperms are noteworthy. Once transported to the stigma, the germination of the pollen grains is markedly affected by the chemical composition of the stigmatic exudate; *e.g.,* pollen of an unrelated species will not germinate on the stigma. There is evidence also that the directional growth of the pollen tubes through the style toward the ovular micropyle is stimulated by a chemical substance produced by the style, ovules, and probably other tissues of the pistil.

It is also clear that both pollination and fertilization have physiological effects on fruit production. In a number of cases pollination alone (not followed by fertilization) is sufficient stimulus to evoke enlargement of the pistil to form a seedless (parthenocarpic) fruit. This phenomenon occurs in the banana and in certain varieties of citrus fruits, grapes, and cucumbers. Parthenocarpy can also be induced by exposure of the stigma to indoleacetic acid, naphthoxyacetic acid, indole butyric acid, naphthaleneacetic acid, or other synthetic hormones. The hormone gibberellin is effective in producing seedless grapes and is the active component in preparations used to prevent premature dropping of fruit.

Also significant in the reproduction of flowering plants are the phenomena of self-sterility and self-incompatibility, which prevail in certain plants with perfect flowers (having both stamens and pistils). Pollen from one flower or from another flower of the same plant or from a plant with identical genetic constitution, when applied to the stigma, fails to germinate or grows so slowly that fertilization does not occur. For example, sweet cherries are self-incompatible: a number of groups of genetically intrasterile types are known, but cross-pollination of trees of different genetic groups results in fruit production. Self-incompatibility occurs also in certain strains of garden plants and in plum, apple, pear, and tobacco.

Self-sterility, usually associated with differences in chromosome number, often occurs in hybrids from widely divergent parents. This is true of cultivated varieties of blackberry and raspberry. Such chromosomal imbalance creates irregularities in meiosis during formation of the pollen, which results in infertility. (H.C.B.)

Pollination

Pollination is the transference of pollen grains from the stamens, the flower parts that produce them, to the ovule-bearing organs or to the ovules (seed precursors) themselves. In conifers and cycads, in which the ovules are exposed, the pollen is simply caught in a drop of fluid secreted by the ovule. In flowering plants, however, the ovules are contained within a hollow organ called the pistil, and the pollen is deposited on the pistil's receptive surface, the stigma. There the pollen germinates and gives rise to a pollen tube, which grows down through the pistil toward one of the ovules in its base. In an act of double fertilization, one of the two sperm cells within the pollen tube fuses with the egg cell of the ovule, making possible the development of an embryo, and the other cell combines with the two subsidiary sexual nuclei of the ovule, which initiates formation of a reserve food tissue, the endosperm. The growing ovule then transforms itself into a seed. As a prerequisite for fertilization, pollination is essential to the production of fruit and seed crops and plays an important part in programs designed to improve

Importance to agriculture

plants by breeding. Furthermore, studies of pollination are invaluable for understanding the evolution of flowering plants and their distribution in the world today. As sedentary organisms, plants usually must enlist the services of external agents for pollen transport. In flowering plants, these are (roughly in order of diminishing importance): insects, wind, birds, mammals, and water.

Pollination by insects probably occurred in primitive seed plants, reliance on other means being a relatively recent evolutionary development. Reasonable evidence indicates that flowering plants first appeared in tropical rain forests during the Mesozoic Era (about 65,000,000 to 225,000,000 years ago). The most prevalent insect forms of the period were primitive beetles; no bees and butterflies were present. Some Mesozoic beetles, already adapted to a diet of spores from primitive plants, apparently became pollen eaters, capable of effecting chance pollination with grains accidentally spared. The visits of such beetles to primitive flowering plants may have been encouraged by insect attractants, such as odours of carrion, dung, or fruit, or by sex attractants. In addition, visits of the insects to the plants could be made to last longer and thus potentially be more valuable to the plant as far as fertilization was concerned, if the flower had a functional, traplike structure. Nowadays, such flowers are found predominantly, although not exclusively, in tropical families regarded as ancient; *e.g.,* the water lily (Nymphaeaceae) and the arum lily (Araceae) families. At the same time, other plants apparently began to exploit the fact that primitive gall-forming insects visited the flowers to deposit eggs. In the ancient genus *Ficus* (figs and banyan trees), pollination still depends on gall wasps. In general, Mesozoic flowering plants could not fully rely on their pollinators, whose presence also depended on the existence of a complete, well-functioning ecological web with dung, cadavers, and food plants always available. More advanced flowers escaped from such dependence on chance by no longer relying on deceit, trapping, and tasty pollen alone; nectar became increasingly important as a reward for the pollinators.

Development of nectar

Essentially a concentrated, aqueous sugar solution, nectar existed in certain ancestors of the flowering plants. In bracken fern even nowadays, nectar glands (nectaries) are found at the base of young leaves. In the course of evolutionary change, certain nectaries were incorporated into the modern flower (floral nectaries), although extrafloral nectaries also persist. Flower colours thus seem to have been introduced as "advertisements" of the presence of nectar, and more specific nectar guides (such as patterns of dots or lines, contrasting colour patches, or special odour patterns) were introduced near the entrance to the flower, pointing the way to the nectar hidden within. At the same time, in a complex pattern of parallel evolution, groups of insects appeared with sucking mouthparts capable of feeding on nectar. In extreme cases, there arose a complete mutual dependence. For example, a Madagascar orchid, *Angraecum sesquipedale,* with a nectar receptacle 20 to 35 centimetres (eight to 14 inches) long, depends for its pollination exclusively on the local race of a hawkmoth, *Xanthopan morganii,* which has a proboscis of 22½ centimetres (nine inches). Interestingly enough, the existence of the hawkmoth was predicted by Charles Darwin and Alfred Russel Wallace, codiscoverers of evolution, about 40 years before its actual discovery.

TYPES: SELF-POLLINATION AND CROSS-POLLINATION

Self-pollination and cross-pollination

An egg cell in an ovule of a flower may be fertilized by a sperm cell derived from a pollen grain produced by that same flower or by another flower on the same plant, in either of which two cases fertilization is said to be due to self-pollination (autogamy); or, the sperm may be derived from pollen originating on a different plant individual, in which case the process is called cross-pollination (heterogamy). Both processes are common, but cross-pollination clearly has certain evolutionary advantages for the species: the seeds formed may combine the hereditary traits of both parents, and the resulting offspring generally are more varied than would be the case after self-pollination. In a changing environment, some of the individuals resulting from cross-pollination still may be found capable of coping with their new situation, ensuring survival of the species, whereas the individuals resulting from self-pollination might all be unable to adjust. Self-pollination, or selfing, although foolproof in a stable environment, thus is an evolutionary cul-de-sac. There also is a more direct, visible difference between selfing and outbreeding (cross-pollination): in those species where both methods work, cross-pollination usually produces more, and better quality, seeds. A dramatic demonstration of this effect is found with hybrid corn (maize), a superior product that results from cross-breeding of several especially bred lines.

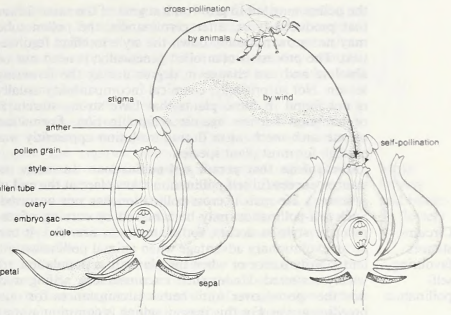


Figure 11: Types of pollination.
Drawing by M. Moran based on B.J.D. Meeuse, *The Story of Pollination*

Mechanisms that prevent self-pollination. *Structural.* Not surprisingly, many species of plants have developed mechanisms that prevent self-pollination. Some—*e.g.,* date palms (*Phoenix dactylifera*) and willows (*Salix* species)—have become dioecious; that is, some plants produce only "male" (staminate) flowers, with the rest producing only "female" (pistillate or ovule-producing) ones. In species in which staminate and pistillate flowers are found on the same individual (monoecious plants) and in those with hermaphroditic flowers (flowers possessing both stamens and pistils), a common way of preventing self-fertilization is to have the pollen shed either before or after the period during which the stigmas on the same plant are receptive, a situation known as dichogamy. The more usual form of dichogamy, which is found especially in such insect-pollinated flowers as fireweed (*Epilobium angustifolium*) and salvias (*Salvia* species), is protandry, in which the stamens ripen before the pistils. Protogyny, the situation in which the pistils mature first, occurs in arum lilies and many wind-pollinated plants, such as grasses—although several grasses are self-pollinated, including common varieties of wheat, barley, and oats. Avocado has both protogynous and protandrous varieties, and these often are grown together to encourage cross-fertilization. A structural feature of flowers that discourages selfing is heterostyly, or variation in the length of the style (neck of the pistil). This occurs in the common primrose (*Primula vulgaris*) and species of wood sorrel (*Oxalis*) and flax. In most British primrose populations, for example, approximately half the individuals have so-called "pin" flowers, which possess short stamens and a long style, giving the stigma a position at the flower's mouth, whereas the other half have "thrum" flowers, in which the style is short and the stamens are long, forming a "thrumhead" at the opening of the flower. Bees can hardly fail to deposit the pollen they receive from one type of flower onto the stigmas of the other type. The genetic system that regulates flower structure in these primroses is so constituted that cross-pollination automatically maintains a 50:50 ratio between pins and thrums. In the flowers of purple loosestrife (*Lythrum salicaria*), the stamens and styles are of three different lengths to limit self-fertilization.

Chemical. Chemical self-incompatibility is another device for preventing self-fertilization. In this phenomenon, which depends on chemical substances within the plant,

the pollen may fail to grow on a stigma of the same flower that produced it or, after germination, the pollen tube may not grow normally down the style to effect fertilization. The process is controlled genetically; it need not be absolute and can change in degree during the flowering season. Not surprisingly, chemical incompatibility usually is not found in those plants that have strong structural or temporal barriers against self-pollination. Formation of one such mechanism during evolution apparently was enough for most plant species.

Mechanisms that permit self-pollination. In many instances, successful self-pollination takes place at the end of a flower's life-span if cross-pollination has not occurred. Such self-pollination may be achieved by curving of stamens or style as occurs, for example, in fireweed. It can be an evolutionary advantage when animal pollinators are temporarily scarce or when the plants in a population are widely scattered. Under such circumstances, selfing may tide the species over until better circumstances for outbreeding arrive. For this reason, selfing is common among annual plants; these often must produce an abundance of seed for the rapid and massive colonization of any bare ground that becomes available. If, in a given year, an annual plant were to produce no seed at all, survival of the species might be endangered. A persistent habit of self-pollination apparently has been adopted successfully by some plant species whose natural pollinators have died out. Continued selfing also is practiced by many food-crop plants. Some of these plants are cleistogamous, meaning that the flowers fail to open, an extreme way of ensuring self-pollination. A similar process is apomixis, the development of an ovule into a seed without fertilization. Apomixis is easily demonstrated in lawn dandelions, which produce seeds even when stamens and styles are cut off just before the flowers open. Consistent apomixis has the same pros and cons as continued selfing. The offspring show very little genetic variability, but there is good survival if the species is well adapted to its habitat and if the environment does not change.

Circumstances favouring self-pollination

AGENTS OF POLLEN DISPERSAL

Insects. *Beetles and flies.* The ancient principle of trapping insects as a means of ensuring pollination was readopted by some advanced families (*e.g.,* orchids and milkweeds), and further elaboration perfected the flower traps of primitive families. The cuckoopint (*Arum maculatum*), for example, attracts minute flies, which normally breed in cow dung, by means of a fetid smell. This smell is generated in early evening, along with considerable heat, which helps to volatilize the odour ingredients. The flies visiting the plant, many of which carry *Arum* pollen, enter the floral trap through a zone of bristles and then fall into a smooth-walled floral chamber from which escape is impossible. Gorging themselves on a nutritious stigmatic secretion produced by the female flowers at the base of the chamber, the flies effect cross-pollination. Late at night, when the stigmas no longer function, the male flowers, situated much higher on the floral column, proceed to bombard the flies with a rain of pollen. The next day, when smell, heat, and food are gone, the prisoners, "tarred" with stigmatic secretion and "feathered" with pollen, are allowed to escape by a wilting of the inflorescence (flower cluster). Usually the escaped flies are soon recaught by another inflorescence, which is still in the smelly, receptive stage, and cross-pollination again ensues. Superb timing mechanisms underlie these events. The heat-generating metabolic process in the inflorescence is triggered by a hormone, calorigen, originating in the male flower buds only under the right conditions. The giant inflorescences of the tropical plant *Amorphophallus titanum* similarly trap large carrion beetles.

Trap flowers

In general, trap flowers victimize beetles or flies of a primitive type. Although beetles most likely were involved as pollinators when flowering plants as a group were born, their later performance in pollination has been disappointing. Some modern beetles do visit smelly flowers of an open type, such as elderberry and hawthorn, but with few exceptions they are still mainly pollen eaters. Flies as a group have become much more diversified in their habits than beetles have. Female short-tongued flies may be deceived by open-type flowers with carrion smells; *e.g.,* the flowers of *Stapelia* and *Rafflesia.* Mosquitoes with their long tongues are effective pollinators of certain orchids (*Habenaria* species) in North American swamps. In Europe, the bee fly (*Bombylius*) is an important long-tongued pollinator. Extremely specialized as nectar drinkers are certain South African flies; for example, *Megistorhynchus longirostris,* which has a tongue that is 60 to 70 millimetres (2.3 to 2.7 inches) long.

Pollen flowers

The voraciousness of flower beetles demonstrates the futility of enticing insect pollinators solely with such an indispensable material as pollen. As a defensive strategy, certain nectar-free flowers that cater to beetles and bees—such as wild roses, peonies, and poppies—produce a superabundance of pollen. Other plants—*e.g., Cassia*—have two types of stamens, one producing a special sterile pollen used by insects as food, the other yielding normal pollen for fertilizing the ovules. Other flowers contain hairs or food bodies that are attractive to insects.

Bees. In the modern world, bees are probably the most important insect pollinators. Living almost exclusively on

Drawing by M. Pahl based on (E) photograph courtesy of B.J.D. Meeuse. (A) reprinted with permission of Macmillan Publishing Co., Inc. from *A Textbook of General Botany,* 3rd ed., by G.M. Smith *et al.* Copyright 1935 by The Macmillan Company, renewed 1963 by Genevieve Allan, Helen P. Smith and Katherine N. Bryan; (B, F, G) B.J.D. Meeuse. *The Story of Pollination* (Copyright © 1961), The Ronald Press Company, New York. (C, D) *An Evolutionary Survey of the Plant Kingdom* by R.F. Scagel, G.E. Rouse, J.R. Stein, R.J. Bandoni, W.B. Schofield, T.M.C. Taylor. © 1965 by Wadsworth Publishing Company, Inc., Belmont, California 94002. Reprinted by permission of the publisher

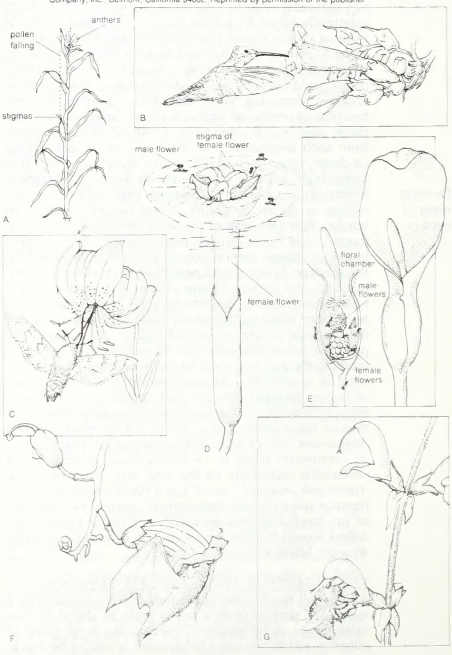


Figure 12: *Examples of different pollination modes:* (A) Pollination by wind in maize, or Indian corn (*Zea mays*). (B) Bird pollination of a trumpet-vine flower by a hummingbird. (C) Hawkmoth taking nectar from petal pouch of lily and pollinating at the same time. (D) Water pollination of *Vallisneria americana* in which floating male flowers come in contact with stigmata of larger female flower. (E) A trap flower (*Arum italicum*) pollinated by insects trapped in the flower structure for a time. (F) Bat pollinating a flower of the sausage tree (*Kigelia*). (G) Pollination of *Salvia pratensis* by bumblebees, which are touched on the back by either the stigma or the anthers, depending on which stage the flower is in.

nectar, they feed their larvae pollen and honey (a modified nectar). To obtain their foods, they possess striking physical and behavioral adaptations, such as tongues as long as 2½ centimetres (one inch), hairy bodies, and (in honeybees and bumblebees) special pollen baskets. The Austrian naturalist Karl von Frisch has demonstrated that honeybees, although blind to red light, distinguish at least four different colour regions, namely, yellow (including orange and yellow green), blue green, blue (including purple and violet), and ultraviolet. Their sensitivity to ultraviolet enables bees to follow nectar-guide patterns not apparent to the human eye. They are able to taste several different sugars and also can be trained to differentiate between aromatic, sweet, or minty odours but not foul smells. Fragrance may be the decisive factor in establishing the honeybee's habit of staying with one species of flower as long as it is abundantly available. Also important is that honeybee workers can communicate to one another both the distance and the direction of an abundant food source by means of special dances.

Bee flowers

Bee flowers, open in the daytime, attract their insect visitors primarily by bright colours; at close range, special patterns and fragrances come into play. Many bee flowers provide their visitors with a landing platform in the form of a broad lower lip on which the bee sits down before pushing its way into the flower's interior, which usually contains both stamens and pistils. The hermaphroditism of most bee flowers makes for efficiency, because the flower both delivers and receives a load of pollen during a single visit of the pollinator, and the pollinator never travels from one flower to another without a full load of pollen. Indeed, the floral mechanism of many bee flowers permits only one pollination visit. The pollen grains of most bee flowers are sticky, spiny, or highly sculptured, ensuring their adherence to the bodies of the bees. Since one load of pollen contains enough pollen grains to initiate fertilization of many ovules, most individual bee flowers produce many seeds.

Examples of flowers that depend heavily on bees are larkspur, monkshood, bleeding heart, and Scotch broom. Alkali bees (*Nomia*) and leaf-cutter bees (*Megachile*) are both efficient pollinators of alfalfa; unlike honeybees, they are not afraid to trigger the explosive mechanism that liberates a cloud of pollen in alfalfa flowers. Certain Ecuadorian orchids (*Oncidium*) are pollinated by male bees of the genus *Centris;* vibrating in the breeze, the beelike flowers are attacked headlong by the strongly territorial males, who mistake them for competitors. Other South American orchids, nectarless but very fragrant, are visited by male bees (*Euglossa* species) who, for reasons not yet understood, collect from the surface of the flowers an odour substance, which they store in the inflated parts of their hindlegs.

Wasp flowers

Wasps. Few wasps feed their young pollen or nectar. Yellow jackets, however, occurring occasionally in large numbers and visiting flowers for nectar for their own consumption, may assume local importance as pollinators. These insects prefer brownish-purple flowers with easily accessible nectar, such as those of figwort. The flowers of some Mediterranean and Australian orchids mimic the females of certain wasps (of the families Scoliidae and Ichneumonidae) so successfully that the males of these species attempt copulation and receive the pollen masses on their bodies. In figs, it is not the pollinator's sexual drive that is harnessed by the plant but the instinct to take care of the young; tiny gall wasps (*Blastophaga*) use the diminutive flowers (within their fleshy receptacles) as incubators.

Butterflies and moths. The evolution of moths and butterflies (Lepidoptera) was made possible only by the development of the modern flower, which provides their food. Nearly all species of Lepidoptera have a tongue, or proboscis, especially adapted for sucking. The proboscis is coiled at rest and extended in feeding. Hawkmoths hover while they feed, whereas butterflies alight on the flower. Significantly, some butterflies can taste sugar solutions with their feet. Although moths, in general, are nocturnal and butterflies are diurnal, a colour sense has been demonstrated in representatives of both. Generally, the colour sense in Lepidoptera is similar to that of bees, but swallowtails and certain other butterflies also respond to red colours. Typically, colour and fragrance cooperate in guiding Lepidoptera to flowers, but in some cases there is a strong emphasis on just one attractant; for example, certain hawkmoths can find fragrant honeysuckles hidden from sight.

Moth flowers and butterfly flowers

Typical moth flowers—*e.g.,* jimsonweed, stephanotis, and honeysuckle—are light-coloured, often long and narrow, without landing platforms. The petals are sometimes fringed; the copious nectar is often in a spur. They are open and overwhelmingly fragrant at night. Butterfly flowers—*e.g.,* those of butterfly bush, milkweed, and verbena—are conspicuously coloured, often red, generally smaller than moth flowers, but grouped together in erect, flat-topped inflorescences that provide landing space for the butterflies.

Important pollinating moths are the various species of the genus *Plusia,* sometimes occurring in enormous numbers, and the hummingbird hawkmoth (*Macroglossa*), which is active in daylight. A small moth, *Tegeticula maculata,* presents an interesting case. It is totally dependent on yucca flowers, in whose ovules its larvae develop. Before depositing their eggs, the females pollinate the flowers, following an almost unbelievable pattern of specialized behaviour, which includes preparing a ball of pollen grains and carrying it to the stigma of the plant they are about to use for egg laying.

Wind flowers

Wind. Although prevalent in the primitive cycads and in conifers, such as pine and fir, wind pollination (anemophily) in the flowering plants must be considered as a secondary development. It most likely arose when such plants left the tropical rain forest where they originated and faced a more hostile environment, in which the wind weakened the effectiveness of smell as an insect attractant and the lack of pollinating flies and beetles also made itself felt. Lacking in precision, wind pollination is a wasteful process. For example, one male plant of *Mercurialis annua,* a common weed, produces 1,250,000,000 grains of pollen to be dispersed by the wind; a male sorrel plant produces 400,000,000. Although, in general, the concentration of such pollen becomes very low about one-fourth mile (0.4 kilometre) from its source, nonetheless in windy areas it can cover considerable distances. Pine pollen, for example, which is naturally equipped with air sacs, can travel up to 500 miles (800 kilometres) although the grains may lose their viability in the process. Statistically, this still gives only a slim chance that an individual stigma will be hit by more than one or two pollen grains. Also relevant to the number of pollen grains per stigma is the fact that the dry, glueless, and smooth-surfaced grains are shed singly. Since the number of fertilizing pollen grains is low, the number of ovules in a single flower is low and, as a consequence, so is the number of seeds in each fruit. In hazel, walnut, beech, and oak, for example, there are only two ovules per flower, and, in stinging nettle, elm, birch, sweet gale, and grasses, there is only one. Wind-pollinated flowers are inconspicuous, being devoid of insect attractants and rewards, such as fragrance, showy petals, and nectar. To facilitate exposure of the flowers to the wind, blooming often takes place before the leaves are out in spring, or the flowers may be placed very high on the plant. Inflorescences, flowers, or the stamens themselves move easily in the breeze, shaking out the pollen, or the pollen containers (anthers) burst open in an explosive fashion when the sun hits them, scattering the pollen widely into the air. The stigmas often are long and divided into arms or lobes, so that a large area is available for catching pollen grains. Moreover, in open areas wind-pollinated plants of one species often grow together in dense populations. The chance of self-pollination, high by the very nature of wind pollination, is minimized by the fact that many species are dioecious or (like hazel) have separate male and female flowers on each plant. Familiar flowering plants relying on wind pollination are grasses, rushes, sedges, cattail, sorrel, lamb's-quarters, hemp, nettle, plantain, alder, hazel, birch, poplar, and oak. (Tropical oaks, however, may be insect-pollinated.)

Birds. Because the study of mechanisms of pollination began in Europe, where pollinating birds are rare, their

importance is often underestimated. In fact, in the tropics and the southern temperate zones, birds are at least as important as pollinators as insects are, perhaps more so. About a third of the 300 families of flowering plants have at least some members with ornithophilous flowers; *i.e.*, flowers attractive to birds. Conversely, about 2,000 species of birds, belonging to 50 families, visit flowers more or less regularly to feed on nectar, pollen, and flower-inhabiting insects or spiders. Special adaptations to this way of life, in the form of slender, sometimes curved, beaks and tongues provided with brushes or shaped into tubes, are found in 1,615 species of eight families: hummingbirds, sunbirds, honey eaters, brush-tongued parrots, white-eyes, flowerpeckers, honeycreepers (or sugarbirds), and Hawaiian honeycreepers. Generally, the sense of smell in birds is poorly developed and not used in their quest for food; instead, they rely on their powerful vision and their colour sense, which resembles that of man (ultraviolet not being seen as a colour, whereas red is). Furthermore, the sensitivity of the bird's eye is greatest in the middle and red part of the spectrum. This is sometimes ascribed to the presence in the retina of orange-red drops of oil, which together may act as a light filter.

Bird flowers

Although other explanations have been forwarded, the special red sensitivity of the bird eye is usually thought to be the reason why so many bird-pollinated flowers are of a uniform, pure red colour. Combinations of complementary colors, such as orange and blue, or green and red, also are found, as are white flowers. As might be expected, bird flowers generally lack smell and are open in the daytime; they are bigger than most insect flowers and have a wider floral tube. Bird flowers also are sturdily constructed as a protection against the probing bill of the visitors, with the ovules kept out of harm's way in an inferior ovary beneath the floral chamber or placed at the end of a special stalk or behind a screen formed by the fused bases of the stamens. The latter, often so strong as to resemble metal wire, are usually numerous, brightly coloured, and protruding, so that they touch a visiting bird on the breast or head as it feeds. The pollen grains often stick together in clumps or chains, with the result that a single visit may result in the fertilization of hundreds of ovules. In America, where hummingbirds usually suck the nectar of flowers on the wing, ornithophilous flowers (*e.g.*, fuchsias) are often pendant and radially symmetrical, lacking the landing platform of the typical bee flower. In Africa and Asia, bird flowers often are erect and do offer their visitors, which do not hover, either a landing platform or special perches in the form of small twigs near the flower. Pollinating birds are bigger than insects and have a very high rate of metabolism. Although some hummingbirds go into a state resembling hibernation every night, curtailing their metabolism drastically, others keep late hours. Thus, in general, birds need much more nectar per individual than insects do. Accordingly, bird flowers produce nectar copiously—a thimbleful in each flower of the coral tree, for example, and as much as a liqueur glassful in flowers of the spear lily (*Doryanthus*). Plants bearing typical bird flowers are: cardinal flower, fuchsia, red columbine, trumpet vine, hibiscus, strelitzia, and eucalyptus, and many members of the pea, orchid, cactus, and pineapple families.

Mammals. In Madagascar, the mouse lemurs (*Microcebus*), which are only ten centimetres (four inches) long, obtain food from flowers, and in Australia the diminutive marsupial honey possums and pygmy possums also are flower specialists. Certain highly specialized tropical bats, particularly *Macroglossus* and *Glossophaga*, also obtain most or all of their food from flowers. The *Macroglossus* (big-tongued) species of southern Asia and the Pacific are small bats with sharp snouts and long, extensible tongues, which carry special projections (papillae) and sometimes a brushlike tip for picking up a sticky mixture of nectar and pollen. Significantly, they are almost toothless. Colour sense and that sonar sense so prominent in other bats, seem to be lacking. Their eyesight is keen but, since they feed at night, they are probably guided to the flowers principally by their highly developed sense of smell. The bats hook themselves into the petals with their thumb claws and stick their slender heads into the flowers, extracting viscid nectar and protein-rich pollen with their tongues. The plants involved have, in the process of evolution, responded to the bats by producing large (sometimes huge) amounts of these foods. One balsa-tree flower, for example, may contain a full 10 grams (0.3 ounce) of nectar, and one flower from a baobab tree has about 2,000 pollen-producing stamens. Some bat flowers also provide succulent petals or special food bodies to their visitors. Another striking adaptation is that the flowers are often placed on the main trunk or the big limbs of a tree (cauliflory); or, borne on thin, ropelike branches, they dangle beneath the crown (flagelliflory). The pagoda shape of the kapok tree serves the same purpose: facilitation of the bat's approach. Characteristics of the flowers themselves include drab colour, large size, sturdiness, bell-shape with wide mouth and, frequently, a powerful rancid or urinelike smell. The giant saguaro cactus and the century plant (*Agave*) are pollinated by bats, although not exclusively, and cup-and-saucer vine (*Cobaea scandens*) is the direct descendant of a bat-pollinated American plant. Calabash, candle tree, and areca palm also have bat-pollinated flowers.

Bat flowers

Water. Although pollen grains can be made to germinate in aqueous sugar solutions, water alone in most cases has a disastrous effect on them. Accordingly, only a very few terrestrial plants, such as the bog asphodel of the Faroes, use rainwater as a means of pollen transport. Even in aquatic plants, water is seldom the true medium of pollen dispersal. Thus, the famous Podostemonaceae, plants that grow only on rocks in rushing water, flower in the dry season when the plants are exposed; pollination occurs with the aid of wind or insects or by selfing. Another aquatic plant, ribbon weed, sends its male and female flowers to the surface separately. There, the former transform themselves into minute sailboats, which are driven by the wind until they collide with the female flowers. In the Canadian waterweed, and also in pondweed (*Potamogeton*) and ditch grass (*Ruppia*), the pollen itself is dispersed on the water's surface; it is, however, still water-repellent. True water dispersal (hydrophily), in which the pollen grains are wet by water, is found only in the hornworts and eelgrasses. (B.J.D.M.)

Seed and fruit

A seed is the characteristic reproductive body of both angiosperms (flowering plants) and gymnosperms (conifers, cycads, and ginkgos). Essentially, it consists of a miniature, undeveloped plant (the embryo), which, alone or in the company of stored food for its early development after germination, is surrounded by a protective coat (the testa). Frequently small in size and making negligible demands upon their environment, seeds are eminently suited to perform a wide variety of functions the relationships of which are not always obvious: multiplication, perennation (surviving seasons of stress such as winter), dormancy (a state of arrested development), and dispersal. Pollination and the "seed habit" are considered the most important factors responsible for the overwhelming evolutionary success of the flowering plants, which number more than 300,000 species.

The superiority of dispersal by means of seeds over the more primitive method involving single-celled spores, lies mainly in two factors: the stored reserve of nutrient material that gives the new generation an excellent growing start and the seed's multicellular structure, which provides ample opportunity for the development of adaptations for dispersal, such as plumes for wind dispersal, barbs, and others.

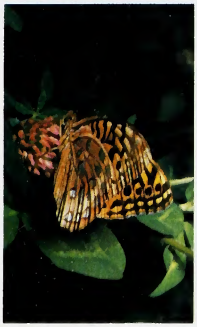
Economically, seeds (and the fruits that contain them) are important primarily because they are sources of a variety of foods; for example, the cereal grains, such as wheat, rice, and maize; the seeds of beans, peas, peanuts, soybeans, almonds, sunflowers, hazelnuts, walnuts, pecans, and Brazil nuts; the fruits of date palm, olive, banana, avocado, apple, and orange. Many fruits, especially those in the citrus family—limes, lemons, oranges, grapefruits—are rich in vitamin C (ascorbic acid); unpolished cereal grains in vitamin B_1 (thiamine); and wheat germ in vita-



Wind pollination: pollen of the lodgepole pine *(Pinus contorta)* and mountain hemlock *(Tsuga mertensiana)* on the surface of Crater Lake, Oregon.



Beetle pollination: pollen-eating longhorn beetle (family Cerambycidae) in a thimbleberry *(Rubus parviflorus)*.



Butterfly pollination: great spangled fritillary *(Speyeria cybele)*, proboscis extended, feeding on red clover *(Trifolium pratense)*.



Bee pollination: pollen-laden honeybee *(Apis mellifera)* on a wall flower *(Erysimum)*.



Wind pollination in grasses: yellow free-hanging anthers (pollen producers) and white feathery stigmas (pollen collectors) of meadow fescue *(Festuca pratensis)* provide maximum wind exposure.



Bird pollination: Anna's hummingbird *(Calypte anna)* feeding on nectar of a fuchsia.



Hawkmoth pollination: hawkmoth (Sphingidae) hovering near a honeysuckle *(Lonicera caprifolium)*.

Agents of pollen dispersal

Plate 1: (Top right, centre) E. S. Ross, (top left) Joe Van Wormer—Bruce Coleman Inc., (centre left) John H. Gerard, (centre right) Mary W. Ferguson, F.P.S.A., (bottom left) Walter E. Harvey—The National Audubon Society Collection/Photo Researchers, (bottom right) Ingmar Holmasen



Trap flower: pollinating psychodid flies trapped in the floral chamber of cuckoopint *(Arum maculatum;* vertical section with spathe cut away).



Deception flower: carrion flower *(Stapelia)* has the appearance and odour of decayed meat. Flies laying eggs crawl over flower's surface and come into contact with stigmas and stamens.



Trap flower: when certain bees land on the hairy, upright petal of the flower of grass-pink orchid *(Calopogon pulchellus),* the petal topples forward, and a visiting bee falls on its back onto the sticky surface of the column.



Nectar guides: spots in throat of foxglove *(Digitalis purpurea* variety *gloxiniaeflora)* provide insects with a visual pathway to nectar glands.



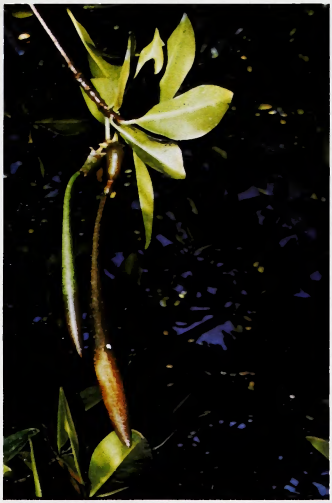




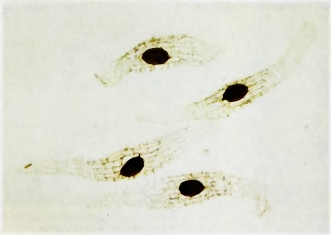
Explosive flower: (top) flower of Scotch broom *(Cytisus scoparius).* (Bottom left) The bumblebee *(Bombus vosnesenskii)* triggers release of the stamens (bottom right) which forcibly strike underside of bee and dust it with pollen.

Flower adaptations related to pollination

Plate 2: (Top left, top right) B.J.D. Meeuse, (centre) Rudolf Schmid, (centre left, bottom left, bottom right) E.S. Ross, (centre right) Derek Fell



Water dispersal: red mangrove seedling *(Rhizophora mangle)* ready to drop into the water after germinating on tree.



Wind dispersal: magnified view of seeds of lady's slipper *(Cypripedium)*, an example of the tiny seeds produced by orchids.



Water dispersal: a coconut *(Cocos nucifera)*, transported by the sea from a distant tropical island, germinating on a mainland beach.



Wind dispersal: woolly seeds produced by the seed pods of the kapok tree *(Ceiba pentandra)*.



Wind dispersal: winged fruits of the silver maple *(Acer saccharinum)*.



Self-dispersal: plumes on the fruits of mountain mahogany *(Cercocarpus)*, which coil and uncoil to drive seeds into the soil.

Wind, water, and self-dispersal of seeds and fruits

Plate 3: (Top left) William E. Ferguson, (top centre) Norman Myers—Bruce Coleman Inc., (top right) Thase Daniel, (centre left) Rudolf Schmid, (bottom left) G.R. Roberts, (bottom right) Dennis Brokaw



Naras melons *(Acanthosicyos horrida)* growing in the Namib Desert, South West Africa. Seeds are dispersed by the gemsbok oryx *(Oryx gazella)*, which feeds on the melons.



Seeds of rosary pea *(Abrus precatorius)*, which mimic fleshy red arils (accessory seed coverings) attractive to seed-eaters.



Acorns stored in the trunk of a digger pine *(Pinus sabiniana)* by the acorn woodpecker *(Melanerpes formicivorus)*.



Fleshy red appendages of fruits of the yew *(Taxus)*, which attract birds.

Seeds and fruit dispersed by animals



Epauletted fruit bat *(Epomophorus wahlbergi)* feeding on wild figs *(Ficus)*.



Curve-bill thrasher *(Toxostoma curvirostre)* with a berry in its beak.

Plate 4: (Top) Anthony Bannister from The Natural History Photographic Agency—EB Inc., (center left) Walter Dawn, (centre) William E. Grenfell, Jr., (centre right) Rudolf Schmid, (bottom left) Jane Burton—Bruce Coleman Inc., (bottom right) Lois Cox

min E. Other useful products provided by seeds and fruits are abundant. Oils for cooking, margarine production, painting, and lubrication are available from the seeds of flax, rape, cotton, soybean, poppy, castor bean, coconut, sesame, safflower, sunflower, the cereal grains of maize, and the fruits of olive and oil palm. Essential oils are obtained from such sources as juniper "berries," used in gin manufacture. Waxes such as those from bayberries (wax myrtles) and vegetable ivory from the hard fruits of a South American palm species (*Phytelephas macrocarpa*) are important products. Stimulants are obtained from such sources as the seeds of coffee, kola, guarana, and cocoa; and drugs, such as morphine, come from opium-poppy fruits. Spices—from mustard and nutmeg seeds; from the aril ("mace") covering the nutmeg seed; from fruits of anise, cumins, caraway, dill, vanilla, black pepper, red or chili pepper, allspice, and others—form a large group of economic products. Dyes (Persian berries, butternut brown, sap green) and ornaments (Job's tears from the grass *Coix*, used for curtains; *Abrus, Adenanthera,* and *Rhynchosia* seeds for necklaces; others for rosaries) are also provided by seeds and fruits.

THE NATURE OF SEEDS AND FRUITS

Seeds. *Angiosperm seeds.* In the typical flowering plant, or angiosperm, seeds are formed from bodies called ovules contained in the ovary, or basal part of the female plant structure, the pistil. The mature ovule contains in its central part a region called the nucellus that in turn contains an embryo sac with eight nuclei, each with one set of chromosomes (*i.e.,* they are haploid nuclei). The two nuclei near the centre are referred to as polar nuclei; the egg cell, or oosphere, is situated near the micropylar ("open") end of the ovule. (These reproductive structures are described above; see *Plant reproductive systems.*)

With very few exceptions (*e.g.,* the dandelion), development of the ovule into a seed is dependent upon fertilization, which in turn follows pollination. Pollen grains that land on the receptive upper surface (stigma) of the pistil will germinate, if they are of the same species, and produce pollen tubes, each of which grows down within the style (the upper part of the pistil) toward an ovule. The pollen tube has three haploid nuclei, one of them, the so-called vegetative, or tube, nucleus seems to direct the operations of the growing structure. The other two, the generative nuclei, can be thought of as nonmotile sperm cells. After reaching an ovule and breaking out of the pollen tube tip, one generative nucleus unites with the egg cell to form a diploid zygote (*i.e.,* a fertilized egg with two complete sets of chromosomes, one from each parent), which, through a limited number of divisions gives rise to an embryo. The other generative nucleus fuses with the two polar nuclei to produce a triploid (three sets of chromosomes) nucleus, which divides repeatedly before cell-wall formation occurs, producing the triploid endosperm, a nutrient tissue that contains a variety of storage materials—such as starch, sugars, fats, proteins, hemicelluloses, and phytic acid (a phosphate reserve).

Double fertilization in angiosperms

The events just described constitute what is called the double-fertilization process, one of the characteristic features of all flowering plants. In the orchids and in some other plants with minute seeds that contain no reserve materials, endosperm formation is completely suppressed. In other cases it is greatly reduced, but the reserve materials are present elsewhere—*e.g.,* in the cotyledons, or seed leaves, of the embryo, as in beans, lettuce, and peanuts, or in a tissue derived from the nucellus, the perisperm, as in coffee. Other seeds, such as those of beets, contain both perisperm and endosperm. The seed coat, or testa, is derived from the one or two protective integuments of the ovule. The ovary, in the simplest case, develops into a fruit. In many plants, such as the grasses and lettuce, the outer integument and ovary wall are completely fused, so that seed and fruit form one entity; thus seeds and fruits can logically be described together as "dispersal units," or diaspores. More often, however, the seeds are discrete units attached to the placenta on the inside of the fruit wall through a stalk, or funiculus.

The hilum of a liberated seed is a small scar marking its former place of attachment. The short ridge (raphe) that sometimes leads away from the hilum is formed by the fusion of seed stalk and testa. In many seeds, the micropyle of the ovule also persists as a small opening in the seed coat. The embryo, variously located in the seed, may be very small (as in buttercups) or may fill the seed almost completely (as in roses and plants of the mustard family). It consists of a root part, or radicle, a prospective shoot (plumule or epicotyl), one or more cotyledons (one or two in flowering plants, several in *Pinus*), and a hypocotyl, which is a region that connects radicle and plumule (see Figure 13). A classification of seeds can be based on size

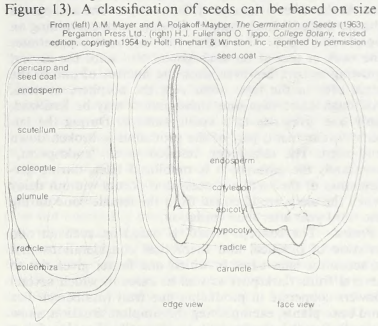


Figure 13: *Longitudinal sections of mature seeds.*
(Left) *Zea mays*, maize or Indian corn, a monocotyledonous caryopsis, or grain, which is actually both seed and fruit through the fusion of the seed coat with the ovary wall. (Centre and right) *Ricinus Communis*, castor bean, a dicotyledonous seed, edge and face views.

and position of the embryo and on the proportion of embryo to storage tissue; the possession of either one or two cotyledons is considered crucial in recognizing two main groups of flowering plants, the Monocotyledones and Dicotyledones.

Seedlings, arising from embryos in the process of germination, are classified as epigeal (cotyledons above ground, usually green and capable of photosynthesis) and hypogeal (cotyledons below ground; see Figure 14). Particularly in the monocots, special absorbing organs may develop that mobilize the reserve materials and withdraw them from the endosperm; *e.g.,* in the grasses, the cotyledon has been modified into an enzyme-secreting scutellum ("shield") between embryo and endosperm.

Gymnosperm seeds. In gymnosperms (plants with "naked seeds"—conifers, cycads, ginkgos) the ovules are not enclosed in an ovary but lie exposed on leaflike structures, the megasporophylls. A long time span separates pollination and fertilization, and the ovules begin to develop into seeds long before fertilization has been accom-

Delayed fertilization in gymnosperms

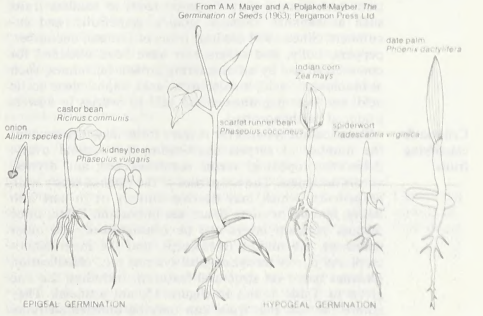


Figure 14: *Seedling morphology and germination modes.*
(Left) Epigeal germination. (Right) Hypogeal germination.

plished; in some cases, in fact, fertilization does not occur until the ovules ("seeds") have been shed from the tree. In the European pine *Pinus sylvestris,* for example, the female cones (essentially collections of megasporophylls) begin to develop in winter and are ready to receive pollen from the male cones in spring. During the first growing season, the pollen tube grows slowly through the nucellus, while within the ovule the megaspore nucleus, through a series of divisions, gives rise to a collection of some 2,000 nuclei, which are then individually enclosed by walls to form a structure called the female gametophyte or prothallus. At the micropylar end of the ovule, several archegonia (bottle-shaped female organs) develop, each containing an oosphere ("egg"). The pollen tube ultimately penetrates the neck of one of the archegonia. Not until the second growing season, however, does the nucleus of one of the male cells in the tube unite with the oosphere nucleus. Although more than one archegonium may be fertilized, only one gives rise to a viable embryo. During the latter's development, part of the prothallus is broken down and used. The remainder, referred to as "endosperm," surrounds the embryo; it is mobilized later, during germination of the seed, a process that occurs without delay when the seeds are liberated from the female cone during the third year after their initiation.

Fruits. The concept "fruit" is based on such an odd mixture of practical and theoretical considerations that it accommodates cases in which one flower gives rise to several fruits (larkspur) as well as cases in which several flowers cooperate in producing one fruit (mulberry). Pea and bean plants, exemplifying the simplest situation, show in each flower a single pistil, traditionally thought of as a megasporophyll or carpel. The carpel is believed to be the evolutionary product of an originally leaflike organ bearing ovules along its margin, but somehow folded along the median line, with a meeting and coalescing of the margins of each half, the result being a miniature, closed but hollow pod with one row of ovules along the suture. In many members of the rose and buttercup families each flower contains a number of similar single-carpelled pistils, separate and distinct, which together represent what is known as an apocarpous gynoecium. In still other cases, two to several carpels (still thought of as megasporophylls, although perhaps not always justifiably) are assumed to have fused to produce a single compound gynoecium (pistil), whose basal part or ovary may be uniloculate (one cavity) or plurilocular (with several compartments), depending on the method of carpel fusion. Most fruits develop from a single pistil. A fruit resulting from the apocarpous gynoecium (several pistils) of a single flower may be referred to as an aggregate fruit; a multiple fruit represents the gynoecia of several flowers. When additional flower parts, such as the stem axis or floral tube, are retained or participate in fruit formation, as in the apple, an accessory fruit results.

Certain plants, mostly cultivated varieties, spontaneously produce fruits in the absence of pollination and fertilization; such natural parthenocarpy leads to seedless fruits such as bananas, oranges, grapes, grapefruits, and cucumbers. Since 1934 seedless fruits of tomato, cucumber, peppers, holly, and others also have been obtained for commercial use by administering growth hormones, such as indoleacetic acid, indolebutyric acid, naphthalene acetic acid, and beta-naphthoxyacetic acid to ovaries in flowers (induced parthenocarpy).

Criteria for classifying fruits

Classification systems for mature fruits take into account the number of carpels constituting the original ovary; dehiscence (opening) versus nondehiscence; and dryness versus fleshiness. The properties of the ripened ovary wall, or pericarp, which may develop entirely or in part into fleshy, fibrous, or stony tissue, are important. Often, three distinct pericarp layers can be distinguished: the outer (exocarp), the middle (mesocarp), and the inner (endocarp). All purely morphological systems (*i.e.,* classification schemes based on structural features), including the one given in Table 1 and in Figure 15, are artificial. They ignore the fact that fruits can only be understood functionally and dynamically.

As strikingly exemplified by the word nut, popular terms often do not properly describe the botanical nature of certain fruits. A Brazil "nut," for example, is a thick-walled seed enclosed in a likewise thick-walled capsule along with several sister seeds. A coconut is a drupe (a stony-seeded fruit; see Table 1) with a fibrous outer part. A walnut is a drupe in which the pericarp has differentiated into a fleshy outer husk and an inner hard "shell"; the "meat" represents the seed—two large, convoluted cotyledons, a minute epicotyl and hypocotyl, and a thin, papery seed coat. A peanut is an indehiscent legume fruit. An almond "nut" is the "stone"—*i.e.,* the hardened endocarp of a drupe usually containing a single seed. Botanically speaking, blackberries and raspberries are not "berries" but aggregates of tiny drupes. A juniper "berry" is comparable to a complete pine cone. A mulberry is a multiple fruit (see Table 1) composed of small nutlets surrounded by fleshy sepals; a strawberry represents a much swollen receptacle (the tip of the flower stalk bearing the flower parts) bearing on its convex surface an aggregation of tiny achenes (small, single-seeded fruits; see Table 1).

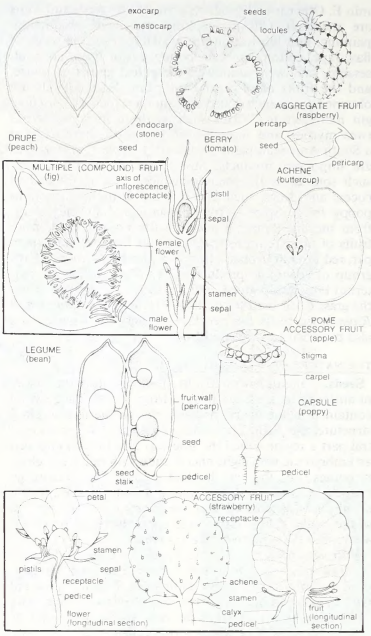


Figure 15: Types of fruit.
From H.J. Fuller and O. Tippo, *College Botany,* revised edition, copyright 1954 by Holt, Rinehart & Winston, Inc.; reprinted by permission

FORM AND FUNCTION

Seed size. In the Late Carboniferous Period (about 280,000,000 to 325,000,000 years ago) some seed ferns produced large seeds (12×6 centimetres [5×2 inches] in *Pachytesta incrassata*). This primitive, ancestral condition of large seeds is reflected in certain gymnosperms (*Cycas circinalis,* 5.5×4 centimetres; *Araucaria bidwillii,* 4.5×3.5 centimetres) and also in some tropical rain-forest trees with nondormant, water-rich seeds (*Mora excelsa,* 12×7 centimetres). The "double coconut" palm *Lodoicea maldivica* represents the extreme, with seeds weighing up to 27 kilograms (about 60 pounds). Herbaceous, nontropical flowering plants usually have seeds weighing in the range of about 0.0001 to 0.01 grams. Within a given

Table 1: Classification of Fruits

major types	structure	
	one carpel	two or more carpels
Dry dehiscent	**Follicle**—at maturity, the carpel splits down one side, usually the ventral suture; milkweed, columbine, peony, larkspur, marsh marigold **Legume**—dehisces along both dorsal and ventral sutures, forming two valves; most members of the pea family	**Capsule**—from compound ovary, seeds shed in various ways—*e.g.*, through holes (*Papaver*—poppies) or longitudinal slits (California poppy) or by means of a lid (pimpernel); flower axis participates in *Iris*; snapdragons, violets, lilies, and many plant families **Silique**—from bicarpellate, compound, superior ovary; pericarp separates as two halves, leaving persistent central septum with seed or seeds attached; dollar plant, mustard, cabbage, rock cress, wall flower **Silicle**—a short silique; shepherd's purse, pepper grass
Dry indehiscent	**Peanut fruit**—(nontypical legume) **Lomentum**—a legume fragmentizing transversely into single-seeded "mericarps"; sensitive plant (*Mimosa*)	**Nut**—like the achene (see below); derived from 2 or more carpels, pericarp hard or stony; hazelnut, acorn, chestnut, basswood **Schizocarp**—collectively, the product of a compound ovary fragmentizing at maturity into a number of one-seeded "mericarps"; maple, mallows, members of the mint family (Lamiaceae or Labiatae), geraniums, carrots, dills, fennels
	Achene—small, single-seeded fruit, pericarp relatively thin; seed free in cavity except for its funicular attachment; buttercup, anemones, buckwheat, crowfoot, water plantain	
	Cypsela—achene-like, but from inferior, compound ovary; members of the aster family (Asteraceae or Compositae), sunflowers	
	Samara—a winged achene; elm, ash, tree-of-heaven, wafer ash	
	Caryopsis—achene-like; from compound ovary; seed coat fused with pericarp; grass family (Poaceae or Graminae)	
Fleshy (pericarp partly or wholly fleshy or fibrous)	**Drupe**—mesocarp fleshy, endocarp hard and stony; usually single-seeded; plum, peach, almond, cherry, olive, coconut	
	Berry—both mesocarp and endocarp fleshy; one-seeded: nutmeg, date; one carpel, several seeds: baneberry, may apple, barberry, Oregon grape; more carpels, several seeds: grape, tomato, potato, asparagus	
	Pepo—berry with hard rind; squash, cucumber, pumpkin, watermelon	
	Hesperidium—berry with leathery rind; orange, grapefruit, lemon	
	two or more carpels of the same flower plus stem axis or floral tube	carpels from several flowers plus stem axis or floral tube plus accessory parts
Fleshy (pericarp partly or wholly fleshy or fibrous)	**Pome**—accessory fruit from compound, inferior ovary; only central part of fruit represents pericarp, with fleshy exocarp and mesocarp and cartilaginous or stony endocarp ("core"); apple, pear, quince, hawthorn, mountain ash **Inferior berry**—blueberry **Aggregate fleshy fruits**—strawberry (achenes borne on fleshy receptacle); blackberry, raspberry (collection of drupelets); magnolia	**Multiple fruits**—fig (a "syconium"), mulberry, osage orange, pineapple, flowering dogwood

family (*e.g.*, the pea family, Fabaceae or Leguminosae) seed size may vary greatly; in others it is consistently large or small, justifying the recognition of "megaspermous" families (*e.g.*, the beech, nutmeg, palm, and soursop families) and "microspermous" ones (*e.g.*, the milkweed, daisy, heather, nettle, and willow families). The smallest known seeds, devoid of food reserves, are found in orchids, saprophytes (nongreen plants that absorb nutrients from dead organic matter—*e.g.*, Indian pipe, *Monotropa*; coral root, *Corallorhiza*), carnivorous plants (sundews, pitcher plants), and total parasites (members of the families Rafflesiaceae and Orobanchaceae, or broomrapes, which latter have seeds weighing about 0.001 milligrams—about 3.5 hundred-millionths of an ounce). Clearly, seed size is related to life-style—total parasites obtain food from their host, even in their early growth stages, and young orchids are saprophytes that receive assistance in absorbing nutrients from fungi that associate closely with their roots. In both cases only very small seeds that lack endosperm are produced. Dodders (*Cuscuta*) and mistletoes (*Viscum, Phoradendron*) live independently when very young and accordingly have relatively large seeds. Many plant species possess seeds of remarkably uniform size, useful as beads (*e.g., Abrus precatorius*) or units of weight—one carat of weight once corresponded with one seed of the carob tree, *Ceratonia siliqua*. In wheat and many other plants, average seed size does not depend on planting density, showing that seed size is under rather strict genetic control. This does not necessarily preclude significant variations among individual seeds; in peas, for example, the seeds occupying the central region of the pod are the largest, probably as the result of competition for nutrients between developing ovules on the placenta. Striking evolutionary changes in seed size, inadvertently created by man, have occurred in the weed *Camelina sativa* subspecies *linicola*, which grows in flax fields. The customary winnowing of flax seeds selects forms of *Camelina* whose seeds are blown over the same distance as flax seeds in the operation, thus staying with their "models." Consequently, *Camelina* seeds in the south of Russia now mimic the relatively thick, heavy seeds of the oil flax that is grown there, whereas in the northwest they resemble the flat, thin seeds of the predominant fibre flax.

The smallest seeds

Seed size and predation. Seeds form the main source of food for many birds, rodents, ants, and beetles. Harvester ants of the genus *Veromessor*, for example, exact a toll of about 15,000,000 seeds per acre per year from the Sonoran Desert of the southwestern United States. In view of the enormous size range of the predators, which include minute weevil and bruchid-beetle larvae that attack the seeds internally, evolutionary "manipulation" of seed size by a plant species cannot in itself be effective in completely avoiding seed attack. With predation inescapable, however, it must be advantageous for a plant species to invest the total reproductive effort in a large number of very small units (seeds) rather than in a few big ones. The mean seed weight of those 13 species of Central American woody legumes vulnerable to bruchid attack is 0.26 gram; for the 23 species invulnerable by virtue of toxic seed constituents it is three grams.

Seed size and germination. Ecologically, seed size is also important in the breaking of dormancy. Being small, a seed can only "sample" that part of the environment immediately adjacent to it, which is not necessarily representative of the generally prevailing conditions. For successful seedling establishment, there is clearly a risk in "venturing out" too soon. The development in seeds of mechanisms acting as "integrating rain gauges" (see below) should be considered in that light.

The shape of dispersal units. Apart from the importance of shape as a factor in determining the mode of dispersal (*e.g.*, wind dispersal of winged seeds, animal dispersal of spiny fruits), shape also counts when the seed or diaspore is seen as a landing device. The flatness of the enormous tropical *Mora* seeds prevents rolling and effectively restricts germination to the spot where they land. In contrast, *Eusideroxylon zwageri* does not grow on steep slopes because its heavy fruits roll downhill. The grains of the grass *Panicum turgidum*, which have a flat and a round side, germinate much better when the flat rather than the convex side lies in contact with wet soil. In very small seeds, the importance of shape can be judged only by taking into account soil clod size and microtopography

Functional aspects of seed shape

of the soils onto which they are dropped. The rounded seeds of cabbage species, for example, tend to roll into crevices, whereas the reticulate ones of lamb's quarters often stay in the positions in which they first fall. Several seeds have appendages (awns, bristles) that promote germination by aiding in orientation and self-burial. In one study, for example, during a six-month period, awned grains of *Danthonia penicillata* gave rise to 12 times as many established seedlings as de-awned ones.

Polymorphism of seeds and fruits. Some plant species produce two or more sharply defined types of seeds that differ in appearance (colour), shape, size, internal structure, or dormancy. In common spurry (*Spergula arvensis*), for example, the seed coat (part of the mother plant) may be either smooth or papillate (covered with tiny nipple-like projections). Here, the phenomenon is genetically controlled by a single factor, so that all the seeds of a given plant are either papillate or smooth. More common is somatic polymorphism, the production by individual plants of different seed types, or "morphs." Somatic polymorphism occurs regularly in *Atriplex* and *Chenopodium*, in which a single plant may produce both large brown seeds capable of immediate germination and small black ones with some innate dormancy. Somatic polymorphism may be controlled by the position of the two (or more) seed types within one inflorescence (flower cluster) or fruit, as in cocklebur, or it may result from environmental effects, as in *Halogeton*, in which imposition of long or short days leads to production of brown and black seeds, respectively. Since the different morphs in seed (and fruit) polymorphism usually have different dispersal mechanisms and dormancies, so that germination is spread out both in space and in time, the phenomenon can be seen as an insurance against catastrophe. The most spectacular example of heterocarpy (*i.e.*, production of differing fruit) is found in the Mediterranean *Fedia cornucopiae* (family Valerianaceae), which has three astonishingly different kinds of fruits that show adaptations to dispersal by wind and water, ants, and larger animals, respectively.

AGENTS OF DISPERSAL

The dispersing agents for seeds and fruits are indicated in such terms as anemochory, hydrochory, and zoochory, which mean dispersal by wind, water, and animals, respectively. Within the zoochorous group further differentiation according to the carriers can be made: saurochory, dispersal by reptiles; ornithochory, by birds; myrmecochory, by ants. Or the manner in which the diaspores are carried can be emphasized, distinguishing endozoochory, diaspores carried within the animal; epizoochory, diaspores accidentally carried on the outside; and synzoochory, diaspores intentionally carried, mostly in the mouth as in birds and ants. See Figure 16 for examples of fruits and seeds that are adapted for dispersal by various means.

Categories of seed dispersal

Dispersal by animals. Snails disperse the small seeds of a very few plant species (*e.g.*, *Adoxa*). Earthworms are more important as seed dispersers. Many intact fruits and seeds can serve as fish bait, those of *Sonneratia*, for example, for the catfish *Arius maculatus*. Certain Amazon River fishes react positively to the audible "explosions" of the ripe fruits of *Eperua rubiginosa*. Fossil evidence indicates that saurochory is very ancient. The giant Galapagos

Drawing by M. Moran based on (A, B, D, beggar's ticks, grapple plant, thistle, unicorn plant) *An Evolutionary Survey of the Plant Kingdom*, by R.F. Scagel, G.E. Rouse, J.R. Stein, R.J. Bandoni, W.B. Schofield, T.M.C. Taylor, © 1965 by Wadsworth Publishing Company, Inc., Belmont, California 94002. Reprinted by permission of the publisher. (C) S.K. von Mariaun, *The Natural History of Plants*, Holt, Rinehart and Winston, Inc.

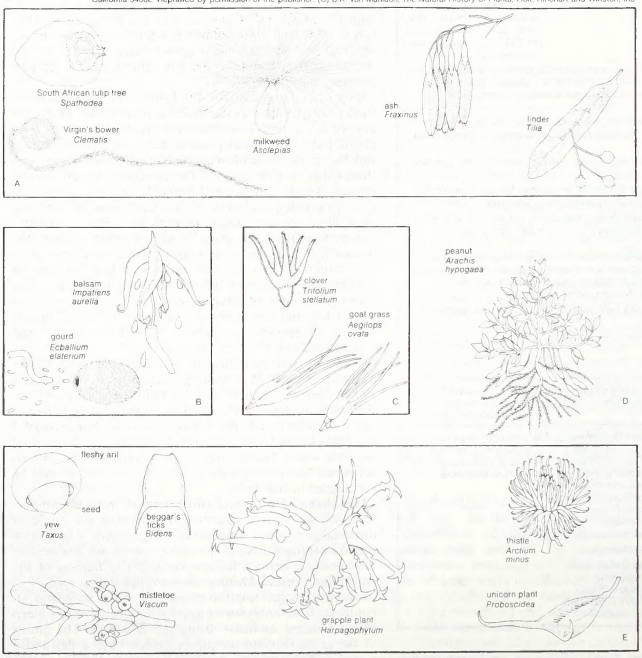


Figure 16: *Seed and fruit dispersal.*
(A) Wind-dispersed fruits and seeds. (B) Mechanically dispersed seeds. (C) Creeping and hopping fruits. (D) Limitation of seed dispersal, geocarpy, underground development of fruits resulting in confinement to areas near parent plant. (E) Animal-dispersed fruits and seeds.

tortoise is important for the dispersal of local cacti and tomatoes. The name alligator apple for *Annona palustris* refers to its method of dispersal, an example of saurochory. Many mammals, ranging in size from mice and kangaroo rats to elephants, eat and disperse seeds and fruits. In the tropics, chiropterochory (dispersal by large bats such as flying foxes, *Pteropus*) is particularly important. Fruits adapted to these animals are relatively large and drab in colour, with large seeds and a striking (often rank) odour; they are accessible to bats because of the pagoda-like structure of the tree canopy, fruit placement on the main trunk, or suspension from long stalks that hang free of the foliage. Examples include mangoes, guavas, breadfruit, carob, and several fig species. In South Africa, a desert melon (*Cucumis humifructus*) participates in a symbiotic relationship with aardvarks—the animals eat the fruit for its water content and bury their own dung, which contains the seeds, near their burrows. Furry terrestrial mammals are the agents most frequently involved in epizoochory, the inadvertent carrying by animals of dispersal units. Burrlike seeds and fruits, or those diaspores provided with spines, hooks, claws, bristles, barbs, grapples, and prickles, are genuine hitchhikers, clinging tenaciously to their carriers. Their functional shape is achieved in various ways—in cleavers, or goose grass (*Galium aparine*), and enchanter's nightshade (*Circaea lutetiana*) the hooks are part of the fruit itself; in common agrimony (*Agrimonia eupatoria*) the fruit is covered by a persistent calyx (the sepals, parts of the flower, which remain attached beyond the usual period) equipped with hooks; in wood avens (*Geum urbanum*) the persistent styles have hooked tips. Other examples are bur marigolds, or beggar's-ticks (*Bidens* species); buffalo burr (*Solanum rostratum*); burdock (*Arctium*); *Acaena;* and many *Medicago* species. The last-named, with dispersal units highly resistant to damage from hot water and certain chemicals (dyes), have achieved wide global distribution through the wool trade. A somewhat different principle is employed by the so-called trample burrs, said to lodge themselves between the hooves of large grazing mammals. Examples are mule grab (*Proboscidea*) and the African grapple plant (*Harpagophytum*). In water burrs, such as those of the water nut *Trapa,* the spines should probably be considered as anchoring devices.

Examples of seed hitchhikers

Dispersal by birds. Birds, being preening animals, rarely carry burrlike diaspores on their bodies. They do, however, transport the very sticky (viscid) fruits of *Pisonia,* a tropical tree of the four-o'clock family, to distant Pacific islands in this way. Small diaspores, such as those of sedges and certain grasses, may also be carried in the mud sticking to waterfowl and terrestrial birds.

Synzoochory, deliberate carrying of diaspores by animals, is practiced when birds carry diaspores in their beaks. The European mistle thrush, *Turdus viscivorus,* deposits the viscid seeds of mistletoe (*Viscum album*) on potential host plants when, after a meal of the berries, it whets its bill on branches or simply regurgitates the seeds. The North American mistletoes (*Phoradendron*) are dispersed by various birds, and the comparable tropical species of the plant family Loranthaceae by flowerpeckers (of the bird family Dicaeidae), which have a highly specialized gizzard that allows seeds to pass through but retains insects. Plants may also profit from the forgetfulness and sloppy habits of certain nut-eating birds that cache part of their food but neglect to recover everything, or drop units on their way to the hiding place. Best known in this respect are the nutcrackers (*Nucifraga*), which feed largely on the "nuts" of beech, oak, walnut, chestnut, and hazel; the jays (*Garrulus*), which hide hazelnuts and acorns; the nuthatches; and the California woodpecker (*Balanosphyra*), which may embed literally thousands of acorns, almonds, and pecan nuts in bark fissures or holes of trees. Secondarily, rodents may aid in dispersal by stealing the embedded diaspores and burying them. In Germany an average jay may transport about 4,600 acorns per season, over distances of up to four kilometres (2.5 miles). Woodpeckers, nutcrackers, and squirrels are responsible for a similar dispersal of *Pinus cembra* in the Alps near the tree line.

Most ornithochores (plants with bird-dispersed seeds) have conspicuous diaspores attractive to such fruit-eating birds as thrushes, pigeons, barbets (members of the bird family Capitonidae), toucans, and hornbills (family Bucerotidae), all of which either excrete or regurgitate the hard part undamaged. Such diaspores have a fleshy, sweet, or oil-containing edible part; a striking colour (often red or orange); no pronounced smell; a protection against being eaten prematurely in the form of acids and tannic compounds that are present only in the green fruit; a protection of the seed against digestion—bitterness, hardness, or the presence of poisonous compounds; permanent attachment; and, finally, absence of a hard outer cover. In contrast to bat-dispersed diaspores, they occupy no special position on the plant. Examples are rose hips, plums, dogwood fruits, barberry, red currant, mulberry, nutmeg fruits, figs, blackberries, and others. The natural and abundant occurrence of *Evonymus* (cardinal's hat), essentially a tropical genus, in temperate Europe and Asia, can be understood only in connection with the activities of birds. Birds also contributed substantially to the repopulation with plants of the island Krakatoa after the catastrophic eruption of 1883. Birds have made *Lantana* (originally American) a pest in Indonesia; the same is true of wild plums (*Prunus serotina*) in parts of Europe and *Rubus* species in Brazil and New Zealand.

Adaptations of seeds and fruits to bird dispersal

Mimicry—the protection-affording imitation of a dangerous or toxic species by an edible, harmless one—is shown in reverse by certain bird-dispersed "coral seeds" such as those of many species in the genera *Abrus, Ormosia, Rhynchosia, Adenanthera,* and *Erythrina.* Hard and often shiny red or black and red, many such seeds deceptively suggest the presence of a fleshy red aril and thus invite the attention of hungry birds.

Dispersal by ants. Mediterranean and North American harvester ants (*Messor, Atta, Tetramorium,* and *Pheidole*) are essentially destructive, storing and fermenting many seeds and eating them completely. Other ants (*Lasius, Myrmica,* and *Formica* species) eat the fleshy, edible appendage (the fat body or elaiosome) of certain specialized seeds, which they disperse (see Figure 17). Most myrmecochorous plants (species of violet, primrose, hepatica, cyclamen, anemone, corydalis, *Trillium,* and bloodroot) belong to the herbaceous spring flora of northern forests. Tree poppy (*Dendromecon*), however, is found in the dry California chaparral; *Melica* and *Centaurea* species in arid

Fritz Schremmer

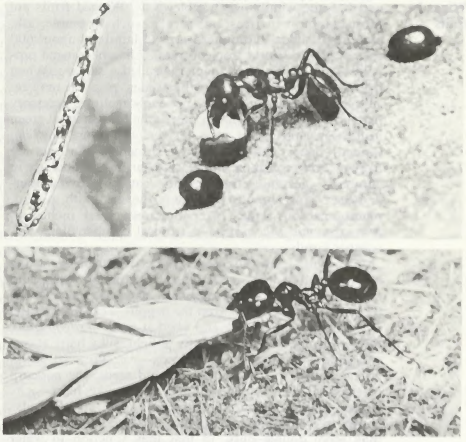


Figure 17: *Seed dispersal by ants.*
(Top left) Fruit of greater celandine (*Chelidonium majus*) opened to show dark seeds with light elaiosomes (fat bodies). (Top right) Ant (*Myrmica laevinodes*) dispersing celandine seeds. (Bottom) Harvester ant (*Messor barbarus*) transporting cereal grains.

Plant and ant interrelationships

Mediterranean regions. The so-called ant epiphytes of the tropics (*i.e.*, species of *Hoya, Dischidia, Aeschynanthus,* and *Myrmecodia*—plants that live in "ant gardens" on trees or offer the ants shelter in their own body cavities) constitute a special group of myrmecochores, providing oil in seed hairs, which in ancestral forms must have served in wind dispersal. The primary ant attractant of myrmecochorous seeds is not necessarily oil; instead, an unsaturated, somewhat volatile fatty acid is suspected in some cases. The myrmecochorous plant as a whole may also have specific adaptations; for example, Cyclamen brings fruits and seeds within reach of ants by conspicuous coiling (shortening) of the flower stalk as soon as flowering is over.

Dispersal by wind. In the modern world, wind dispersal (although numerically important) reflects the climatic and biotic poverty of certain regions; it is essentially a feature of pioneer vegetations. The flora of the Alps is 60 percent anemochorous, that of the Mediterranean garrigue 50 percent. By making certain assumptions (*e.g.*, for average wind velocity and turbulence) the "average limits of dispersal"—that is, the distance that 1 percent of the diaspores can reach—can be calculated for dispersal units of various construction and weight. This calculation yields values of 10 kilometres (six miles) for dandelion (*Taraxacum officinale*) and 0.5 kilometre (0.3 mile) for European pine (*Pinus sylvestris*). Storms result in higher values—30 kilometres (20 miles) for poplar and 200 kilometres (125 miles) for *Senecio congestus*.

Too much success in dispersal may be ecologically futile, as exemplified by certain Florida orchids that arise from windblown West Indian seeds but do not multiply because of the lack of specific pollinators; usually certain bees or wasps. Anemochorous diaspores can be subdivided into flyers, dust diaspores, balloons, and plumed or winged diaspores; rollers, chamaechores or tumbleweeds; and throwers, ballistic anemochores. Dispersal by means of minute dust diaspores produced in huge quantities is comparable to spore dispersal in lower plants—a "saturation bombing" is required to find the very limited number of targets, or favourable growth habitats, that exist. Not surprisingly, it is practiced mostly by total parasites, such as broomrapes (in which the finding of the specific host is a problem), and saprophytes. The inflated, indehiscent pods of *Colutea arborea*, a steppe plant, represent balloons capable of limited air travel before they hit the ground and become windblown tumbleweeds. Winged fruits are most common in trees and shrubs, such as maple, ash, elm, birch, alder, and dipterocarps (a family of about 600 species of Old World tropical trees). The one-winged propeller type, as found in maple, is called a samara. When fruits have several wings on their sides, rotation may result, as in rhubarb and dock species. Sometimes accessory parts form the wings—for example, the bracts (small green leaflike structures that grow just below flowers) in *Tilia* (linden). Seeds with a thin wing formed by the testa are likewise most common in trees and shrubs, particularly in climbers—jacaranda, trumpet vine, catalpa, yams, butter-and-eggs. Most famous of these is the seed with a giant, membranaceous wing (15 centimetres, or six inches) of *Macrozanonia macrocarpa*, a tropical climber of the cucumber family.

Winged fruits

Many fruits form plumes, some derived from persisting and ultimately hairy styles, as in clematis, avens, and anemones; some from the perianth, as in the sedge family (Cyperaceae); and some from the pappus, a calyx structure, as in dandelion and Jack-go-to-bed-at-noon (*Tragopogon*). Plumed seeds usually have tufts of light, silky hairs at one end (rarely both ends) of the seeds—*e.g.*, fireweed, milkweeds, dogbane. In woolly fruits and seeds, the pericarp or the seed coat is covered with cotton-like hairs—*e.g.*, willow, poplar or cottonwood, kapok, cotton, balsa, silk-cotton tree, and some anemones. In some cases, the hairs may serve double duty, in that they function in water dispersal as well as wind dispersal. In tumbleweeds, the whole plant or its fruiting portion breaks off and is blown across open country, scattering seeds as it goes; examples include Russian thistle, pigweed, tumbling mustard, perhaps rose of Jericho, and "windballs" of the grass *Spinifex* of Indonesian shores and Australian steppes. Poppies have a mechanism in which the wind has to swing the slender fruitstalk back and forth before the seeds are thrown out through pores near the top of the capsule.

Dispersal by water. Many beach, pond, and swamp plants have waterborne seeds, which are buoyant by being enclosed in corky fruits or air-containing fruits or both; examples of these plants include water plantain, yellow flag, sea kale, sea rocket, sea beet, and all species of Rhizophoraceae, a family of mangrove plants. Sea dispersal of the coconut palm has been well proved; the fibrous mesocarp of the fruit, a giant drupe, provides buoyancy. Once the nuts are ashore, the mesocarp also aids in the aboveground germination process by collecting rainwater; in addition, the endosperm has in its "milk" a provision for seedling establishment on beaches without much freshwater. A sea rocket species with seeds highly resistant to seawater is gaining a foothold on volcanic Surtsey Island south of Iceland. Purple loosestrife, monkey flower, *Aster tripolium*, and *Juncus* species (rushes) are often transported by water in the seedling stage. Rainwash down mountain slopes may be important in tropical forests. A "splashcup mechanism," common in fungi for spore dispersal, is suggested by the open fruit capsule with exposed small seeds in the pearlwort (*Sagina*) and mitrewort (*Mitella*). Hygrochasy, the opening of fruits in moist weather, is displayed by species of *Mesembryanthemum, Sedum*, and other plants of dry environments.

Self-dispersal. Best known in this category are the active ballists, which forcibly eject their seeds by means of various mechanisms. In the fruit of the dwarf mistletoe (*Arceuthobium*) of the western United States, a very high osmotic pressure (pressure accumulated by movement of water across cell membranes principally in only one direction) builds up that ultimately leads to a lateral blasting out of the seeds over distances of up to 15 metres (49 feet) with an initial velocity of about 95 kilometres (60 miles) per hour. Squirting cucumber (*Ecballium elaterium*) also employs an osmotic mechanism. In Scotch broom and gorse, however, drying out of the already dead tissues in the two valves of the seed pod causes a tendency to warp, which, on hot summer days, culminates in an explosive and audible separation of these valves, with violent seed release. Such methods may be coupled with secondary dispersal mechanisms, effected by ants in the case of Scotch broom and gorse or by birds and mammals, to which sticky seeds may adhere, in the case of *Arceuthobium* and squirting cucumber. Other active ballists are species of geranium, violet, wood sorrel, witch-hazel, touch-me-not (*Impatiens*), and acanthus; probable champions are *Bauhinia purpurea*, with a distance of 15 metres and the sandbox tree (*Hura crepitans*) with 14 metres. Barochory, the dispersal of seeds and fruits by gravity alone, is demonstrated by the heavy fruits of horse chestnut.

Explosive fruits

Creeping diaspores are found in grasses such as *Avena sterilis* and *Aegilops ovata*, the grains of which are provided with bristles capable of hygroscopic movements (coiling and flexing, in response to changes in moisture). The mericarps (a fruit fragment—see the Table: *Schizocarp*) or stork's bill (an *Erodium* species), when moistened, bury themselves with a corkscrew motion by unwinding a multiple-barbed, beak-shaped appendage, which, in the dry state, was coiled.

Atelechory, the dispersal over a very limited distance only, represents a waste-avoiding, defensive "strategy" that functions in further exploitation of an already occupied, favourable site. This aim is often achieved by synaptospermy, the sticking together of several diaspores, which makes them less mobile, as in beet and spinach; also, by geocarpy, which is either the production of fruits underground, as in the arum lilies *Stylochiton* and *Biarum*, in which the flowers are already subterranean, or the active burying of fruits by the mother plant, as in the peanut, *Arachis hypogaea*. In the American hog peanut (*Amphicarpa bracteata*), pods of a special type are buried by the plant and are cached by squirrels later on. Kenilworth ivy (*Cymbalaria*), which normally grows on stone or brick walls, stashes its fruits away in crevices after strikingly extending the flower stalks. Not surprisingly, geocarpy, like

synaptospermy, is most often encountered in desert plants; however, it also occurs in violet species, in subterranean clover (*Trifolium subterraneum*)—even when it grows in France and England—and in *Begonia hypogaea* of the African rain forest.

GERMINATION

Dormancy and life-span of seeds. Diaspore dormancy has at least three functions: (1) immediate germination must be prevented even when circumstances are optimal so as to avoid exposure of the seedling to an unfavourable period (*e.g.*, winter), which is sure to follow; (2) the unfavourable period has to be survived; and (3) the various dispersing agents must be given time to act. Accordingly, the wide variation in diaspore longevity can be appreciated only by linking it with the various dispersal mechanisms employed, as well as with the climate and its seasonal changes. Thus, the downy seeds of willows, blown up and down rivers in early summer with a chance of quick establishment on newly exposed sandbars, have a life-span of only one week. Tropical rain forest trees frequently have seeds of low life expectancy also. Intermediate are seeds of sugarcane, tea, and coco palm, among others, with life-spans of up to a year. *Mimosa glomerata* seeds in the herbarium of the Muséum National d'Histoire Naturelle in Paris were found viable after 221 years. In general, viability is better retained in air of low moisture content. Some seeds, however, remain viable under water—those of certain rush (*Juncus*) species and *Sium cicutaefolium* for at least seven years. Salt water can be tolerated for years by the pebble-like but floating seeds of *Caesalpinia* (*Guilandina*) *bonduc* and *C. bonducella*, species that, in consequence, possess an almost pantropical distribution. Seeds of the sacred lotus (*Nelumbo nucifera*) found in a peat deposit in Manchuria and estimated by radioactive-carbon dating to be 1,400 (±400) years old, rapidly germinated (and subsequently produced flowering plants) when the seeds were filed to permit water entry. In 1967 seeds of the arctic tundra lupine (*Lupinus arcticus*), found in a frozen lemming burrow with animal remains established to be at least 10,000 years old, germinated within 48 hours when returned to favourable conditions. The problem of differential seed viability has been approached experimentally by various workers, one of whom buried 20 species of common Michigan weed seeds, mixed with sand, in inverted open-mouthed bottles for periodic inspection. After 80 years, three species still had viable seeds.

Lack of dormancy. In some plants, the seeds are able to germinate as soon as they have matured on the plant, as demonstrated by wheat, sweet corn, peas, and beans in a very rainy season. Certain mangrove species normally form foot-long embryos on the trees; these later drop down into the mud or seawater. Such cases, however, are exceptional. The lack of dormancy in cultivated species, contrasting with the situation in most wild plants, is undoubtedly the result of conscious selection by man.

Immature embryos. In plants whose seeds ripen and are shed from the mother plant before the embryo has undergone much development beyond the fertilized egg stage (orchids, broomrapes, ginkgo, dogtooth violet, ash, winter aconite, and buttercups), there is an understandable delay of several weeks or months, even under optimal conditions, before the seedling emerges.

Role of the seed coat. There are at least three ways in which a hard testa may be responsible for seed dormancy: it may (1) prevent expansion of the embryo mechanically (pigweed); or (2) block the entrance of water; or (3) impede gas exchange so that the embryos lack oxygen. Resistance of the testa to water uptake is most widespread in the bean family, the seed coats of which, usually hard, smooth, or even glassy, may, in addition, possess a waxy covering. In some cases water entry is controlled by a small opening, the strophiolar cleft, which is provided with a corklike plug; only removal or loosening of the plug will permit water entry. Similar seeds not possessing a strophiolar cleft must depend on abrasion, which in nature may be brought about by microbial attack, passage through an animal, freezing and thawing, or mechanical means. In horticulture and agriculture, the coats of such seeds are deliberately damaged or weakened by man (scarification). In chemical scarification, seeds are dipped into strong sulfuric acid, organic solvents such as acetone or alcohol, or even boiling water. In mechanical scarification, they may be shaken with some abrasive material such as sand or be scratched with a knife.

Frequently seed coats are permeable to water yet block entrance of oxygen; this applies, for example, to the upper of the two seeds normally found in each burr of the cocklebur plant. The lower seeds germinate readily under a favourable moisture and temperature regime, but the upper ones fail to do so unless the seed coat is punctured or removed or the intact seed is placed under very high oxygen concentrations.

Afterripening, stratification, and temperature effects. The most difficult cases of dormancy to overcome are those in which the embryos, although not underdeveloped, remain dormant even when the seed coats are removed and conditions are favourable for growth. Germination in these takes place only after a series of little-understood changes, usually called afterripening, have taken place in the embryo. In this group are many forest trees and shrubs such as pines, hemlocks, and other conifers; some flowering woody plants such as dogwood, hawthorn, ash, linden, tulip poplar, holly, and viburnum; fruit trees such as apples, pears, peaches, plums, and cherries; and flowering herbaceous plants such as iris, Solomon's seal, and lily-of-the-valley. In some species, one winter suffices for afterripening. In others, the process is drawn out over several years, with some germination occurring each year. This can be viewed as an insurance of the species against flash catastrophes that might completely wipe out certain year classes.

Significance of afterripening requirement

Many species require moisture and low temperatures: for example, in apples, when the cold requirement is insufficiently met, abnormal seedlings result. Others (cereals, dogwood) afterripen during dry storage. The seeds of certain legumes—for example, the seeds of the tree lupin, the coats of which are extremely hard and impermeable—possess a hilum with an ingenious valve mechanism that allows water loss in dry air but prevents re-uptake of moisture in humid air. Of great practical importance is stratification, a procedure aimed at promoting a more uniform and faster germination of cold-requiring, afterripening seeds. In this procedure, seeds are placed for one to six months, depending on the species, between layers of sand, sawdust, sphagnum, or peat and kept moist as well as reasonably cold (usually 0° to 10° C [32° to 50° F]). A remarkable "double dormancy" has thus been uncovered in lily-of-the-valley and false Solomon's seal. Here, two successive cold treatments separated by a warm period are needed for complete seedling development. The first cold treatment eliminates the dormancy of the root; the warm period permits its outgrowth; and the second cold period eliminates epicotyl or leaf dormancy. Thus, almost two years may be required to obtain the complete plant. The optimal temperature for germination, ranging from 1° C (34° F) for bitterroot to 42° C (108° F) for pigweed, may also shift slightly as a result of stratification.

Many dry seeds are remarkably resistant to extreme temperatures, some even to that of liquid air (−140° C or −220° F). Seeds of Scotch broom and some *Medicago* species can be boiled briefly without losing viability. Ecologically, such heat resistance is important in vegetation types periodically ravaged by fire, such as in the California chaparral, where the germination of *Ceanothus* seeds may even be stimulated. Also important ecologically is a germination requirement calling for a modest daily alternation between a higher and a lower temperature. Especially in the desert, extreme temperature fluctuations are an unavoidable feature of the surface, whereas with increasing depth these fluctuations are gradually damped out. A requirement for a modest fluctuation—*e.g.*, from 20° C (68° F) at night to 30° C (86° F) in the daytime (as displayed by the grass *Oryzopsis miliacea*)—practically ensures germination at fair depths; and this is advantageous because a seed germinating in soil has to strike a balance between two conflicting demands, both depending on depth—on the one hand, germination in deeper layers

is advantageous because a dependable moisture supply simply is not available near the surface; but, on the other hand, closeness to the surface is desirable because it allows the seedling to reach air and light rapidly and become self-supporting.

Plants stimulated by light to germinate

Light and seed germination. Many seeds are insensitive to light, but in a number of species germination is stimulated or inhibited by exposure to continuous or short periods of illumination. So stimulated are many grasses, lettuce, fireweed, peppergrass (*Lepidium*), mullein, evening primrose, yellow dock, loosestrife, and Chinese lantern plant. Corn (maize), the smaller cereals, and many legumes, such as beans and clover, germinate as well in light as in darkness. Inhibition by light is found in chive, garlic, and several other species of the lily family, jimson weed, fennel flower (*Nigella*), *Phacelia, Nemophila,* and pigweed (*Amaranthus*). Sometimes, imbibed (wet) seeds that do not germinate at all in darkness may be fully promoted by only a few seconds or minutes of white light. The best studied case of this type, and one that is a milestone in plant physiology, concerns seeds of the Grand Rapids variety of lettuce, which is stimulated to germination by red light (wavelength about 660 nanometres) but inhibited by "far red" light (wavelength about 730 nanometres). Alternations of the two treatments to almost any extent indicate that the last treatment received is the decisive one in determining whether the seeds will germinate.

Ecological role of light. Laboratory experiments and field observations indicate that light is a main controller of seed dormancy in a wide array of species. The absence of light, for example, was found in one study to be responsible for the nongermination of seeds of 20 out of 23 weed species commonly found in arable soil. In regions of shifting sands, seeds of Russian thistle germinate only when the fruits are uncovered, often after a burial period of several years. Conversely, the seeds of *Calligonum comosum* and the melon *Citrullus colocynthis,* inhabiting coarse sandy soils in the Negev Desert, are strongly inhibited by light. The survival value of this response, which restricts germination to buried seeds, lies in the fact that at the surface fluctuating environmental conditions may rapidly create a very hostile micro-environment. The seeds of *Artemisia monosperma* have an absolute light requirement but respond to extremely low intensities, such as is transmitted by a two-millimetre- (0.08-inch-) thick sand filter. In seeds buried too deeply, germination is prevented. The responsiveness to light, however, increases with the duration of water imbibition. Even when full responsiveness to light has been reached, maximal germination occurs only after several light-exposures are given at intervals. In the field, this combined response mechanism acts as an integrating (cumulative) rain gauge because the seeds (as indicated) become increasingly responsive to light, and thus increasingly germinable, the longer the sand remains moistened. Certain *Juncus* seeds have an absolute light requirement over a wide range of temperatures; consequently, they do not germinate under dense vegetation or in overly deep water. In combination with temperature, light (in the sense of day length) may also restrict germination to the most suitable time of year. In birch, for example, seeds that have not gone through a cold period after imbibing water remain dormant after release from the mother plant in the fall and will germinate only when the days begin to lengthen the next spring.

Significance of seed light requirements

Stimulators and inhibitors of germination. A number of chemicals (potassium nitrate, thiourea, and ethylene chlorhydrin) and plant hormones (gibberellins and kinetin) have been used experimentally to break seed dormancy. Their mode of action is obscure, but it is known that in some instances thiourea, gibberellin, and kinetin can substitute for light.

Natural inhibitors, which completely suppress germination (coumarin, parasorbic acid, ferulic acid, phenols, protoanemonin, transcinnamic acid, alkaloids, essential oils, and the hormone dormin) may be present in the pulp or juice of fruits or in various parts of the seed. The effect of seed coat phenols, for example, may be indirect—being highly oxidizable, they may screen out much-needed oxygen. Ecologically, such inhibitors are important in at least three ways. Their slow disappearance with time may spread germination out over several years (a protection against catastrophes). Furthermore, when leached out by rainwater, they often serve as agents inhibiting the germination of other competitive plants nearby. Finally, the gradual leaching out of water-soluble inhibitors serves as an excellent integrating rain gauge. Indeed, it has been shown that the germination of certain desert plants is not related to moisture as such but to soil water movement—*i.e.,* to the amount and duration of rain received.

(B.J.D.M.)

ANIMAL REPRODUCTION

The role of reproduction is to provide for the continued existence of a species; it is the process by which living organisms duplicate themselves. Animals compete with other individuals in the environment to maintain themselves for a period of time sufficient to enable them to produce tissue nonessential to their own survival, but indispensable to the maintenance of the species. The additional tissue, reproductive tissue, usually becomes separated from the individual to form a new, independent organism.

This section describes the reproductive systems in metazoans (multicelled animals) from sponges to mammals, exclusive of man. It focusses on the gonads (sex organs), associated ducts and glands, and adaptations that aid in the union of gametes—*i.e.,* reproductive cells, male or female, that are capable of producing a new individual by union with a gamete of the opposite sex. Brief mention is made of how the organism provides for the development of embryos and of the regulatory role of gonads in vertebrate cycles. (Human reproduction is treated separately below.)

Unlike most other organ systems, the reproductive systems of higher animals have not generally become more complex than those of lower forms. Asexual reproduction (*i.e.,* reproduction not involving the union of gametes), however, occurs only in the invertebrates, in which it is common, occurring in animals as highly evolved as the sea squirts, which are closely related to the vertebrates. Temporary gonads are common among lower animals; in higher animals, however, gonads are permanent organs. Hermaphroditism, in which one individual contains functional reproductive organs of both sexes, is common among lower invertebrates; yet separate sexes occur in such primitive animals as sponges, and hermaphroditism occurs in animals more highly evolved—*e.g.,* the lower fishes. Gonads located on or near the animal surface are common in the lowest invertebrates, but in higher animals they tend to be more deeply situated and often involve intricate duct systems. In echinoderms, which are among the highest invertebrates, the gonads hang directly into the sea and spill their gametes into the water. In protochordates, gametes are released into a stream of respiratory water that passes directly into the sea. Duct systems of the invertebrate flatworms (Platyhelminthes) are relatively complex, and those of specialized arthropods (*e.g.,* insects, spiders, crabs) are more complex than those of any vertebrate. Copulatory organs occur in flatworms, but copulatory organs are not ubiquitous among vertebrates other than reptiles and mammals. The trend toward fewer eggs and increased parental care in higher animals may account for the relative lack of complexity in the reproductive systems of some advanced forms. Whereas trends toward increasing structural complexity have often been reversed during evolution, reproductive behaviour patterns in many phylogenetic (*i.e.,* evolutionary) lines have become more complicated in order to enhance the opportunity for fertilization of eggs and maximum survival of offspring (see SEX AND SEXUALITY).

A direct relationship exists between behaviour and the functional state of gonads. Reproductive behaviour in-

duced principally but not exclusively by organic substances called hormones promotes the union of sperm (spermatozoa) and eggs, as well as any parental care accorded the young. There are a number of reasons why behaviour must be synchronized with gonadal activity. Chief among these are the following:

Synchronization of behaviour and gonadal activity

Individuals of a species must congregate at the time the gonads contain mature gametes. This often entails migration, and some members of all major vertebrate groups migrate long distances to gather at spawning grounds or rookeries.

Individuals with gametes ready to be shed must recognize members of the opposite sex. Recognition is sometimes by external appearance or by chemical substances (pheromones), but sex-linked behaviour is often the only signal.

Geographical territories frequently must be established and aggressively defended.

The building of nests, however simple, is essential reproductive behaviour in many species.

When fertilization of aquatic forms is external, sperm and eggs must be discharged at approximately the same time into the water, since gametes may be quickly dispersed by currents. Courtship, often involving highly intricate behaviour patterns, serves to release the gametes of both mating individuals simultaneously.

When fertilization is internal, willingness of the female to mate is often essential. Female mammals not in a state of willingness to mate not only will not mate but may injure or even kill an aggressive male. The unwillingness of a female mammal to mate when mature eggs are not present prevents loss of sperm needed to preserve the species.

Parental care of fertilized eggs by one parent or the other has evolved in many species. Parental behaviour includes fanning the water or air around the eggs, thereby maintaining appropriate temperature and oxygen levels; secretion of oxygen from a parent's gills; transport of eggs on or in the parental body (including the mouth of some male parents); and brooding, or incubation, of eggs.

Some species extend parental care into the postnatal period, feeding and protecting the offspring. Such behaviour patterns are adaptations for survival and thus are essential; all are induced by the nervous and endocrine systems and are typically cyclical, because gonadal activity is cyclical (see also BEHAVIOUR, ANIMAL: *Reproductive behaviour.*)

Reproductive systems of invertebrates

GONADS, ASSOCIATED STRUCTURES, AND PRODUCTS

Although asexual reproduction occurs in many invertebrate species, most reproduce sexually. The basic unit of sexual reproduction is a gamete (sperm or egg), produced by specialized tissues or organs called gonads. Sexual reproduction does not necessarily imply copulation or even a union of gametes. As might be expected of such a large and diverse group as the invertebrates, many variations have evolved to ensure survival of species. In many lower invertebrates, gonads are temporary organs; in higher forms, however, they are permanent. Some invertebrates have coexistent female and male gonads; in others the same gonad produces both sperm and eggs. Animals in which both sperm and eggs are produced by the same individual (hermaphroditism) are termed monoecious. In dioecious species, the sexes are separate. Generally, the male gonads ripen first in hermaphroditic animals (protandry); this tends to ensure cross-fertilization. Self-fertilization is normal, however, in many species, and some species undergo sex reversal.

Sponges, coelenterates, flatworms, and aschelminths. Sponges are at a cellular level of organization and thus do not have organs or even well-developed tissues; nevertheless, they produce sperm and eggs and also reproduce asexually. Some species of sponge are monoecious, others are dioecious. Sperm and eggs are formed by aggregations of cells called amoebocytes in the body wall; these are not considered gonads because of their origin and transitory nature.

In hydrozoan coelenterates, temporary gonads are formed by groups of cells in either the epidermis (outer cell layer) or gastrodermis (gut lining), depending on the species; scyphozoan and anthozoan coelenterates generally have gonads in the gastrodermis. The origin and development of gonads in coelenterates, particularly freshwater species, are often associated with the seasons. Freshwater hydrozoans, for example, reproduce asexually until the onset of cold weather, which stimulates them to form testes and ovaries. Colonial hydrozoans asexually produce individuals known as polyps. Polyps, in turn, give rise to free swimming stages (medusae), in which gonads develop (Figure 18). The body organization of sponges and coelenterates is such that most of their cells are in intimate contact with the environment; consequently, gametes are shed into the water, and no ducts are necessary to convey them to the outside.

Polyps

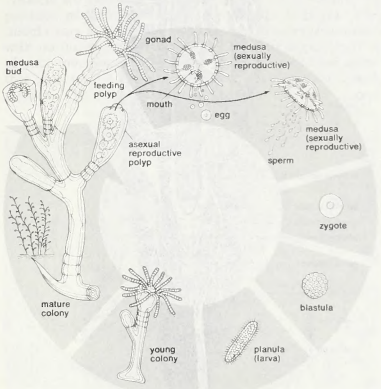


Figure 18: Life cycle of the colonial hydrozoan *Obelia*.

In contrast to sponges and coelenterates, platyhelminths generally have well-developed organ systems of a permanent nature and, in addition, have evolved secondary reproductive structures to convey sex products. One exception is the acoels, a group of primitive turbellarians; they lack permanent gonads, and germinal cells develop from amoebocytes in much the same manner as in sponges. The majority of flatworms, however, are monoecious, the primary sex organs consisting of one or more ovaries and testes (Figure 19). The tube from the ovary to the outside is called the oviduct; it often has an outpocketing (seminal receptacle) for the storage of sperm received during copulation. In many species the oviduct receives a duct from yolk (vitelline) glands, whose cells nourish the fertilized egg. Beyond the entrance of the duct from the yolk glands the oviduct may be modified to secrete a protective capsule around the egg before it is discharged to the outside. The male organs consist of testes, from which extend numerous tubules (vasa efferentia) that unite to form a sperm duct (vas deferens); the latter becomes an ejaculatory duct through which sperm are released to the outside. The sperm duct may exhibit expanded areas that store sperm (seminal vesicles), and it may be surrounded by prostatic cells that contribute to the seminal fluid. The sperm duct eventually passes through a copulatory organ. The same basic structural pattern, somewhat modified, is found in most higher invertebrates.

Aschelminthes (roundworms) are mostly dioecious; frequently there are external differences between males and females (sexual dimorphism). The males are generally smaller and often have copulatory spicules. Nematodes have relatively simple reproductive organs, a tubular testis or ovary being located at the end of a twisted tube. The portion of the female tract nearest the ovary forms a uterus for temporary storage of fertilized eggs. Some species lay eggs, but others retain the egg in the uterus until the larva hatches. The sperm are released into a cav-

Reproductive organs of nematodes

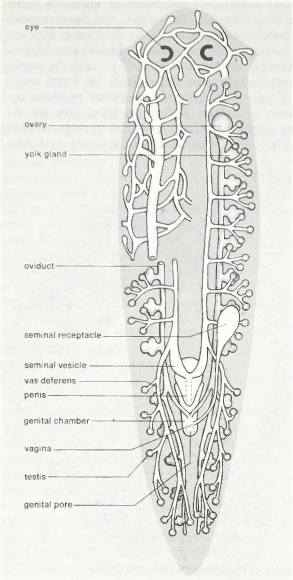


Figure 19: Reproductive system of a planarian flatworm, a monoecious animal.

Reprinted with permission of The Macmillan Company from *College Zoology* by K.A. Stiles, R.W. Hegner, and R.A. Boolootian. © Copyright by The Macmillan Company, 1968

ity called the cloaca. A number of free-living nematodes are capable of sex reversal—if the sex ratio in a given population is not optimal or if environmental conditions are not ideal, the ratio of males to females can be altered. This sometimes results in intersexes; *i.e.*, females with some male characteristics. Hermaphroditism occurs in nematodes, and self-fertilization in such species is common. Unisexual reproduction among rotifers is described below (see *Parthenogenesis*).

Annelids and mollusks. Annelids have a well-developed body cavity (coelom), a part of the lining of which gives rise to gonads. In some annelids, gonads occur in several successive body segments. This is true, for example, in polychaetes, most of which are dioecious. Testes and ovaries usually develop, though not invariably, in many body segments; and the sperm and eggs, often in enormous numbers, are stored in the coelom. Fertilization is external. In oligochaetes (all of which are monoecious) on the other hand, the gonads develop in a few specific segments. Sperm are stored in a seminal vesicle and eggs in an egg sac, rather than in the coelom. A portion of the peritoneum, the membrane lining the coelom, becomes a saclike seminal receptacle that stores sperm received from the mate. The earthworm, *Lumbricus terrestris*, is an example of a specialized annelid reproductive system. Female organs consist of a pair of ovaries in segment 13; a pair of oviducts that open via a ciliated funnel (*i.e.*, with hairlike structures) into segment 13 but open to the exterior in segment 14; an egg sac near each funnel; and a pair of seminal receptacles in segment 9 and another in segment 10. Male organs consist of two pairs of minute testes in segments 10 and 11, each associated with a ciliated sperm funnel leading to a tiny duct, the vas efferens. The two ducts on each side lead to a vas deferens that opens in segment 15. Testes and funnels are contained within two of three pairs of large seminal vesicles that occupy six body segments. Leeches (Hirudinea), also monoecious, have one pair of ovaries and a segmentally arranged series of testes with duct systems basically similar to those of earthworms.

Differences between mollusk and annelid systems

Although mollusks have a close evolutionary kinship to annelids, they have reduced or lost many structures characteristic of segmented worms. The coelom persists only as three regional cavities: gonadal, nephridial (kidney), and pericardial (heart). In ancestral forms these were interconnected so that gametes from the gonad passed through the pericardial cavity, the nephridial cavity, then to the outside through a nephridial pore. The various groups of mollusks have tended to modify this arrangement, with the result that gonads have their own pore; among amphineurans, for example, the sexes are usually separate, and there is one gonad with an associated pore. Gastropods show considerable variability, but generally one gonad (ovary, testis, or ovotestis—a structure combining the functional gonads of both sexes) is located in the visceral hump and connected to the outside by a remnant of the right kidney. In hermaphroditic forms, one duct carries sperm as well as eggs. The gonadal ducts of gastropod females often secrete a protective capsule around the fertilized eggs; in males, the terminal portion of the duct is sometimes contained in a copulatory organ. Pelecypods may be either monoecious or dioecious, but the gonads are usually paired. In mussels and oysters, the gonads open through the nephridial pore, but in clams the reproductive system opens independently. The cephalopods are all dioecious. The single testis or ovary releases its products into the pericardial cavity and this, in turn, leads to a gonopore, the external opening. The oviduct of the squid is terminally modified to form a shell gland. The male system is more complex—the gonoduct leads into a seminal vesicle where a complicated torpedo-shaped sperm case (spermatophore) is secreted and contains the sperm. Spermatophores are then stored in a special structure (Needham's sac) until copulation occurs.

A remarkable characteristic of some mollusks is the ability to alter their sex. Some species are clearly dioecious; however, among the monoecious species there is considerable variability in their hermaphroditic condition. In some species, male and female gonads, although in the same individual, are independent functionally and structurally. In others, an ovotestis produces both sperm and eggs. Oysters display a third condition; young oysters have a tendency toward maleness, but, if water temperature or food availability is altered, some individuals develop into females. Later, a reversal to the male condition may occur. The sexual makeup of an entire oyster population also has a seasonal aspect; in harmony with the group, an individual may undergo several alterations in the course of a year. A similar phenomenon, called consecutive sexuality, occurs in limpets. These gastropods stack themselves in piles, with the younger animals on top. The animals on top are males with well-developed testes and copulatory organs; those in the middle are hermaphroditic; those on the bottom are females, having lost the testes and copulatory organ (penis) by degeneration. A decrease in the number of females in a stack induces males to assume female characteristics, but the transition is retarded when an excess of females is present. The degree of maleness or femaleness is probably controlled in part by environmental and internal factors.

Arthropods. The phylum Arthropoda includes a vast number of organisms of great diversity. Most arthropods are dioecious, but many are hermaphroditic, and some reproduce parthenogenetically (*i.e.*, without fertilization). The primary reproductive organs are much the same as in other higher invertebrates, but the secondary structures are often greatly modified. Such modifications depend on whether fertilization is internal or external, whether the egg or zygote (*i.e.*, the fertilized egg) is retained or immediately released, and whether eggs are provided some means of protection after they have left the body of the female. The mandibulate arthropods (*e.g.*, crustaceans, insects) include more species than any other group and have invaded most habitats, a fact reflected in their reproductive processes.

Crustaceans (*e.g.*, crabs, crayfish, barnacles) are for the most part dioecious. The primary reproductive organs generally consist of paired gonads that open through paired

Reproductive variation among arthropod groups

ventral (bottom side) gonopores. Females often have a seminal receptacle (spermatheca) in the form of an outpocketing of the lower part of the female tract or as an invagination (inpocketing) of the body near the gonopore. Males have appendages modified for clasping the female during copulation or for guiding sperm. A number of groups have members that reproduce parthenogenetically. Branchiopods (*e.g.,* water fleas, fairy shrimp) have simple paired gonads. The female gonopore often opens dorsally (on the back side) into a brood chamber; the male gonopore opens near the anus. Males have appendages for clasping females during copulation. Ostracods, or seed shrimp, have paired, tubular gonads. The eggs may be brooded by the female, or they may be released into the water via a gonoduct and gonopore. The terminal portion of the male gonoduct is enclosed in a single or paired penis. Many species reproduce parthenogenetically. Some experts contend that this is the only method employed, even though functional males may be present in the population. Copepods (*e.g., Cyclops*) have paired ovaries and an unpaired testis. The terminal portion of the oviduct constitutes an ovisac for storage of eggs. The male deposits sperm in a spermatophore that is transferred to the female. Sexual dimorphism is particularly evident among parasitic copepods. Frequently, parasitic females can hardly be recognized as copepods except for the distinctive ovisacs. Males, on the other hand, are free-living and are recognizable as copepods.

The hermaphroditic Cirripedia (*e.g.,* barnacles) are among the exceptions to the generalization that crustaceans are dioecious. It has been suggested that hermaphroditism in barnacles is an adaptation to their sessile, or stationary, existence, but cross-fertilization is more common than self-fertilization. The ovaries lie either in the base or in the stalk of the animal, and the female gonopore is near the base of the first pair of middle appendages (cirri). The testes empty into a seminal vesicle through a series of ducts; from the vesicle extends a long sperm duct within a penis that may be extended to deposit sperm in the mantle cavity of an adjacent barnacle. The terminal portion of the oviduct secretes a substance that forms a kind of ovisac within the mantle cavity, where fertilized eggs undergo early development. Although most barnacles are hermaphrodites, some display a peculiar adaptation in that they contain parasitic dwarf or accessory males. Dwarf males are much smaller than the host barnacle in which they live and are degenerate, except for the testes. In some species they live in the mantle cavity of hermaphroditic forms and produce accessory sperm; in other species only the female organs persist in the host animal, and the accessory male is a necessity.

Amphipods and isopods (*e.g.,* pill bugs, sow bugs), like most crustaceans, are dioecious and have paired gonads. Females of both groups have a ventral brood chamber (marsupium) formed by a series of medially directed (*i.e.,* toward the body midline) plates (oostegites) in the region of the thorax, the region between head and abdomen. Many isopods are parasitic and have developed unusual sex-related activities. Certain species are parasitic on other crustaceans. After a series of molts (*i.e.,* shedding of the body covering) a parasitic larval (immature) isopod attaches to the shell of a crab. If it is the only larva to do so, it increases in size and develops into an adult female. If another larva subsequently attaches, the new arrival becomes a male. It has been demonstrated that the testes of the functioning male larva will change to ovaries if the larva is removed to a new, uninfected host. Thus, the larvae of these species apparently are intersexual and can develop into either sex. This phenomenon, reminiscent of that in mollusks, demonstrates the way in which similar adaptations have evolved in diverse groups of organisms.

The gonads of crabs and lobsters are paired, as are the gonopores. The females of many species have external seminal receptacles on the ventral part of the thorax; those of other species have internal receptacles in the same region. In some species, seminal receptacles are absent, and the male simply attaches a spermatophore to the female. Thus, males either have appendages (gonopods) by which sperm are inserted in the body of the female or produce spermatophores for sperm transfer. The sexual dimorphism of many decapods can be altered by parasitism. An example of this is the crab that is parasitized by a barnacle. A barnacle infection in male crabs induces the secondary sex characters of the crab to resemble those of a female; however, masculinization does not occur in parasitized females. At each molt a parasitized male crab increasingly resembles a female even though the testes may be completely unaffected. Feminization results from a hormonal alteration of the parasitized crab.

Parthenogenesis in insects

Insects are rarely hermaphroditic, but many species reproduce parthenogenetically (without fertilization). The insect ovary is composed of clusters of tubules (ovarioles) with no lumen, or cavity (Figure 20). The upper portion of each ovariole gives rise to oocytes (immature eggs) that mature and are nourished by yolk from the lower portion. The oviduct leads to a genital chamber (copu-

From R. Snodgrass, *Principles of Insect Morphology*. © 1935 by McGraw-Hill Book Co., used with permission of McGraw-Hill Book Co.

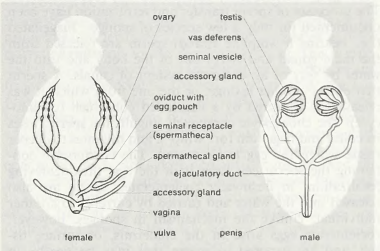


Figure 20: Reproductive organs of insects.

latory bursa, or vagina), with which are often associated accessory glands and a seminal receptacle. Some accessory glands form secretions by which eggs become attached to a hard surface; others secrete a protective envelope around the egg. The eighth and ninth body segments are often modified for egg-laying. The paired testes consist of a series of seminal tubules that form primary spermatogonia (immature spermatozoa) at their upper ends. As the spermatogonia mature a covering is secreted around them. Eventually they enter a storage area (seminal vesicle). The terminal portion of the male system is an ejaculatory duct that passes through a copulatory organ. A pair of accessory glands, often associated with the ejaculatory duct, contributes to the semen (fluid containing sperm) or participates in spermatophore formation. The ninth body segment and sometimes the tenth bear appendages for sperm transfer. Scorpions and spiders have tubular or saclike gonads; the female system is equipped to receive and store sperm, and, in some species, the female retains the eggs long after fertilization has occurred. Male spiders may have a cluster of accessory glands associated with the terminal portion of the reproductive system for the manufacture of spermatophores, or they may have expanded seminal vesicles for the retention of sperm until copulation takes place. Often specific appendages are adapted for sperm transfer.

Echinoderms and protochordates. Echinoderms (*e.g.,* sea urchins), hemichordates (including acornworms), urochordates (*e.g.,* sea squirts), and cephalochordates (amphioxus) are restricted to a marine habitat. As with many other marine animals, their gametes are shed into the water. In echinoderms, the gonads are generally suspended from the arms directly into the sea; with few exceptions, the sexes are separate. Female starfishes have been known to release as many as 2,500,000 eggs in two hours; 200,000,000 may be shed in a season. Males produce many times that number of sperm. Acornworms reproduce only sexually, and the sexes are generally separate. The gonads lie on each side of the gut as a paired series of simple or lobed sacs. Each opens to the exterior, either directly or via a short duct. The eggs, when shed, are in coiled mucous masses, each of which contains 2,500 to 3,000 eggs.

Gonads in urochordates and cephalochordates

In urochordates and cephalochordates the gonads develop in the wall of a cavity (atrium) that receives respiratory water after it passes over the gills. Gametes are released into the cavity, then carried into the sea by the water flowing from the cavity. Most urochordates are hermaphroditic. One ovary and one testis may lie side by side, each with its own duct to the atrium; some species have many pairs of ovaries and testes. The eggs develop in so-called ovarian follicles consisting of two layers of cells, as in many vertebrates. The inner layer remains with the ovulated, or shed, egg, and the cells become filled with air spaces, which apparently help the eggs to float. Amphioxus, the highest animals lacking vertebral columns, are dioecious. They have 24 or more pairs of ovaries or testes lacking ducts. When ripe, the gonads rupture, spilling their gametes directly into the atrium.

MECHANISMS THAT AID IN THE UNION OF GAMETES

Sponges, coelenterates, flatworms, and aschelminthes. The processes of sperm transfer and fertilization have been documented for only a few species of sponges. Flagellated (*i.e.*, bearing a whiplike strand) sperm are released from the male gonad and swept out of the body and into the water by way of an elaborate system of canals. A sperm that enters another sponge, or the one from which it was released, is captured by a flagellated collar cell (choanocyte). The choanocyte completely engulfs the sperm, loses its collar and flagellum (or "whip"), and migrates to deeper tissue where the egg has matured. The choanocyte containing the sperm cell fuses with the egg, thus achieving fertilization. In freshwater coelenterates, sperm are also released into the water and carried by currents to another individual. Unlike the mechanism in sponges, however, coelenterate eggs arise in the epidermis, or surface tissue, and are exposed to sperm that may be nearby in the water; thus, no intermediate transport cell is needed. Many species of marine coelenterates expel both sperm and eggs into the water, and fertilization takes place there. Some medusoid coelenterates (jellyfish), however, offer some protection to the egg. After leaving the gonad, the egg becomes temporarily lodged in the epidermis on the underside of the organism, where fertilization and early development occur.

In all flatworms, fertilization is internal. Among species with no female duct, sperm are injected, and fertilization occurs in the inner layer of tissue. Most flatworms, however, have an elaborate system of male and female ducts. Generally, the male gonoduct passes through a penis-like organ, and sperm are transferred during copulation. In parasitic species, which often cannot find a mate, self-fertilization serves as the means for reproduction. Sperm and ova unite in the oviduct, which then secretes yolk around the zygote.

Male nematodes (roundworms) are usually equipped with a pair of copulatory structures (spicules) that guide the sperm during copulation. The posterior end of some males also exhibits a lateral (sideward) expansion (copulatory bursa) that clasps the female during copulation. Other males loop their tail around the female in the region of her gonopore. Unlike many other aschelminthes, nematodes have sperm cells that are amoeboid; *i.e.*, their cell contents seem to flow. Some male rotifers have a copulatory organ.

Annelids and mollusks. In some species of annelid polychaetes (marine worms) reproductive activity is synchronized with lunar cycles. At breeding time the body of both sexes differentiates into two regions, an anterior atoke and a posterior epitoke, in which gonads develop. When the moon is in a specific phase, the epitoke separates from the rest of the body and swims to the surface. The female epitoke apparently stimulates the male epitoke to release sperm, and sperm release, in turn, evokes expulsion of eggs. Fertilization is external. So well coordinated is this phenomenon that tremendous numbers of epitokes appear on the surface at about the same time.

Sexually mature oligochaetes have a clitellum, which is a modification of a section of the body wall consisting of a glandular, saddlelike thickening near the gonopores. During copulation, the clitellum secretes a mucus that keeps the worms paired while sperm are being exchanged. Following copulation, the clitellum secretes substance for a cocoon, which encircles the worm and into which eggs and sperm are deposited. The worm then manipulates the cocoon until it slips off over the head. Thereupon, the ends of the cocoon become sealed, and fertilization and development take place inside. Many leeches also form a cocoon; but the males of some species have a penis that can be inserted into the female gonopore. In other leeches, a spermatophore is thrust into the body of the mate during copulation.

Union of gametes among mollusks is effected in a number of ways. Marine pelecypods synchronously discharge sperm and eggs into the sea; some freshwater clams are apparently self-fertilizing. One of the more unusual types of reproductive diversity occurs in marine gastropods of the family Scalidae that produce two kinds of sperm cells. A large sperm with a degenerate nucleus acts as a transport cell for carrying numerous small fertilizing sperm through the water and into the oviduct of another individual. Cephalopod males have modified arms for the transfer of spermatophores. The right or left fourth arm of the squid, for example, is so modified. Following an often elaborate courtship, the male squid uses the modified appendage to remove spermatophores from their storage place in his body and place them in the mantle cavity of the female. A cementing substance, which is released from the spermatophore, firmly attaches the spermatophore to the female's body near the oviduct. In some species, the male loses the arm. Manipulation of the eggs by the female's arms may also occur.

Sexual behaviour among mollusks

Some unusual behaviour patterns have evolved in conjunction with sperm transfer in mollusks. Prior to copulation of certain land snails, a dart composed of calcium carbonate is propelled forcefully from the gonopore of each of the mating individuals and lodges in the viscera of the mate. Even though the snails have assumed a mating posture, sperm transfer cannot occur until each snail has been stimulated by a dart.

Arthropods. Arthropods are as varied as mollusks in their methods of effecting union of sperm and eggs. They have relatively few devices for sperm transfer, but many display a high degree of behavioral complexity.

The male and female scorpion participate in a courtship ritual involving complicated manoeuvres. In some species the male produces spermatophores that are anchored to the ground. In the course of the ritual dance the female is positioned over the spermatophore. The male then presses her down until the sperm packet is forced into her genital chamber, where it becomes attached by means of small hooks. Thus, ultimately, fertilization takes place internally.

Among some spiders the male's pedipalp, a grasping or crushing appendage, contains a bulb and an extensible, coiled structure (embolus). As mating begins, the male dips the pedipalp into semen from his gonopore. The embolus is then placed in the female gonopore, and the sperm are transferred to her seminal receptacle. The female deposits the sperm along with her eggs into a silken cocoon, which she attaches to her body or to an object such as a rock or a leaf.

Sperm transfer in copepods, isopods, and many decapods, often preceded by courtship, is effected by modified appendages, gonopods, or spermatophores. Copepods clasp the female with their antennae while placing a spermatophore at the opening of the seminal receptacle. In some decapods fertilization occurs as eggs are being released into the water.

Fertilization among insects is always internal; there is much variation in the manner in which sperm are transferred to the female. Males of some species form spermatophores that are deposited in a copulatory bursa (vagina) of the female; the wall of the spermatophore breaks down, and the sperm swim to the seminal receptacle. In other species, sperm are introduced directly into the seminal receptacle by an intromittent organ. In still others, sperm, but no spermatophores, are deposited in the copulatory bursa and migrate to the seminal receptacle. In all instances, sperm are retained in the seminal receptacle until after fertilization. An exception to the usual route of sperm transfer occurs in insects that inject sperm into

the female's hemocoel (*i.e.,* the space between the body organs). The sperm then migrate to the ovary and oviduct and unite with eggs before the eggshell is formed.

PARTHENOGENESIS

Most frequently, parthenogenesis is the development of a new individual from an unfertilized gamete. Often referred to as unisexual reproduction, it has been observed in almost every major invertebrate group, with the exception of protochordates (including hemichordates), and frequently occurs alternately with bisexual reproduction (reproduction by union of gametes). Some species, in which males are completely unknown, apparently reproduce only by parthenogenesis. Species that alternate between parthenogenesis and bisexual reproduction (heterogenetic species) often do so in response to changes in population density, food availability, or other environmental conditions.

Examples of parthenogenetic reproduction

The best known examples of parthenogenetic reproduction are found among rotifers. Males are completely unknown in some genera; in others, they appear in the population only for brief periods and more or less seasonally. Females are the dominant form or are the only sex present in a population throughout most of the year. Because no reductional division (meiosis) occurs in the course of egg maturation, the eggs are diploid—that is, they have the full number of chromosomes; they give rise to new diploid individuals with no chromosomal contribution from a male gamete (diploid parthenogenesis). Even if males were present, sperm could not fertilize the eggs because the latter are already diploid. Under conditions of environmental stress such as seasonal changes, some females form eggs that undergo reductional division, resulting in eggs with the haploid number of chromosomes; such eggs must be fertilized by a male gamete to produce a new female. When the new individual matures, it will probably reproduce parthenogenetically. If, however, there are no males in the population, the haploid eggs can develop into haploid males (haploid parthenogenesis), which then participate in bisexual reproduction. Bisexually produced eggs are often referred to as winter eggs since they have a thick covering that protects the embryo during adverse environmental conditions. Summer eggs, produced parthenogenetically, are thin shelled. Bisexual reproduction occurs, therefore, only often enough to ensure survival of the species.

Nematodes, especially free-living species such as some dioecious soil nematodes, exhibit a type of parthenogenesis known as gynogenesis. In this type of reproduction, the sperm produced by males do not unite with the haploid female egg but merely activate it to begin development. The result is haploid females.

Parthenogenesis, which apparently occurs only rarely in the annelids and mollusks, is found more frequently among the arthropods. The cladocerans (*e.g.,* water fleas), for example, have a reproductive cycle much like that of rotifers—so long as environmental conditions are optimal and food is plentiful, females produce other females by diploid parthenogenesis. When conditions become adverse, males begin to appear in the population, and bisexual reproduction follows. The precise trigger for the appearance of males is not yet known. Fertilized eggs, covered with a highly resistant case, enter a resting stage (ephippium) and can withstand severe temperatures and drying out. The return of favourable conditions leads to the emergence of females that reproduce parthenogenetically. The ability to form a resting stage regulates population density. Whenever the food supply becomes short because of overpopulation by parthenogenetic females, bisexual reproduction is induced, and a dormant stage ensues. During periods of food shortage, the excess females die from lack of food, but the ephippia remain to restore the population.

Insects provide numerous examples of parthenogenesis of varying degrees of complexity. One of the most notable is that of the honeybee. Unfertilized eggs develop into drones, which are males. Fertilized eggs become worker females, which are kept in a nonreproductive state by secretions from the reproductive female, the queen bee.

Life cycles involving alternation between parthenogenesis and bisexual reproduction can be found in many species of Homoptera and Diptera (flies). Aphids (Homoptera) have a seasonal cycle consisting of a bisexual winter phase and a parthenogenetic summer phase; some species spend each phase on a different host plant. Temperature change, length of day, and food availability play major roles in initiating the phases. In the midge, a type of fly, the bisexual phase occurs in adults, and parthenogenesis takes place among the larvae (paedogenesis). Adult female midges deposit fertilized eggs, from which hatch larvae whose ovaries develop while the rest of the body retains a larval form. The ovaries of the larvae release eggs that enter the larval hemocoel (the space between body organs), where they undergo development while feeding on larval tissue. When sufficiently developed, the parthenogenetically produced young emerge either as larvae that continue parthenogenetic reproduction, forming larvae like themselves, or as male or female larvae that mature to become bisexually reproducing adults.

PROVISIONS FOR THE DEVELOPING EMBRYO

Invertebrates have developed a great many methods for protecting the fertilized egg and young embryo and for providing nutrients for the developing young. This is especially true of freshwater and terrestrial forms. Sponges and freshwater coelenterates, exposed to seasonal drying out, provide a tough covering for the eggs that prevents water loss. Many turbellarians envelop the eggs with a capsule and attach it to a hard surface, where it remains until the young emerge. Other turbellarians retain encapsulated eggs in the body until development is complete and the young emerge. All parasitic flatworms enclose their eggs in a protective capsule within which development occurs after it has left the parent's body. Most nematodes and rotifers do likewise, but a few species are ovoviviparous; *i.e.,* the egg hatches in the mother's body. In many forms the amount of yolk provided in the egg and the nature of the egg capsule are correlated with annual seasons—summer eggs generally have less yolk and thinner capsules than do winter eggs. This is true also in a number of crustaceans. Freshwater and terrestrial annelids provide a cocoon for their young and often deposit it in a moist place. One group of leeches, however, does not form a cocoon; instead, the egg, surrounded by a protective membrane, is attached to the underside of the parent. As the young develop, the adult leech undulates its body so that water currents flow over the young. Presumably this serves as a means of aeration. Mollusks that live in freshwater may provide a protective covering for the eggs, or the eggs may be brooded by the female. Some pelecypods (bivalves) release mature eggs into their gill chambers; here the eggs are fertilized, and embryonic development is completed in a protected location. Cephalopods (*e.g.,* squid, octopus) attach the eggs to a surface, then continuously force jets of water over the egg masses, thereby keeping them free of debris and perhaps aerating them. Some echinoderms also brood the eggs until the young emerge.

Differences in amount of yolk

Arthropods have a particularly wide range of methods for ensuring offspring survival. Brood pouches, common in branchiopods, isopods, and amphipods, are sometimes part of the carapace, or back plate. In other instances, expanded plates on the lower side (sternum) form the pouches. Crayfish cement the fertilized eggs to their swimmerets (modified appendages) and carry them about as they are brooded by the female. The most elaborate provisions for the embryo are found among terrestrial arthropods, especially insects. Although some species simply deposit their eggs and abandon them, many retain the encapsulated egg within the body during early development. Some are viviparous; that is, they bear living young. The eggs of certain species of scorpions have little or no yolk; the embryo is nourished by the parent in a manner similar to that in mammals—part of the scorpion oviduct becomes modified as a uterus for the embryo; another part lies close to the female's gut and absorbs nutritive substances that are conveyed to the developing young. A similar arrangement has evolved in some insects. Other viviparous insects nourish the larvae by glandular secretions from the uterine lining.

Reproductive systems of vertebrates

GONADS, ASSOCIATED STRUCTURES, AND PRODUCTS

The reproductive organs of vertebrates consist of gonads and associated ducts and glands. In addition, some vertebrates, including some of the more primitive fishes, have organs for sperm transfer or ovipository (egg-laying) organs. Gonads produce the gametes and hormones essential for reproduction. Associated ducts and glands store and transport the gametes and secrete necessary substances. In addition to these structures, most male and female vertebrates have a cloaca, a cavity that serves as a common terminal chamber for the digestive, urinary, and reproductive tracts and empties to the outside. In lampreys and most ray-finned fishes in which the cloaca is small or absent, the alimentary canal has a separate external opening, the anus. In some teleosts the alimentary, genital, and urinary tracts open independently. Hagfishes, which are closely related to the lampreys, have a short cloaca. In many vertebrates other than mammals, especially reptiles and birds, the cephalic, or head, end of the cloaca is partitioned by folds into a urinogenital chamber (urodeum) and an alimentary chamber (coprodeum) that open into a common terminal chamber (proctodeum). Above monotremes (*e.g.*, platypus, echidna) the embryonic cloaca becomes completely partitioned into a urinogenital sinus conveying urine and the products of the gonads, and an alimentary pathway; the two open independently to the exterior.

Embryonic development of gonads

Gonads arise as a pair of longitudinal thickenings of the coelomic epithelium and underlying mesenchyme (unspecialized tissue) on either side of the attachment of a supporting membrane, the dorsal mesentery, to the body wall. At first, gonadal ridges bulge into the coelom and are continuous with the embryonic kidney. The germinal epithelium covering the gonadal ridges gives rise to primary sex cords (medullary cords) that invade the underlying mesenchyme. These cords establish within the gonadal blastema (a tissue mass that gives rise to an organ) a potentially male component, the medulla. Secondary sex cords grow inward, spreading just beneath the germinal epithelium to form a cortex. If the gonad is to become a testis, only the medullary component differentiates. If the gonad is to become an ovary, only the cortex differentiates.

The length of an adult gonad depends, in part, upon the extent of gonadal-ridge differentiation. In cyclostomes (lampreys and hagfish), elasmobranchs (sharks, skates, and rays), and teleosts most of it differentiates, and the gonads extend nearly the length of the body trunk. In tetrapods (amphibians, reptiles, birds, and mammals), the cranial portion, at the anterior end, generally does not differentiate; in toads only the more caudal, or posterior, portion does so. The middle segment in toads of both sexes gives rise to a Bidder's organ containing immature eggs. In anurans (frogs and toads) and some lizards of both sexes, one segment of the gonadal ridge gives rise to yellow fat bodies that, especially in anurans, diminish in size just prior to the breeding season. In mammals, only the middle portion of the gonadal ridge differentiates.

Some vertebrate species have only one gonad, which may lie in the midline or on one side; the condition is more common among females. Adult cyclostomes of both sexes have one gonad. In lampreys it is in the middle of the body; in hagfishes it is on the right side. Birds are the only other major group of vertebrates in which most females have one gonad, the right ovary being typically absent. Male birds have a pair of testes, however. Exceptions to the condition of single ovaries among birds include members of the falcon family, in which more than 50 percent of mature hawks have two well-developed ovaries. In all bird species a small percentage of females probably have two ovaries; reported instances include owls, parrots, sparrows, and doves, with estimates for doves ranging from 5 percent to 25 percent. A few teleosts and viviparous elasmobranchs have only one ovary; in sharks the right one is usually present, in rays, the left. In amniotes (*i.e.*, reptiles, birds, and mammals) unpaired gonads are unusual. Some lizards have one testis, and some female crocodiles have one ovary. Among mammals, the platypus usually has only a left ovary, and some bat species (family Vespertilionidae) have only the right.

Theories about unpaired gonads

One of two explanations may account for unpaired gonads: the paired embryonic gonadal ridges may fuse to form a median gonad—as in lampreys and the perch—or only one gonadal ridge may receive immigrating primordial germ cells (immature sperm or eggs), with the result that the opposite gonad does not develop—as in chickens and ducks. Both gonadal ridges have been reported to exhibit an equal number of primordial germ cells in embryonic hawks, and these typically have two ovaries.

Among lower vertebrates, mature gonads sometimes produce both sperm and eggs. Hermaphroditism is more general in cyclostomes and teleosts than in other fishes. A teleost may function as a male during the early part of its sexual life and as a female later. In some teleost families sperm and eggs mature simultaneously but in different regions of the gonad. These fish normally function as males during one season and as females the next. Cyclostomes generally are ambisexual during juvenile life—*i.e.*, immature male and female sex cells exist side by side, or, as in *Myxine*, the anterior part of the immature gonad may be ovary and the caudal part, the testis. It is thought that cyclostomes normally become unisexual at maturity. Hermaphroditism is uncommon among amphibians, although it frequently occurs as an anomaly. In vertebrates above amphibians, true hermaphroditism probably does not exist.

Both male and female duct systems are occasionally absent. In cyclostomes, a few elasmobranchs, and some teleosts, such as salmon, trout, and eels, the gametes are propelled toward the posterior within the coelom, often by cilia (minute hairlike structures), and exit via a pair of funnel-like genital pores near the base of the tail. In cyclostomes, the pores lead to a sinus, or cavity, within a median papilla (*i.e.*, a fingerlike structure) and are open only during breeding seasons.

Seminiferous tubules

Male systems. *Testes.* In anurans, amniotes (reptiles, birds, and mammals), and even some teleosts, testes are composed largely of seminiferous tubules—coiled tubes, the walls of which contain cells that produce sperm—and are surrounded by a capsule, the tunica albuginea. Seminiferous tubules may constitute up to 90 percent of the testis. The tubule walls consist of a multilayered germinal epithelium containing spermatogenic cells and Sertoli cells, nutritive cells that have the heads of maturing sperm embedded in them. Seminiferous tubules may begin blindly at the tunic, or outermost tissue layer, and pass toward the centre, becoming tortuous before emptying into a system of collecting tubules, the rete testis. Such an arrangement is characteristic of frogs. In certain amniotes—the rat, for example—the tubules may be open ended, running a zigzag course from the rete to the periphery and back again. The average length of such tubules is 30 centimetres (12 inches), and they seldom communicate with each other. In many mammals the tubules are grouped into lobules separated by connective-tissue septa, or walls. The arrangement permits the packing of an extensive amount of germinal epithelium into a small space. In immature males and in adult males between breeding seasons, the tubules are inconspicuous and the epithelium is inactive; in some species, however, spermatogenesis, or production of sperm, proceeds at a variable pace throughout the year. An active epithelium may exhibit all stages of developing sperm. The lumen, or tubule cavity, contains the tails of many sperm (the heads of which are embedded in Sertoli cells), free sperm, and fluid that is probably resorbed. In mammals, in any single zone along a tubule, all sperm are at the same stage of maturation; adjacent zones contain different generations of sperm, and a period of sperm formation and discharge is followed by an interval of inactivity.

In cyclostomes, most fishes, and tailed amphibians the germinal epithelium is arranged differently. Instead of seminiferous tubules there are large numbers of spermatogonial cysts (also called spermatocysts, sperm follicles, ampullae, crypts, sacs, acini, and capsules) in which sperm develop, but in which the epithelium is not germinal. Spermatogenic cells migrate into the cysts from a per-

manent germinal layer, which, depending on the species, may lie among cysts at the periphery of the testes or in a ridge along one margin of the testis. After invading the thin nongerminal epithelium of a cyst, spermatogenic cells multiply, producing enormous numbers of sperm. The cysts become greatly swollen and whitish in colour; the entire testis also swells and has a granular appearance. As sperm mature, they separate from the epithelium and move freely in the cystic fluid. Finally, the cysts burst, and the sperm are shed into ducts. In the case of cyclostomes and a few teleosts the sperm are shed into the coelom. The cysts, totally emptied, collapse. Then either they are replaced by new ones, or they become repopulated by additional spermatogenic cells. It is not yet known which of these processes occurs.

Testicular stroma, which fills the spaces between seminiferous tubules or spermatogenic cysts, consists chiefly of connective tissue, blood and lymphatic vessels, and nerves; it is more abundant in some vertebrates than in others. Glandular Leydig (interstitial) cells are also present in most, if not all, vertebrates. Thought to be a primary source of androgens, or male hormones, Leydig cells are not always readily distinguishable, and, in some bird species, they may be seen only with the electron microscope. The capillary system of the rat testis, and probably that of many other vertebrates, is such that blood that has bathed the Leydig cells flows to the tubules; it is thus probable that Leydig cell hormones have an immediate effect on the germinal epithelium.

Location of testes in mammalian orders

Testes in vertebrates below mammals lie within the body. This is also true of many, sometimes all, members of the mammalian orders Monotremata, Insectivora, Hyracoidea, Edentata, Sirenia, Cetacea, and Proboscidea. Some male mammals—most marsupials, ungulates, carnivores, and primates after infancy—have a special pouch (scrotum) that the testes occupy permanently. A few mammals have a pouch into which the testes descend and from which they can be retracted by muscular action. These include a few rodents such as ground squirrels; most, if not all, bats; and some primitive primates (loris, potto). The scrotum consists of two scrotal sacs, each connected to the abdominal cavity by an inguinal canal lined with the peritoneal membrane. The canals are the path of descent (and retraction) of the testes to the sacs. In descending, the testes carry along a spermatic duct, blood and lymphatic vessels, and a nerve supply wrapped in peritoneum and constituting, collectively, the spermatic cord. Rabbits, most rodents, and some insectivores, which lack scrotal sacs, have instead a wide inguinal canal into which the testes may be drawn and from which they are retracted when in danger of injury. In these mammals, descended testes cause a temporary bulge in the perineal region (*i.e.*, between the anus and the urinogenital opening). In a small number of mammals, the testes permanently occupy the perineal location.

The scrotum is a temperature-regulating device. Warm blood approaching the testis comes close to the vessels carrying cool blood leaving the testis, so that the blood approaching the testis is cooled; the vessels form an intricate vascular network (pampiniform plexus) within the spermatic cord. Failure of both testes to enter the scrotal sacs (cryptorchidism) results in permanent sterility. In cold weather two sets of muscles, the dartos and cremasteric, pull the testes close to the body. The dartos lies between the two scrotal sacs and is attached to the scrotal skin. The cremaster, wrapped around the spermatic cord, is an extension of the abdominal wall musculature. It retracts the testis. Birds, like mammals, are homoiothermic (warm-blooded), and their testes are near air sacs (extensions of incurrent respiratory tubes). Air in the sacs may help regulate the temperature of the testes.

Ducts. The male duct system begins as the rete testis, a network within the testis of thin-walled ductules, or minute ducts, that collects sperm from the seminiferous tubules. The rete is drained by a number of small ducts—usually fewer than ten—called the vasa efferentia, which are modified kidney tubules. In some fishes and amphibians the vasa efferentia connect the testes with the cranial (anterior) end of the kidneys (Figure 21). In anamniotes

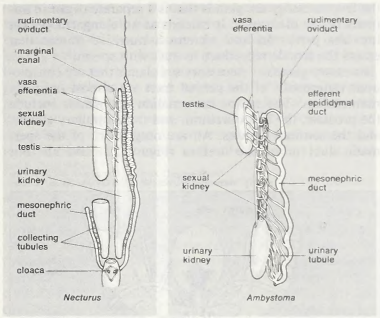


Figure 21: *Reproductive systems of two male amphibians.* (Left) The sexual kidney is composed of one row of modified tubules that are ciliated for sperm transport and drain into the adult (mesonephric) kidney duct. This duct also collects urine from the kidney. (Right) The mesonephric duct collects only sperm.

(Left and right) From George C. Kent, Jr., *Comparative Anatomy of the Vertebrates*, 2nd ed. (1969) The C.V. Mosby Co., St. Louis. (left) redrawn from C.J. Baker and W.W. Taylor *Journal of the Tennessee Academy of Science*, vol. 39, no. 1 (1964)

(*e.g.*, fish and amphibians), therefore, except teleosts, the ducts that drain the kidneys usually drain the testes also. In most amphibians these ducts pass caudad, or posteriorly, to empty independently into the cloaca; in some fishes they pass through a median urinogenital papilla.

Although drainage of the testis and the kidney by the same duct is a basic pattern, there has been a tendency in many vertebrates toward separate spermatic and urinary ducts. This tendency is manifested in one of two ways among anamniotes. In many sharks and in some amphibians (Plethodontidae, Salamandridae, Ambystomatidae), the embryonic kidney duct ultimately drains the testis, and one or more new ducts (ureters) drain the adult kidney. On the other hand, in the primitive fish *Polypterus* and in most teleosts, the embryonic kidney duct drains the adult kidney, and a new duct arises to drain the testis. Many degrees of separation of the two ducts occur in anamniotes, from the condition of the sturgeon, in which the spermatic duct unites with the urinary duct far toward the head, to the condition in *Esox* (a pike), in which spermatic and urinary ducts empty independently to the exterior.

In amniotes, the mesonephric kidney is a temporary structure confined to the embryo, but the mesonephric duct persists in the adult male as a sperm duct. A separate ureter drains the adult kidney. The spermatic and urinary ducts empty independently into the cloaca except in mammals above monotremes, in which they are confluent with the urethra. The epididymis of amniotes, a highly tortuous duct draining the vasa efferentia, usually serves as a temporary storage place for sperm; it is small in birds and large in turtles. In mammals, the first part of the epididymis consists of a head, body, and tail that wrap around the testis; it gradually straightens to become the spermatic duct. The epididymis secretes substances that prolong the life of stored sperm and increase their capacity for motility.

Sperm storage

In all vertebrates certain regions of the spermatic duct are lined by cilia and a variety of secretory epithelial cells. One end may enlarge to form a sperm reservoir or secrete seminal fluid. In the catfish *Trachycorystes murabilis* secretions of the spermatic duct form a gelatinous plug in the female similar to the vaginal plug of mammals. A seminal glomulus in birds functions as a sperm reservoir. In some mammals an enlargement of the spermatic duct called the ampulla contributes to the seminal fluid and stores sperm. Small mucous glands (of Littré) and other glandular structures open into the urethra along its length. Cloacal glands in basking sharks and many salamanders form a jelly that encloses sperm in a spermatophore. Cloacal glands of some lizards produce secretions called pheromones. The siphon sac of elasmobranchs is one of

the few accessory sex glands that is a separate organ in animals below mammals. It extends as an elongated pocket into the pelvic fin and secretes a nutritive mucus that enters the female reproductive tract with sperm.

Accessory glands. Accessory sex glands that are conspicuous outgrowths of the genital tract are almost uniquely mammalian. The major mammalian sex glands include the prostate, the bulbourethral, and the ampullary glands, and the seminal vesicles. All are outgrowths of the spermatic duct or of the urethra (Figure 22) and all four

From George C. Kent, Jr., *Comparative Anatomy of the Vertebrates*, 2nd ed. (1969); The C.V. Mosby Co., St. Louis

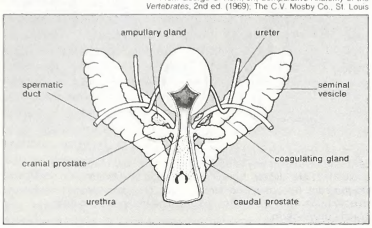


Figure 22: Accessory sex organs of a male golden hamster. Bulbourethral glands arise more toward the tail end and are not shown. The bladder and urethra have been opened to show entrances of ducts.

occur in elephants and horses and in most moles, bats, rodents, rabbits, cattle, and primates. A few members of these groups lack ampullary glands, or ampullary glands and seminal vesicles. Cetaceans (whales, porpoises) have only the prostate, as do some carnivores, including dogs, weasels, ferrets, and bears.

The prostate gland

The prostate, the most widely distributed mammalian accessory sex gland, is absent only in *Echidna* (a marsupial) and a few carnivores. It empties into the urethra by multiple ducts. Many rodents, insectivores, and lagomorphs have three separate prostatic lobes; in a few mammals (some primates and carnivores) the prostate is a single mass with lobules and encircles the urethra at the base of the bladder. In a few mammals (*e.g.,* opossum), the prostate is not a compact mass but a partly diffuse gland. In many rodents (*e.g.,* rat, guinea pig, mouse, hamster) and some other mammals, the semen coagulates quickly after ejaculation as a result of a secretion from a male coagulating gland, which is usually considered part of the prostatic mass. Coagulated semen forms a vaginal plug that temporarily prevents copulation.

Bulbourethral (Cowper's) glands arise from the urethra near the penis and are surrounded by the muscle of the urethra or penis. Typically, there is one pair, but as many as three (marsupials) may be found. The glands, small in man, large in rodents, elephants, and some ungulates including pigs, camels, and horses, are absent in cetaceans, mustelids (*e.g.,* mink, weasel), sirenians (manatees, dugongs), pholidotans (pangolins), some edentates, and carnivores such as walrus, sea lion, bear, and dog.

Although many mammals have an ampullary swelling on the spermatic duct near the urethra, only a small number form a separate ampullary gland as an outgrowth of the duct. It is very large in some bats, absent in many mammalian orders, and variable in the rest. Although common in rodents, it is absent in guinea pigs and some strains of mice.

Seminal vesicles are paired, typically elongated and coiled fibromuscular sacs that empty into either the spermatic duct or the urethra. Absent in monotremes, marsupials, carnivores, cetaceans, and in some insectivores, chiropterans, and primates, seminal vesicles are exceptionally large in rhesus monkeys and small in man. They are absent in domesticated rabbits, small or rudimentary in cottontails, large in armadillos, and variable in sloths. They contribute the sugar fructose and citric acid to the semen but do not serve as sperm reservoirs.

Female systems. *Ovaries.* Ovaries lie within the body cavity and are suspended by a dorsal mesentery (mesovarium), through which pass blood and lymph vessels and nerves. Primitive vertebrate ovaries occur in the hagfish, in which a mesentery-like fold of gonadal tissue stretches nearly the length of the body cavity. Unique in the hagfish is the fact that functional ovarian tissue occupies only the forward half of the gonadal mass, the rear part containing rudimentary testicular tissue. In most fishes except very primitive forms, the ovaries are similarly elongated. In tetrapods other than mammals, the ovaries are usually confined to the middle third or half of the body cavity, particularly during nonbreeding seasons. The ovaries of mammals undergo moderate caudal displacement, finally coming to lie between the kidney and the pelvis.

Characteristics of the ovary

The appearance of an ovary depends on many factors—*e.g.,* whether one egg or thousands are discharged (ovulated); whether the eggs are immature or ripe; whether mature eggs are small or large; or whether pigments occur in the egg cytoplasm, such as those responsible for yellow yolk. Other factors also affect the appearance of the ovary: the season of the year in seasonal breeders (the ovary enlarges during breeding seasons, diminishes in size between seasons); the age of the animal (whether juvenile, reproductively active, or senile, particularly in birds and mammals); and the fate of ovulated, or discharged, egg follicles, or sacs.

The ovaries are covered with a germinal epithelium that is continuous with the peritoneum lining the body cavity. The term germinal epithelium is inappropriate because in most adults it contains no germ cells, these having moved deeper into the ovary. In hagfishes and amphibians, cells that give rise to eggs are known to occur in the germinal epithelium, and it may be that the germinal epithelium in a few other vertebrates contains similar cells. The germinal epithelium undergoes cell division, however. This is particularly true of species in which enormous expansion of the ovary occurs each breeding season. Beneath the epithelium is a layer of connective tissue, the tunica albuginea, which is much thinner than that surrounding the testes.

A typical vertebrate ovary consists of cortex and medulla. The cortex, immediately internal to the tunica albuginea, contains future eggs and, at one time or another, eggs in ovarian follicles (*i.e.,* developing eggs); it undergoes fluctuations in size and appearance that correlate with stages of the reproductive cycle. The cortex also contains remnants of ovulated follicles and, in mammals, clusters of interstitial cells that, in some species, are glandular. The cortical components are embedded in a supportive framework of connective, vascular, and neural tissue constituting the stroma. Internal to the cortex is the medulla, consisting of blood and lymph vessels, nerves, and connective tissue. The medulla, which contains no germinal elements, exhibits no significant cyclical activity, is usually inconspicuous, is continuous with the dorsal mesentery, and, in cyclostomes, is hardly distinguishable from the latter. The mammalian medulla, on the contrary, is almost completely surrounded by cortex and converges on the mesovarium (*i.e.,* the part of the peritoneum that supports the ovary) at a narrow hilus, at which nerves and vessels enter the ovary. In the medulla of the mammalian ovary near the hilus are small masses of blind tubules or solid cords—the rete ovarii—which are homologous (*i.e.,* of the same embryonic origin) with the rete testis in the male. The microscopic right ovary of birds usually consists only of medullary tissue.

Ovaries are characterized as saccular, hollow, lacunate (*i.e.,* compartmented), or compact. The ovary of many teleosts, especially viviparous ones, contains a permanent cavity, which is formed during ovarian development when an invagination of the ovarian surface traps a portion of the coelom. The cavity is therefore unique in that it is lined by germinal epithelium. The lining develops numerous ovigerous folds that project into the lumen and greatly increase the surface area for proliferation of eggs. In most other teleosts, a temporary ovarian cavity develops after each ovulation, when the shrinking cortex withdraws from the outside ovarian wall along one side of the ovary. The resulting cavity is obliterated as eggs of the next generation enlarge. The permanent and temporary cavities of teleost ovaries and a similar cavity in garfish ovaries are

continuous with the lumen of the oviduct, and eggs are shed into them. The ovaries of other fishes lack cavities and are characterized as compact. The amphibian ovary, which contains six or more central, hollow sacs that give it a lobed appearance, is characterized as saccular. The sacs are formed when the embryonic medullary and rete cords become hollow and coalesce. Maturing eggs bulge into the sacs but are not shed into them. The ovaries of reptiles, birds, and monotremes have cavities homologous to those in amphibians; the number of medullary spaces in the adults is considerably larger, however, so that the ovaries contain an extensive network of fluid-filled cavities (lacunae). Such ovaries are characterized as lacunate. The ovaries of mammals above monotremes are compact, having no medullary cavities.

Formation and fate of ovarian follicles

An ovarian follicle consists of an oocyte, or immature egg, surrounded by an epithelium, the cells of which are referred to variously as follicular, nurse, or granulosa cells. In cyclostomes, teleosts, and amphibians, the epithelium is one layer thick. In the hagfish and those vertebrates in which the oocyte receives heavy deposits of yolk (elasmobranchs, reptiles, birds, and monotremes), the epithelium appears to be two cells thick, apparently the result of layering of nuclei in a simple columnar epithelium (*i.e.,* epithelium consisting of relatively "tall" cells). Above monotremes the follicular epithelium appears to be many cells thick; in at least one species, however, this is considered an artifact, and all granulosa cells are said to extend between the outer boundary of the epithelium and the oocyte.

The follicular epithelium originates as a few flattened cells derived from the germinal epithelium. Primary follicles are usually situated just under the tunica albuginea; secondary follicles lie deeper in the cortex. The primitive role of the follicular cells appears to be the secretion of the yolk-forming material onto or into the oocyte. Evidence from mammals indicates that the follicular cells may also have a role in converting substances produced elsewhere into female hormones, or estrogens. In some hibernating bats the granulosa cells are filled with glycogen, or animal starch, which may be a source of energy. Mammalian follicles above monotremes are unique in that they develop a fluid-filled cavity (antrum) within the granulosa layer. During antrum formation cell division of the granulosa cells increases, and fluid-filled spaces develop among the cells. The spaces coalesce to form the antrum. Under the influence of pituitary gonadotropic hormones, many antral follicles thereafter continue to grow, forming large so-called Graafian follicles—less than 400 microns, or 0.4 millimetre (0.16 inch), in diameter in large mammals, 150–200 microns, or 0.15–0.2 millimetre (0.006–0.008 inch), in small ones. Graafian follicles contain mature eggs and appear as large blisters on the ovary. At this stage the ovum, suspended within the fluid of the antrum (liquor folliculi) by a slender stalk of granulosa cells, is surrounded by a cluster of these cells, the cumulus oophorus, or discus proligerus. The remaining follicular cells form a thin wall surrounding the antrum. Antra are lacking in a few insectivores (*Hemicentetes, Euriculus*) because the granulosa cells swell and multiply to form corpora lutea, masses of yellow tissue. In the bat *Myotis* the antrum is likewise compressed and disappears just before discharge of the egg, or ovulation.

In all vertebrates, oocytes that have begun to grow and mature may, at any time until just before ovulation, cease development and undergo atresia, or degeneration. This is a normal process that reduces the number of eggs ovulated. In small laboratory rodents, atresia takes place in 50 percent of the Graafian follicles in each ovary one or two days before ovulation, thus reducing the number of ovulatable eggs by 50 percent. A similar reduction takes place in hagfish prior to ovulation. Atretic follicles eventually become lost in the stroma of the cortex of the ovary. In mammals especially, follicles lacking oocytes and antra, called anovular follicles, as well as polyovular follicles (*i.e.,* containing more than one oocyte), occasionally occur.

The theca

The ovarian follicle of vertebrates, commencing with hagfish, is surrounded by a theca, or sheath, composed of two concentric layers of stromal cells. The outer layer (theca externa) is chiefly connective tissue but may contain smooth muscle fibres. The inner layer (theca interna) has more blood vessels and, in vertebrates that produce heavily yolked eggs, the largest vessels carry venous blood. In these species the cell membranes of the oocyte and granulosa cells have many microvilli (*i.e.,* fingerlike projections), which probably facilitate transport of substances important in yolk formation from the blood vessels to the egg. Mature follicles in the marsupial *Dasyatus* are said to lack theca, and in some bats only one thecal layer has been described.

During the growth phase, eggs in species with massive amounts of yolk may increase in size 10^6 (1,000,000) or more times as a result of vitellogenesis (deposit of yolk). In goldfish, on the other hand, when vitellogenesis commences, the egg has a diameter of 150 microns (0.15 millimetre [0.006 inch]); that of the mature egg is only 500 microns (0.5 millimetre [0.02 inch]). Mammalian eggs contain little yolk and vary little in size. Oogonia (*i.e.,* cells that form oocytes) of the golden hamster average 15 microns (0.015 millimetre [0.0006 inch]) in diameter, and eggs in Graafian follicles average 70 microns (0.07 millimetre [0.003 inch]). The mature eggs of horses and humans are approximately the same size—somewhat less than 150 microns. In seasonally breeding oviparous fishes and amphibians, all eggs are usually in the same stage of development, and the ovary grows to a mature state quite rapidly as a result of growth of the eggs, which frequently number more than 1,000,000. Such ovaries distend the body wall when mature; following spawning, the ovaries shrink rapidly to inconspicuous bodies consisting mainly of oogonia, immature oocytes, and a few stromal cells. In reptiles and birds, ovarian weight also is high in proportion to body weight during egg-laying seasons. The weight of the ovary of the starling, for example, may increase from eight milligrams in early winter to 1,400 milligrams immediately before ovulation. The mature eggs of reptiles and birds are unique in that they are suspended from the ovary by a short stalk (pedicle). The stalk contains a cortex with additional oocytes in various stages of development and extensions of vessels and nerves. Full growth of the follicle in reptiles and birds requires only a few days or weeks (nine days in the domestic hen). In mammals, the ratio of ovarian weight to body weight varies insignificantly throughout the reproductive life of the female, and follicles in many stages of development are constantly present.

Vertebrate eggs are almost universally shed into the coelom or into a subdivision thereof, from which they enter the female reproductive tract. Even in those teleosts in which the eggs are shed into an ovarian cavity, the latter is often of coelomic origin. In many mammals a membranous sac of peritoneum, the ovarian bursa, traps part of the coelom in a chamber along with the ovary. The bursal cavity (periovarian space) may be broadly open to the main coelom, completely closed off from the coelom, or in communication with the coelom by a narrow, slitlike passage. The bursa, moderately developed in lower primates and catarrhines (Old World monkeys), is poorly developed in man. In horses, one edge of the ovary contains a long groove (ovulation fossa) into which all eggs are shed; the groove is found in a cleftlike ovarian bursa. The ovarian bursa increases the probability that all ovulated eggs will enter the oviduct.

Ovulation

The process of ovulation has been described for all vertebrate classes. Elasmobranchs, reptiles, and birds have massively yolked eggs. As ovulation approaches, the fimbria (*i.e.,* frills, or fringes) of the membranous and muscular funnel surrounding the entrance to the oviduct wave in a gentle, undulating motion. An egg that is nearly free of the ovary is grasped and partially encompassed by the fimbria; when the egg is freed, the fimbria draw the egg into the funnel. At this time, the egg has little shape and is partly squirted and partly flows into the oviduct; never completely free in the coelom, its chances of not entering the oviduct are small. In the case of moderately or poorly yolked eggs cilia help to sweep the eggs into the ostium, or opening, of the oviduct. During ovulation in Japanese rice fish, *Oryzias latipes,* a tiny papilla, or fingerlike process, develops on the surface of a bulging mature follicle

in the centre (stigma) of the follicle. The follicle becomes thin at the stigma, an aperture appears, and the egg rolls out. In rabbits this process differs only in detail. During the final 20 minutes before ovulation in rabbits, some of the tiny blood vessels surrounding the stigma rupture, and a small pool of blood forms under the apex of the cone-shaped papilla. The follicular wall shortly gives way at the apex, and follicular fluid oozes from the opening, followed soon after by the egg. The ovulated mammalian egg typically is surrounded by a layer of columnar follicular cells, the corona radiata; but it is naked in some insectivores and some marsupials. Following ovulation in all vertebrates, the ovary may become smaller, become modified for maintenance of pregnancy, or proceed to form additional eggs.

The process of ovulation in vertebrates has been documented, but the immediate causes remain to be clarified. It is almost certain that an ovulatory hormone is secreted by the pituitary gland (*i.e.*, the so-called master endocrine gland) of all vertebrates. It is highly probable that breakdown of very small fibres that bind the follicular cells together may occur at the stigma, weakening the follicular wall at that location. Hormones from the ovary and other sources may play a role, as may neurohormones, which are hormones released at nerve endings. Rhythmic contractions of the entire ovary occur at ovulation in many vertebrates and have been described in rabbits. The role of mechanical pressure within the follicle, however, is not understood. Ovulation in most mammals (spontaneous ovulators) occurs cyclically as a result of the spontaneous release of the ovulatory hormone. In a few mammals (reflex ovulators) the stimulus of copulation is essential for release of the ovulatory hormone.

Striking postovulatory changes take place in the follicles of mammals and, to lesser degrees, of lower vertebrates. Blood vessels from the theca interna invade the ovulated follicles; the granulosa cells divide, enlarge, accumulate fats, and obliterate any remnants of the collapsed antra. Thereafter, they are known as lutein cells. Theca interna cells undergo changes identical to those of the granulosa cells. The result in mammals is the formation of solid masses called corpora lutea, recognizable as prominent reddish-yellow bulges on the ovary. Corpora lutea produce the hormone progesterone, which is essential for the maintenance of pregnancy. The conversion of postovulatory follicles into structures more or less resembling mammalian corpora lutea has been demonstrated in numerous viviparous reptiles, amphibians, and elasmobranchs; in certain other fishes, including cyclostomes; and in some oviparous amphibians and reptiles. In birds, the postovulatory follicle shrinks, and identifiable corpora lutea do not develop, although some granulosa cells accumulate lipids of unknown significance.

Tracts. The female reproductive tract consists of a pair of tubes (gonoducts) extending from anterior, funnel-like openings (ostia) to the cloaca (Figure 23), except as noted below. The gonoducts are specialized along their length for secretion of substances added to the eggs; for transport, storage, nutrition, and expulsion of eggs or the products of conception; and, in species with internal fertilization, for receipt, transport, storage, and nutrition of inseminated sperm. The predominately muscular tracts are lined by a secretory epithelium and ciliated over at least part of their length. Fusion of the caudal (tail) ends of the paired ducts may occur. Gonoducts are absent in cyclostomes and a few gnathostome fishes that have abdominal pores. A few vertebrates have only one functional gonoduct.

From George C. Kent, Jr., *Comparative Anatomy of the Vertebrates*, 2nd ed. (1969). The C.V. Mosby Co., St. Louis

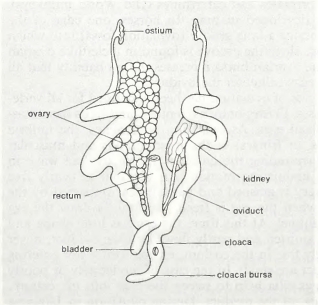


Figure 23: Reproductive tract of female turtle, *Trionyx euphraticus*. One ovary has been removed.

Gonoducts of fishes and amphibians

Gonoducts in lungfishes and amphibians are coiled muscular tubes that are ciliated over most of their length. Only occasionally do they unite caudally in a genital papilla before opening into the cloaca. During breeding seasons their diameter increases severalfold because of the highly active secretory epithelium. Between breeding seasons they are small. In some anurans (frogs, toads), such as *Rana*, the lower end of each gonoduct is expanded to form an ovisac, in which ovulated eggs are stored until spawning; the tube between the ostium (funnel-like opening) and ovisac is the oviduct. In viviparous amphibians the young develop in the ovisac. In amphibians, numerous multicellular glands extend deep into the lining of the female tract. Six successive glandular zones have been described in some urodeles, and these secrete six different gelatinous substances upon the egg. Female urodeles often have convoluted tubular outpocketings of the cloaca called spermatheca; they temporarily store sperm liberated from the male spermatophore.

The two gonoducts of elasmobranchs share a single ostium, a trait found only in Chondrichthyes. The ostium is a wide caudally directed funnel supported in the falciform ligament, which is attached to the liver. The role of the fimbria of the ostium at ovulation has been described (see above *Ovaries*). Two oviducts pass forward from the ostium to the septum transversum (*i.e.*, between the heart and abdominal cavities), curve around one end of the liver, then pass posteriorly on each side. Approximately midway between ostium and uterus each oviduct has a shell (nidamental) gland. Fertilization takes place above the shell gland, which may be immense or almost undifferentiated. Half of the shell gland secretes a substance high in protein content (albumen), and the other half secretes the shell—delicate in viviparous forms, thick and horny in most oviparous species. Horny shells may have spiral ridges and many long tendrils, which entwine about an appropriate surface after the egg is deposited. In the viviparous shark *Squalus acanthias* several eggs pass one after the other through the shell gland, where they are enclosed in one long delicate membranous shell that soon disintegrates. Beyond the shell gland the oviducts terminate in an enlargement, which, in viviparous species, serves as a uterus. An oviducal valve may be found at the junction of oviduct and uterus. Although the two uteri usually open independently into the cloaca, they occasionally unite to form a bicornuate (two-horned) structure. In immature females, the uterus may be separated from the cloaca by a hymen, or membrane. The tract enlarges enormously during the first pregnancy and does not thereafter fully regress to its original size.

The gonoducts of most lower ray-finned fishes resemble those of lungfish, but those of gars and teleosts are exceptional in that the oviducts are usually continuous with the ovarian cavities. A median genital papilla receives the oviducts in teleosts, and the papilla is sometimes elongated to form an ovipositor. European bitterlings deposit their eggs in a mussel by means of the ovipositor, and female pipefish and sea horses deposit them in the brood pouch of a male.

Gonoducts of reptiles, birds, and mammals

With certain modifications, the gonoducts of reptiles and birds are comparable to those of lower vertebrates. Crocodilians, some lizards, and nearly all birds have one gonoduct; the other is not well developed. Even in birds of prey having two functional ovaries, the right oviduct is sometimes undeveloped. The tracts of reptiles generally show less regional differentiation than do those of birds. The oviduct funnel (ostium) in birds forms the chalazae—two coiled, springlike cords extending from the yolk to the

ends of the egg. In both reptiles and birds, much of the length of the female tract is oviduct. This region, called the magnum in birds, secretes albumen; lizards and snakes do not form albumen. Behind the albumen-secreting region is a shell gland. In lizards, the gland is midway along the tract. In birds, the shell gland is at the posterior end, has thick muscular walls, and is often inappropriately called a uterus. It is preceded by a narrow region, or isthmus, which secretes the noncalcareous, or soft, membranes of the shell. The shell gland leads to a narrow muscular vagina that empties into the cloaca. The vagina secretes mucus that seals the pores of the shell before the egg is expelled. Special vaginal tubules (spermatheca) store sperm over winter in some snakes and lizards; seminal receptacles have been described in the oviduct funnel in some snakes. In birds, sperm storage glands (sperm nests) often occur in the funnel and at the uterovaginal junction. In lizards and birds, ovulation does not usually occur into a tract already containing an egg. Some lizards shed very few eggs per season; the gecko, for example, sheds only two.

The female reproductive tracts of monotremes, the egg-laying mammals, consist of two oviducts, the lower ends of which are shell glands. These open into a urinogenital sinus, which, in turn, empties into a cloaca. Marsupials have two oviducts, two uteri (duplex uterus), and two vaginas. The upper parts of the vaginas unite to form a median vagina that may or may not be paired internally. Beyond the median vagina, the vaginas are again paired (lateral vaginas) and lead to a urinogenital sinus. The posterior end of the pouchlike median vagina is separated from the forward end of the urinogenital sinus by a partition. When the female is delivering young, the fetuses are usually forced through the partition and into the urinogenital sinus, bypassing the lateral vaginas. The ruptured partition may remain open thereafter, resulting in a pseudovagina. It closes in opossums, and in kangaroos both the median and lateral routes may serve as birth canals. The lateral vaginas in marsupials receive the forked tips of the male penis. Fertilization in all mammals takes place in the oviducts (Fallopian tubes).

In eutherian mammals (*i.e.*, all mammals except monotremes and marsupials), with exceptions noted below, female reproductive tracts beyond the ostia (oviduct funnels) consist of two narrow and somewhat tortuous Fallopian tubes, two large uterine horns (each of which receives a Fallopian tube), a uterine body, and one vagina. Fallopian tubes often have a short dilated ampulla, or saclike swelling, just beyond the ostium. Implantation of the egg occurs only in the uterine horns; the embryos become spaced equidistant from one another in both horns even if only one ovary has ovulated. In some species one horn is rudimentary—the left in the impala (an African antelope)—and the embryos become implanted in the other horn, even though both ovaries ovulate. The body of the uterus in some mammals (*e.g.*, rabbits, elephants, aardvarks; some rodents, bats, insectivores) contains two separate canals (bipartite uterus). In other mammals (ungulates, many cetaceans, most carnivores and bats) the body of the uterus has one chamber into which the two horns empty (bicornuate uterus). There are numerous intermediate conditions between the bipartite and bicornuate condition. Apes, monkeys, and man have no horns, and the Fallopian tubes empty directly into the body of the uterus (simplex uterus). In all mammals, the uterine body tapers to a narrow neck (cervix). The opening (os uteri) into the vagina is guarded by fleshy folds (lips of the cervix). The vagina in eutherian mammals other than rodents and primates terminates in a urinogenital sinus that opens to the exterior by a urinogenital aperture. In some rodents and in higher primates the vagina opens directly to the exterior. In the young of many species a membrane, the hymen, closes the vaginal opening. In guinea pigs the hymen reseals the opening after each reproductive period. Sperm are stored over winter in the uterus of some bats and in vaginal pouches in others.

Accessory glands. Female mammals have fewer accessory sex glands than males, the most prominent being Bartholin's glands and prostates. Bartholin's (bulbovestibular) glands are homologues of the bulbourethral glands of males. One pair usually opens into the urinogenital sinus or, in primates, into a shallow vestibule at the opening of the vagina. Prostates develop as buds from the urethra in many female embryos but often remain partially developed. They become well developed, however, in some insectivores, chiropterans, rodents, and lagomorphs, although their function is obscure. A variety of glands (labial, preputial, urethral) are found in the mucosa, or mucous membrane. Glands in the uterine mucosa provide nourishment for embryos before implantation. Cervical uterine glands secrete mucus that lubricates the vagina, which has no glands.

ADAPTATIONS FOR INTERNAL FERTILIZATION

Fertilization among vertebrates may be external or internal, but internal fertilization is not always correlated with viviparity or the presence of intromittent (copulatory) organs. The latter, uncommon among fishes, amphibians, and birds, are present in all reptiles (except *Sphenodon*) and mammals.

Copulatory organs in fishes and amphibians

A considerable number of fishes are viviparous; in them, fertilization is internal, and the males have intromittent organs. The claspers of most male elasmobranchs are usually paired extensions of pelvic fins that are inserted into the female's uterus for transfer of sperm. The clasper, supported by modified fin cartilages, contains a groove along which sperm are conveyed into the uterus and is raised, or erected, by muscles at its base. Gonopodia of male teleosts are fleshy, often elongated modifications of pelvic or anal fins that are directed posteriorly, have a genital pore at the end, and often serve as intromittent organs. In some teleosts, a large penis-like papilla located under the throat is supported by bones. The spermatic duct opens on one side of the papilla. In a few teleosts, hemal spines (ventral projections of vertebrae) form the skeleton of an intromittent organ. Occasionally, the intromittent organ is an asymmetrical tube that matches the asymmetrical genital opening of the female. Still other teleosts have uncomplicated fleshy genital papillae that can be erected. In at least one teleost species, the female has a copulatory organ that she inserts into the genital pore of the male for receiving sperm.

Certain amphibians have internal fertilization but no intromittent organs. The muscular cloaca of the male caecilian, however, can be everted (turned outward) to protrude into that of the female. The male urodele deposits a spermatophore that the female picks up with the lips of her cloaca. Among anurans, *Nectophrynoides* (a viviparous frog) and *Ascaphus* (a toad) have internal fertilization, but only *Ascaphus* has an intromittent organ. It is a permanent tubular extension of the cloaca and resembles a tail. Other anurans have external fertilization and no intromittent organs.

Copulatory organs in reptiles, birds, and mammals

The provision of an eggshell in reptiles requires that fertilization be internal, and all reptiles have intromittent organs except *Sphenodon*. Reptilian intromittent organs are of two types. Crocodilians and chelonians (turtles) have a penis (phallus), a median thickening in the floor of the cloaca consisting of two cylinders of spongy vascular erectile tissue, the corpora spongiosa. The caudal tip of the penis protrudes into the cloaca as a genital tubercle, or glans penis. The penis is held in the cloacal floor by retractor muscles. When the blood vessels within the spongy bodies are filled with blood, the penis swells, the retractor muscle relaxes, and the genital tubercle protrudes from the vent to serve as an intromittent organ. A longitudinal groove on the surface of the penis directs the flow of sperm. When the spongy bodies are no longer filled with blood, the retractor muscle returns the penis to the cloacal floor. Snakes and lizards have hemipenes, paired elongated outpocketings of the caudal wall of the cloaca that extend under the skin at the base of the tail. Each hemipenis is held in place by a retractor muscle. During copulation the muscle relaxes, the pocket turns inside out and protrudes through the vent in an erect condition. Semen passes along grooves on its surface. Except in pythons, erectile tissue is lacking in hemipenes. Hemipenes protrude independently of each other and are often covered with spines. Very small hemipenes of unknown function are usually present in females.

All birds have internal fertilization, although they are not viviparous; most lack intromittent organs. Male swans, ducks, geese, tinamous, ostriches, and some other ratites (flightless birds), however, have an erectile median penis like that of crocodiles and turtles. Chickens have an organ consisting of a small amount of erectile tissue, but lymph vessels, rather than blood vessels, become engorged. Some birds have a vestigial penis.

All mammals have internal fertilization and an erectile penis. That of monotremes is of the reptilian type, nonprotrusible and in the cloacal floor. In higher mammals the penis has been modified. The groove on the surface of the embryonic penis becomes enclosed in a tube along with the corpus spongiosum and two additional erectile masses, the corpora cavernosa. The proximal ends (crura) of the corpora cavernosa are anchored laterally to the pubic and ischial bones by various muscles and constitute the root of the penis. The crura converge in the midline to enter the body of the penis, which also contains the urethra, surrounded by the corpus spongiosum. The latter begins on the pelvic floor as the bulb of the penis and contains a dilation of the urethra (urethral bulb). The body of the penis extends a variable distance beyond the body of the mammal, in contrast to the short genital tubercle of reptiles. Except in ruminants (*i.e.,* cud-chewing animals, such as cattle and deer), cetaceans, and some rodents, the penis terminates in a glans penis, a swelling of the corpus spongiosum that caps the ends of the corpora cavernosa and contains the urinogenital aperture. The glans is supplied with nerve endings and is partly or wholly sheathed, except during erection, by a circular fold of skin, the prepuce. The inner surface of the prepuce is moistened by preputial glands, and the external surface is sometimes covered with spines or hard scales, as in the cat, guinea pig, and wombat. The glans penis of the male Virginia opossum (*Didelphis virginiana*), the bandicoot, and some other species is bifid (*i.e.,* with two equal tips), correlated with the paired vaginas of females. In boars, the glans penis is corkscrew-shaped, and in goats, rams, and many antelopes a urethral (vermiform) process of much smaller diameter extends three or four centimetres (about an inch to an inch and a half) beyond the glans. In some cattle, a sigmoid, or S-shaped, flexure bends the penis, which otherwise would be too long to fit into the preputial sac. The penis of marsupials is directed backward, and that of cats and rodents is directed backward, except during copulation. In some mammals (*e.g.,* bats) it is pendulous; and in armadillos it may extend one third the length of the body during copulation.

Erectile tissue

Erection of the mammalian penis is initiated typically by an increase in the volume of blood reaching the cavernous and spongy bodies, engorgement of the vessels, and consequent compression of the veins leaving the organ. When a retractor muscle is present (wolf, fox, dog), it relaxes as erection occurs. The amount of erectile tissue in bovines (cattle) is small, and the penis has much fibroelastic tissue. Erection in such species results primarily from relaxation of the retractor muscle, and vascular engorgement provides only rigidity. Among mechanisms that reverse the erectile state are disgorgement of blood from the cavernous spaces, elasticity of the walls of the spaces, and action of a retractor muscle. A penis bone (baculum, os priapi) is present in various degrees of development in many mammals.

Female mammals have an erectile penile organ known as the clitoris in the floor of the urinogenital sinus or vagina. In the young spider monkey *Ateles,* the clitoris is six or seven centimetres (2.4 to 2.8 inches) long. In a few mammals (some rodents, insectivores, lemurs, and hyenas) the urethral canal becomes enclosed within the clitoris, as in males. In hyenas, the clitoris is large and often mistaken for a penis, and female scrotal pouches, lacking gonads, are present. So much do the male and female external genitalia resemble each other that the ancients regarded the hyena as a hermaphrodite. The clitoris of female mammals often contains cartilage or bone. A specialized clitoris is present in female turtles, crocodiles, alligators, and a few species of birds in which the male has a penis.

The spermatic duct of male mammals between the seminal vesicle and the prostatic urethra has a heavy muscular coat and serves as an ejaculatory duct. In mammals in which the seminal vesicles empty directly into the urethra, the latter is ejaculatory. In birds, the terminal segments of the spermatic ducts are erectile and ejaculatory, and in fish the posterior end of whatever duct transports sperm may be ejaculatory.

ROLE OF GONADS IN HORMONE CYCLES

Neurosecretions formed in the brain in response to environmental stimuli regulate the synthesis and release of hormones known as gonadotropins, which, in turn, stimulate the gonads. Cyclical intervals of illumination (photoperiods) may be the principal environmental factor regulating gonadal activity. Although cyclical temperature changes are experienced by many species, as are fluctuations in food supply, rainfall, and salinity, their precise effects and those of many other stimuli, independently or in combination, have not yet been defined for any species. Photoperiodicity, temperature, and perhaps all other cycles are attributable to the seasons, and to the 24-hour day.

As a result of rhythmic stimulation by gonadotropins secreted by the pituitary gland, the gonads grow, mature, and produce gametes and hormones. Certain of these hormones, known as androgens, are thought to be produced chiefly by interstitial cells and are more abundant in males. Hormones known as estrogens are probably produced chiefly by ovarian follicles and their thecas. Circulating progestins are produced in greatest quantities by corpora lutea. Although the gonadal hormones of different species vary somewhat in structure, their effects are essentially the same. As the quantity of pituitary gonadotropins decreases, the activity of the gonads slows and may temporarily cease.

The effects of gonadal hormones may be summarized as follows:

Effects of gonadal hormones

Gonadal hormones induce growth of and maintain the cyclical function of the reproductive tracts, accessory sex glands, and copulatory or ovipository organs. They thereby provide for the storage, nutrition, and transport of gametes; the secretion of necessary substances onto the surface of gametes; and the ultimate extrusion of sperm, eggs, or the products of conception. In mammals, therefore, they prepare the vagina for copulation and the uterus for implantation of eggs: in addition, gonadal hormones maintain pregnancy until birth or until placental hormones can take over their function. The hormonal basis for the maintenance of viviparity in vertebrates below mammals is almost unknown.

Gonadal hormones participate in the maturation of gametes still in the gonads by augmenting the metabolic effects of other hormones.

Gonadal hormones are essential for the differentiation of many secondary sex characters—the physical differences between the sexes—facilitate amplexus (copulatory embrace) and provide for the protection or nutrition of young. Secondary sex characters include scent glands; sexually linked pigmentation of the skin or its appendages; the nature of any vocal apparatus; hardened areas on the appendages that facilitate amplexus; distribution of hair; body size; mammary gland development; and other features.

Gonadal hormones participate in the induction of behaviour necessary for the union of sperm and eggs; this includes migratory phenomena, heat (estrus) in mammals, courtship, territorial defense, mating, and care of eggs or young.

Gonadal hormones participate in a mechanism that affects the pituitary, thereby imposing certain restraints on the secretion of gonadotropins.

The role of light in establishing periodic reproductive activity

The effects of a cyclical environment on gonads is illustrated in mammals that ovulate spontaneously. Ovulation is induced by ovulatory hormones released rhythmically from the pituitary gland. Newborn mice maintained during the first week of life in regular, natural photoperiods will, on reaching maturity, ovulate regularly. Newborn mice kept in continuous light during this interval will not ovulate regularly. The photoperiods in which these animals live as neonates, or newborn, establish an intrinsic brain rhythm that subsequently results in cyclical

reproductive activity. If mature female mice that have been ovulating regularly are subjected to continuous light, ovulation ultimately becomes arrhythmical. This suggests that the rhythmical environment is the ultimate regulator of the gonads. Because of the effects of cyclical photoperiods, spontaneous ovulation occurs about the same time of day or night in all members of species intensively studied thus far. Golden hamsters ovulate shortly after midnight; chickens and Japanese rice fish ovulate in the morning. Not all mammals ovulate spontaneously, however. In those that do not (*e.g.*, reflex ovulators), including some cats, rodents, weasels, shrews, rabbits, the act of mating substitutes for the environmental effects on the pituitary gland in releasing ovulatory hormones (see BIOCHEMICAL COMPONENTS OF ORGANISMS: *Hormones*).

PROVISIONS FOR THE DEVELOPING EMBRYO

Among the requirements of developing embryos are nutrients, oxygen, a site in which to discharge metabolic wastes, and protection from the environment. These needs exist whether the embryo is developing outside the body of the female parent (oviparity), or within, so that she delivers living young (viviparity). Combinations of yolk, albumen, jellies, and shells contributed by the female parent, as well as membranes constructed from the tissues of the embryo meet the embryo's needs.

Oviparous eggs are usually supplied with enough nutrients to last until the new individual is able to obtain food from the environment. The alternative, postnatal parental feeding, is uncommon. Oviparous animals that develop from yolk-laden eggs are not hatched until they resemble adults. Those that develop from eggs with moderate amounts of yolk hatch sooner, usually into free-living larvae; in this case the larvae transform, or undergo metamorphosis, into adults. The eggs of amphioxus, an oviparous protochordate, contain almost no nutrients; the embryos hatch in an extremely undeveloped but self-sustaining state as few as eight hours after fertilization. The yolk mass is large in some animals and becomes surrounded by a membrane called the yolk sac, the vessels of which convey yolk to the embryo. In some species, yolk also passes from the yolk sac directly into the fetal intestine.

Oviparous fishes and amphibians develop in an aquatic environment, and exchange of oxygen and carbon dioxide and elimination of metabolic wastes occur through the egg membranes. Oviparous reptiles, birds, and monotremes develop on land, and gaseous exchange is accomplished by two membranes (allantois, chorion) applied closely to the shell. The allantois also receives some wastes. Drying out or mechanical injury of embryos of reptiles, birds, and mammals is prevented by still another membrane, the amnion, which is a fluid-filled sac immediately surrounding the embryo.

Types of viviparity

Viviparity has evolved in some members of all vertebrate classes except birds. When eggs heavily laden with yolk and surrounded by a well-formed shell develop within the female, the parent may provide the developing young only with shelter and oxygen (ovoviviparity). At the opposite extreme, if eggs contain only enough nutrients to supply energy for a few cell divisions after fertilization, the female provides shelter, oxygen, and nourishment, and, in addition, excretes all metabolic wastes produced by the developing organism (euviviparity). Between these extremes are numerous intermediate degrees of dependence on the parent.

Teleosts have evolved many unusual adaptations for viviparity. In some viviparous teleosts the eggs are fertilized in the ovarian follicle, where development occurs. The granulosa cells either form a membrane that secretes nutrients and assists in respiratory and excretory functions or they may be ingested along with follicular fluid, nearby eggs, and other ovarian tissue. A common site for development is the ovarian cavity, which may become distended with as many as nine series of embryos of different ages. Embryos in this location are bathed with nutritive fluids secreted by the epithelium of the cavity. In some species, mortality rates of intraovarian young are high, and surviving individuals ingest those that die. In still other species, extensions of villi in the ovarian lining invade the mouth and opercular (gill) openings of the embryo, filling the opercular chamber, mouth, and pharynx with surfaces that secrete nutrients. The embryos also develop specialized surfaces for nutrition, respiration, and excretion. An enlarged pericardial (heart) sac or an expansion of the hindgut of the embryo may occur next to the blood-vessel containing (vascular) follicular wall. Vascular extensions may grow out of the anus, urinogenital pore, or gills of the embryo. Other embryonic surfaces—including ventral body wall, fins, and tail—may participate in the support of viviparity. These embryonic surfaces may lie in contact with the follicular or ovarian epithelium, or they may simply be bathed by ovarian fluids. One or more combinations of the maternal and embryonic specializations described above, as well as many others, make viviparity possible among teleost fishes. In a number of teleosts the eggs are incubated, or brooded, in the mouth of the male for periods as long as 80 days. The oral epithelium becomes vascular and highly glandular. In sea horses and pipefish the female deposits her eggs in a ventral brood pouch of the male, and the embryos develop there.

Oral incubation in fish

In viviparous elasmobranchs development takes place in the uterus, the lining of which develops parallel ridges or folds covered with villi or papillae (trophonemata) that constitute a simple placenta (site of fetal–maternal contact). In contact with this region is the yolk sac of the embryo, which serves as a respiratory and nutritive membrane. Trophonemata secrete uterine fluids that supplement the yolk as a source of energy. In one shark (*Pteroplatea micrura*), trophonemata extend into the spiracular chamber (an opening for the passage of respiratory water) of the young and secrete nutrients into the fetal gut. In another (*Mustelus antarcticus*), the uterine folds form fluid-filled compartments for each embryo. The yolk sac may lie in contact with the uterine lining, or projections of the sac may extend into uterine pits. When the stored yolk is used up before birth, the yolk sac may serve for the absorption of nutrients; *i.e.*, as a placenta. In a few species, immature eggs that enter the oviduct are eaten by the developing young.

Very few amphibians bear living young. In the viviparous frog *Nectophrynoides*, all development, including larval stages, occurs in the uteri and the young are born fully metamorphosed; *i.e.*, except for size they resemble adults. *N. occidentalis*, an African species, has a nine-month gestation period. There is almost no yolk in the egg and no placenta, so it is probable that uterine fluids provide nourishment and oxygen. In *N. vivipara* there are as many as 100 larvae in the uteri, each with long vascular tails that may function as respiratory membranes. *Gastrotheca marsupiata* is an ovoviviparous anuran with a gestation period of three to four months. In certain viviparous salamanders the extent of the nutritional dependence on the mother varies. After depleting their own yolk supply, the larvae of some forms eat other embryos and blood that escapes from the uterine lining. Conventional viviparity is rare among amphibians; however, they have evolved unusual alternatives. In some anurans the young develop in such places as around the legs of the male (*Alytes*), or in pouches in the skin of the back (some females of the genera *Nototrema*, *Protopipa*, and *Pipa*). In *Pipa*, vascular partitions in the skin pouch separate developing young, and the larvae have vascular tails that absorb substances. In *Nototrema* larval gills have vascular extensions with a similar function. The male Chilean toad (*Rhinoderma darwinii*) carries developing eggs in the vocal sac until the young frogs emerge.

Some snakes and lizards and all mammals except monotremes exhibit viviparity to some degree. The same extra-embryonic membranes found in oviparous reptiles and mammals (yolk sac, chorioallantoic membrane, amnion) function in viviparous ones. Here, the extra-embryonic membranes lie against the uterine lining instead of against an egg shell. At special sites of fetal–maternal contact (placentas), viviparous young receive oxygen and give up carbon dioxide; metabolic wastes are transferred to maternal fluids and tissues; and, in euviviparous species, the young receive all their nutrients. Yolk-sac placentas are common in marsupials with short gestation periods (opos-

Placentas

sum, kangaroo) and in lizards. Chorioallantoic placentas (*i.e.*, a large chorion fused with a large allantois) occur in certain lizards, in marsupials with long gestation periods, and in mammals above marsupials. The yolk-sac placenta does not invade maternal tissues, but intimate interlocking folds may occur between the two. The chorioallantoic membranes of reptiles and mammals exhibit many degrees of intimacy with maternal tissues, from simple contact to a deeply rooted condition (deciduate placentas). Chorioallantoic or chorionic placentas represent specializations in a chorionic sac surrounding the embryo. The entire surface of the sac may serve as a placenta (diffuse placenta, as in pigs); numerous separate patches of placental thickenings may develop (cotyledonary placenta, as in sheep); a thickened placental band may develop at the equator of the chorionic sac (zonary placenta, as in cats); or there may be a single oval patch of placental tissue (discoidal placenta, as in higher primates). (G.C.K.)

THE HUMAN REPRODUCTIVE SYSTEM

Man belongs to that group of mammals characterized by the bearing of live offspring that have attained considerable development within the uterus, or womb. Provided all organs are present, normally constructed, and functioning properly, the essential features of human reproduction are (1) liberation of an egg from the ovary at the right time in the reproductive cycle; (2) internal fertilization by spermatozoa (sperm, or male sex cells) of the ovum in the uterine tube; (3) transport of the fertilized ovum along the uterine tube to the uterus; (4) implantation of the blastocyst, the early embryo that develops from the fertilized ovum, in the wall of the uterus; (5) formation of a placenta and maintenance of the intra-uterine existence of the unborn child; (6) birth of the child and expulsion of the placenta; and (7) suckling and care of the child, with an eventual return of the maternal organs virtually to their original state.

For this biological process to be carried out, certain organs and structures are required in both male and female bodies. The source of the ova, the female germ cells, is the female gonad or ovary; that of spermatozoa is the testis. In human beings, the two ovaries are situated in the pelvic cavity, and the two testes are enveloped in a sac of skin, the scrotum, lying below and outside the abdomen. Besides producing the germ cells, or gametes, the ovaries and testes are the source of hormones that cause full development of secondary sexual characteristics and also the proper functioning of the genital (reproductive) tracts. These tracts comprise the uterine tube, the uterus, vagina, and associated structures in females, and, in males, the penis, the sperm channels—epididymis, ductus deferens, and ejaculatory ducts—and other related structures and glands. The function of the uterine tube is to convey an ovum, which is fertilized in the tube, to the uterus, where gestation (development before birth) takes place. The function of the male ducts is to convey spermatozoa from the testis, to store them, and, when ejaculation occurs, to eject them with secretions from the male glands through the penis.

At copulation the erect penis is inserted into the vagina and spermatozoa contained in the seminal fluid are ejaculated into the female genital tract. Spermatozoa then pass from the vagina through the uterus to the uterine tube to fertilize the ovum in the outer part of the tube. Human females exhibit a periodicity in the activity of their ovaries and uterus, which starts at puberty and ends at the menopause. The periodicity is manifested by menstruation at intervals of about 28 days; important changes occur in the ovaries and uterus during each reproductive or menstrual cycle. Periodicity is suppressed during pregnancy and lactation.

The sex of a human child is determined at the time of fertilization of the ovum by the spermatozoon. The differences between a man and a woman are genetically determined by the chromosomes that each possesses in the nuclei of the cells. This stage in the development of the individual is detailed in the article GROWTH AND DEVELOPMENT: *Human embryology*.

Once the genetic sex has been determined there normally follows a succession of changes that will result, finally, in the development of an adult male or female. There is, however, no external indication of the sex of a human embryo during the first eight weeks of its life within the uterus. This is a neutral or indifferent stage during which the sex of an embryo can be ascertained only by examination of the chromosomes in its cells. The next phase, one of differentiation, begins first in gonads that are to become testes, and a week or so later in those destined to be ovaries. Embryos of the two sexes are initially alike in possessing similar duct systems linking the undifferentiated gonads with the exterior and in having similar external genitalia, represented by three simple protuberances. The embryos each have four ducts, the subsequent fate of which is of great significance in the eventual anatomical differences between men and women. Two ducts closely related to the developing urinary system are called mesonephric, or wolffian, ducts. In males, each mesonephric duct becomes differentiated into four related structures: a duct of the epididymis, a ductus deferens, an ejaculatory duct, and a seminal vesicle (see below). In females, the mesonephric ducts are largely suppressed. The other two ducts, called the paramesonephric or müllerian ducts, persist, in females, to develop into the uterine tubes, the uterus, and part of the vagina; in males they are largely suppressed. Differentiation also occurs in the primitive external genitalia, which in males become the penis and scrotum, and in females the clitoris and labia.

Development of sexual differences in embryo

At birth the organs appropriate to each sex have developed and are in their adult positions but are not functioning. Various abnormalities can occur during development of sex organs in human embryos, leading to hermaphroditism, pseudohermaphroditism, and other chromosomally induced conditions. During childhood until puberty there is steady growth in all reproductive organs and a gradual development of activity. Puberty marks the onset of increased activity in the sex glands and the steady development of secondary sexual characteristics.

In males at puberty the testes enlarge and become active, the external genitalia enlarge, and the capacity to ejaculate develops. Marked changes in height and weight occur as hormonal secretion from the testes increases. The larynx, or voice box, enlarges, with resultant deepening of the voice. Certain features in the skeleton, as seen in the pelvic bones and skull, become accentuated. The hair in the armpit and the pubic hair becomes abundant and thicker. A beard, a moustache, and cheek hair develop, as well as hair on the chest, abdomen, and limbs. Hair at the temple recedes. Skin glands become more active, especially apocrine glands (a type of sweat gland that is found in the armpit and groin and around the anus).

These secondary sex characteristics do not develop in individuals castrated before puberty, but the administration of androgens (male sex hormones) to such persons and to males having poorly developed testes can correct, in large measure, some of the poorly developed secondary characteristics. Large amounts of androgen, however, by preventing production of the hormone gonadotrophin by the pituitary, suppress testicular activity, thus depressing formation and release of sperm. Some derivatives of the male sex hormone testosterone can promote general bodily development.

In females at puberty, the external genitalia enlarge and the uterus commences its periodic activity with overt menstruation. The mammary glands develop, and there is a deposition of body fat in accordance with the usual contours of the mature female. Growth of axillary (armpit) and pubic hair is more abundant, and the hair becomes thicker. In a female receiving androgens, the typical male secondary sex characteristics may develop, menstruation may be suppressed, and the mammary glands may atrophy.

The male reproductive system

The male gonads are the testes; they are the source of spermatozoa and also of male sex hormones called androgens. The other genital organs are the epididymides, the ductus or vasa deferentia, the seminal vesicles, the ejaculatory ducts, and the penis, as well as certain accessory structures, the prostate and the bulbourethral (Cowper's) glands. The principal functions of these structures are to transport the spermatozoa from the testes to the exterior, to allow their maturation on the way, and to provide certain secretions that help form the seminal fluid.

EXTERNAL GENITALIA

The penis. The penis, the male organ of copulation, is partly inside and partly outside the body. The inner part, attached to the bony margins of the pubic arch (that part of the pelvis directly in front and at the base of the trunk), is called the root of the penis. The second, or outer, portion is free, pendulous, and enveloped all over in skin; it is termed the body of the penis. The organ is composed chiefly of cavernous or erectile tissue that becomes engorged with blood to produce considerable enlargement and erection. The penis is traversed by a tube, the urethra, which serves as a passage both for urine and for semen.

Adapted from H. Gray, *Anatomy of the Human Body*, 28th ed. by C.M. Goss (1966). Lea & Febiger

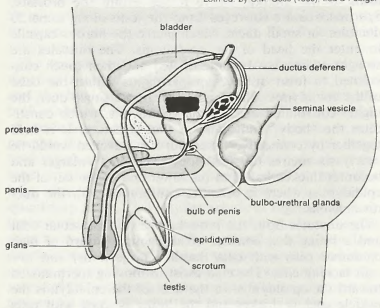


Figure 24: Male reproductive organs.

The body of the penis, sometimes referred to as the shaft, is cylindrical in shape when flaccid but when erect is somewhat triangular in cross section, with the angles rounded. This condition arises because the right and left corpora cavernosa penis, the masses of erectile tissue, lie close together in the dorsal part of the penis, while a single body, the corpus spongiosum penis, which contains the urethra, lies in a midline groove on the under surface of the corpora cavernosa. The dorsal surface of the penis is that which faces upward and backward during erection.

The corpus spongiosum and the corpora cavernosa

The slender corpus spongiosum reaches beyond the extremities of the erectile corpora cavernosa, and at its outer end is enlarged considerably to form a soft, conical, sensitive structure called the glans penis. The base of the glans has a projecting margin, the corona, and the groove where the corona overhangs the corpora cavernosa is referred to as the neck of the penis. The glans is traversed by the urethra, which ends in a vertical, slitlike, external opening. The skin over the penis is thin and loosely adherent and at the neck is folded forward over the glans for a variable distance to form the prepuce or foreskin. A median fold, the frenulum of the prepuce, passes to the under surface of the glans to reach a point just behind the urethral opening. The prepuce can usually be readily drawn back to expose the glans.

The root of the penis comprises two crura, or projections, and the bulb of the penis. The crura and the bulb are attached respectively to the edges of the pubic arch and to the perineal membrane (the fibrous membrane that forms a floor of the trunk). Each crus is an elongated structure covered by the ischiocavernosus muscle, and each extends

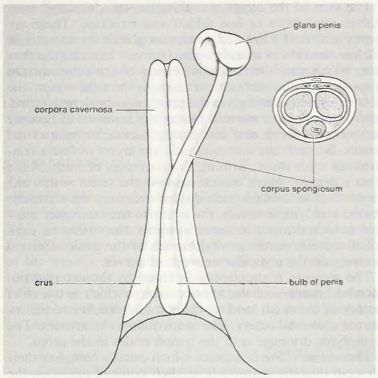


Figure 25: Interior view of the penis with corpus spongiosum detached and turned to one side.

Adapted from H. Gray, *Anatomy of the Human Body*, 28th ed. by C.M. Goss (1966). Lea & Febiger

forward, converging toward the other, to become continuous with one of the corpora cavernosa. The oval bulb of the penis lies between the two crura and is covered by the bulbospongiosus muscle. It is continuous with the corpus spongiosum. The urethra enters it on the flattened deep aspect that lies against the perineal membrane, traverses its substances, and continues into the corpus spongiosum.

The two corpora cavernosa are close to one another, separated only by a partition in the fibrous sheath that encloses them. The erectile tissue of the corpora is divided by numerous small fibrous bands into many cavernous spaces, relatively empty when the penis is flaccid but engorged with blood during erection. The structure of the tissue of the corpus spongiosum is similar to that of the corpora cavernosa, but there is more smooth muscle and elastic tissue. A deep fascia, or sheet of connective tissue, surrounding the structures in the body of the penis, is prolonged to form the suspensory ligament, which anchors the penis to the pelvic bones at the mid point of the pubic arch.

The penis has a rich blood supply from the internal pudendal artery, a branch of the internal iliac artery, which supplies blood to the pelvic structures and organs, the buttocks, and the inside of the thighs. Erection is brought about by distension of the cavernous spaces with blood, which is prevented from draining away by compression of the veins in the area.

The penis is amply supplied with sensory and with autonomic (involuntary) nerves. Of the autonomic nerve fibres the sympathetic fibres cause constriction of blood vessels, and the parasympathetic fibres cause their dilation. It is usually stated that ejaculation is brought about by the sympathetic system, which at the same time inhibits the desire to urinate and also prevents the seminal fluid from entering the bladder.

The scrotum. The scrotum is a pouch of skin lying below the pubic symphysis and just in front of the upper parts of the thighs. It contains the testes and lowest parts of the spermatic cord. A scrotal septum or partition divides the pouch into two compartments and arises from a ridge, or raphe, visible on the outside of the scrotum. The raphe turns forward onto the under surface of the penis and is continued back onto the perineum (the area between the legs and as far back as the anus). This arrangement indicates the bilateral origin of the scrotum from two genital swellings that lie one on each side of the base of the phallus, the precursor of the penis or clitoris in the embryo. The swellings are also referred to as the labioscrotal swellings because in females they remain separate to form the labia majora, while in males they unite to form the scrotum.

Septum and raphe

The skin of the scrotum is thin, brown, devoid of fatty tissue, and more or less folded and wrinkled. There are some scattered hairs and sebaceous glands on its surface. Below the skin is a layer of involuntary muscle, the dartos, which can alter the appearance of the scrotum. On exposure of the scrotum to cold air or cold water, the dartos contracts and gives the scrotum a shortened, corrugated appearance; warmth causes the scrotum to become smoother, flaccid, and less closely tucked in around the testes. Beneath the dartos muscle are layers of fascia continuous with those forming the coverings of each of the two spermatic cords, which suspend the testes within the scrotum and contain each ductus deferens, the testicular blood and lymph vessels, the artery to the cremaster muscle (which draws the testes upward), the artery to each ductus deferens, the genital branch of the genito-femoral nerve, and the testicular network of nerves.

The scrotum is supplied with blood by the external pudendal branches of the femoral artery, which is the chief artery of the thigh, and by the scrotal branches of the internal pudendal artery. The veins follow the arteries. The lymphatic drainage is to the lymph nodes in the groin.

The testes. The two testes, which usually complete their descent into the scrotum from their point of origin on the back wall of the abdomen in the seventh month after conception, are suspended in the scrotum by the spermatic cords. Each testis is four to five centimetres (about one to one and one-half inches) long and is enclosed in a fibrous sac, the tunica albuginea. This sac is lined internally by the tunica vasculosa, containing a network of blood

Adapted from H. Gray, *Anatomy of the Human Body*, 28th ed. by C.M. Goss (1966); Lea & Febiger

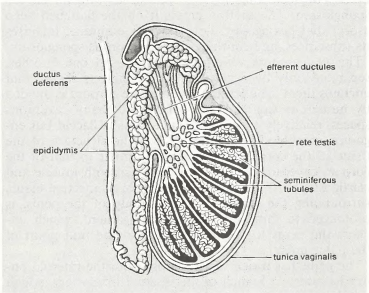


Figure 26: Longitudinal cross section of testis.

vessels, and is covered by the tunica vaginalis, which is a continuation of the membrane that lines the abdomen and pelvis. The tunica albuginea has extensions into each testis that act as partial partitions to divide the testis into approximately 250 compartments, or lobules.

Each lobule contains one or more convoluted tubules, or narrow tubes, where sperm are formed. The tubules, if straightened, would extend about 70 centimetres (about 28 inches). The multistage process of sperm formation, which takes about 60 days, goes on in the lining of the tubules, starting with the spermatogonia, or primitive sperm cells, in the outermost layer of the lining. Spermatozoa (sperm) leaving the tubules are not capable of independent motion, but undergo a further maturation process in the ducts of the male reproductive tract; the process may be continued when, after ejaculation, they pass through the female tract. Maturation of the sperm in the female tract is called capacitation; little is known about it.

Each spermatozoon is a slender elongated structure with a head, a neck, a middle piece, and a tail. The head contains the cell nucleus. When the spermatozoon is fully mature, it is propelled by the lashing movements of the tail.

Production of hormones

The male sex hormone testosterone is produced by the cells of Leydig. These cells are located in the connective (interstitial) tissue that holds the tubules together within each lobule. The tissue becomes markedly active at puberty under the influence of the interstitial-cell-stimulating hormone of the anterior lobe of the pituitary gland; this hormone in women is called luteinizing hormone. Testosterone stimulates the male accessory sex glands (prostate, seminal vesicles) and also brings about the development of male secondary sex characteristics at puberty. The hormone may also be necessary to cause maturation of sperm and heighten the sex drive of the male. The testis is also the source of some of the female sex hormone estrogen, which may exert an influence on pituitary activity.

Each testis is supplied with blood by the testicular arteries, which arise from the front of the aorta just below the origin of the renal (kidney) arteries. Each artery crosses the rear abdominal wall, enters the spermatic cord, passes through the inguinal canal, and enters the upper end of each testis at the back. The veins leaving the testis and epididymis form a network, which ascends into the spermatic cord. The lymph vessels, which also pass through the spermatic cord, drain to the lateral and preaortic lymph nodes. Nerve fibres to the testis accompany the vessels; they pass through the renal and aortic nerve plexuses, or networks.

STRUCTURES OF THE SPERM CANAL

The epididymis, ductus deferens, and ejaculatory ducts form the sperm canal. Together they extend from the testis to the urethra, where it lies within the prostate. Spermatozoa are conveyed from the testis along some 20 ductules, or small ducts, which pierce the fibrous capsule to enter the head of the epididymis. The ductules are straight at first but become dilated and then much convoluted to form distinct compartments within the head of the epididymis. They each open into a single duct, the highly convoluted duct of the epididymis, which constitutes the "body" and "tail" of that structure. It is held together by connective tissue but if unravelled would be nearly six metres (20 feet) long. The duct enlarges and becomes thicker walled at the lower end of the tail of the epididymis, where it becomes continuous with the ductus deferens.

The ductules from the testis have a thin muscular coat and a lining that consists of alternating groups of high columnar cells with cilia (hairlike projections) and low cells lacking cilia. The cilia assist in moving spermatozoa toward the epididymis. In the duct of the epididymis the muscle coat is thicker and the lining is thick with tufts of large nonmotile cilia. There is some evidence that the ductules and the first portion of the duct of the epididymis remove excess fluid and extraneous debris from the testicular secretions entering these tubes. The blood supply to the epididymis is by a branch from the testicular artery given off before that vessel reaches the testis.

The ductus deferens is the continuation of the duct of the epididymis. It commences at the lower part of the tail of the epididymis and ascends along the back border of the testis to its upper pole. Then, as part of the spermatic cord, it extends to the deep inguinal ring. Separating from the other elements of the spermatic cord—the blood vessels, nerves, and lymph vessels—at the ring, the ductus deferens makes its way through the pelvis toward the base of the prostate, where it is joined by the seminal vesicle to form the ejaculatory duct. A part of the ductus that is dilated and rather tortuous, near the base of the urinary bladder, is called the ampulla.

The ductus deferens has a thick coat of smooth muscle that gives it a characteristic cordlike feel. The longitudinal muscle fibres are well developed, and peristaltic contractions (contractions in waves) move the spermatozoa toward the ampulla. The mucous membrane lining the interior is in longitudinal folds and is mostly covered with nonciliated columnar cells, although some cells have nonmotile cilia. The ampulla is thinner walled and probably acts as a sperm store.

ACCESSORY ORGANS

The prostate, seminal vesicles, and bulbourethral glands. These structures are the male accessory reproductive organs and provide secretions to form the bulk of the seminal fluid of an ejaculate. The prostate is in the lesser or true

pelvis, centred behind the lower part of the pubic arch. It lies in front of the rectum. The prostate is shaped roughly like an inverted pyramid; its base is directed upward and is immediately continuous with the neck of the urinary bladder. The urethra traverses its substance. The two ejaculatory ducts (see below) enter the prostate near the upper border of its posterior surface. The prostate is of a firm consistency, surrounded by a capsule of fibrous tissue and smooth muscle. It measures about four centimetres across, three centimetres in height, and two centimetres front to back (about 1.6 by 1.2 by 0.8 inch) and consists of glandular tissue contained in a muscular framework. It is imperfectly divided into three lobes. Two lobes at the side form the main mass and are continuous behind the urethra. In front of the urethra they are connected by an isthmus of fibromuscular tissue devoid of glands. The third, or median, lobe is smaller and variable in size and may lack glandular tissue. There are three clinically significant concentric zones of prostatic glandular tissue about the urethra. A group of short glands that are closest to the urethra and discharge mucus into its channel are subject to simple enlargement. Outside these is a ring of submucosal glands (glands from which the mucosal glands develop), and farther out is a large outer zone of long branched glands, composing the bulk of the glandular tissue. Cancer of the prostate is almost exclusively confined to the outer zone. The glands of the outer zone are lined by tall columnar cells that secrete prostatic fluid under the influence of androgens from the testis. The fluid is thin, milky, and slightly acid.

Structure of prostate

The seminal vesicles are two structures, about five centimetres (two inches) in length, lying between the rectum and the base of the bladder. Their secretions form the bulk of the seminal fluid. Essentially, each vesicle consists of a much-coiled tube with numerous diverticula or outpouches that extend from the main tube, the whole being held together by connective tissue. At its lower end the tube is constricted to form a straight duct or tube that joins with the corresponding ductus deferens to form the ejaculatory duct. The vesicles are close together in their lower parts, but they are separated above where they lie close to the deferent ducts. The seminal vesicles have longitudinal and circular layers of smooth muscle, and their cavities are lined with mucous membrane, which is the source of the secretions of the organs. These secretions are ejected by muscular contraction during ejaculation. The activity of the vesicles is dependent on the production of the hormone androgen by the testes; castration causes atrophy of the seminal vesicles. The secretion is thick, sticky, and yellowish: it contains the sugar fructose and is slightly alkaline.

The bulbourethral glands, often called Cowper's glands, lie on the underside of the urethra between the prostate and the bulk of the penis. They are hardly larger than a pea. Each has a slender duct that runs forward and toward the centre to open on the floor of the spongy portion of the urethra. These glands are poorly developed in man. Their secretion is liberated during sexual excitement and may help to lubricate and coat the urethra to assist passage of the ejaculate.

Reduction of Cowper's glands in man

Ejaculatory ducts. The two ejaculatory ducts lie on each side of the midline and are formed by the union of the duct of the seminal vesicle, which contributes secretions to the seminal fluid, with the end of the ductus deferens at the base of the prostate. Each duct is about two centimetres (about 0.8 inch) long and passes between a lateral and the median lobe of the prostate to reach the floor of the prostatic urethra. This part of the urethra has on its floor (or posterior wall) a longitudinal ridge called the urethral crest. On each side is a depression, the prostatic sinus, into which open the prostatic ducts. In the middle of the urethral crest is a small elevation, the colliculus seminalis, on which the opening of the prostatic utricle is found. The prostatic utricle is a short diverticulum or pouch lined by mucous membrane; it may correspond to the vagina or uterus in the female. The small openings of the ejaculatory ducts lie on each side of or just within the opening of the prostatic utricle. The ejaculatory ducts are thin-walled and lined by columnar cells.

The female reproductive system

The female gonads or sexual glands are the ovaries; they are the source of ova and of the female sex hormones estrogens and progestogens. The uterine tubes conduct ova to the uterus, which lies within the lesser or true pelvis. The uterus connects through the cervical canal with the vagina. The vagina opens into the vestibule about which lie the external genitalia, collectively known as the vulva.

EXTERNAL GENITALIA

The female external genitalia include the structures placed about the entrance to the vagina and external to the hymen, the membrane across the entrance to the vagina. They are the mons pubis (also called the mons veneris), the labia majora and minora, the clitoris, the vestibule of the vagina, the bulb of the vestibule, and the greater vestibular glands.

The mons pubis is the rounded eminence, made by fatty tissue beneath the skin, lying in front of the pubic symphysis. A few fine hairs may be present in childhood; later, at puberty, they become coarser and more numerous. The upper limit of the hairy region is horizontal across the lower abdomen.

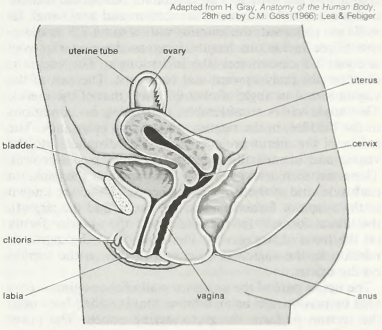


Figure 27: Female reproductive system.

The labia majora are two marked folds of skin that extend from the mons pubis downward and backward to merge with the skin of the perineum. They form the lateral boundaries of the vulval or pudendal cleft, which receives the openings of the vagina and the urethra. The outer surface of each labium is pigmented and hairy; the inner surface is smooth but possesses sebaceous glands. The labia majora contain fat and loose connective tissue and sweat glands. They correspond to the scrotum in the male and contain tissue resembling the dartos muscle. The round ligament (see below *The uterus*) ends in the tissue of the labium. The labia minora are two small folds of skin, lacking fatty tissue, that extend backward on each side of the opening into the vagina. They lie inside the labia majora and are some four centimetres (about 1.6 inches) in length. In front, an upper portion of each labium minus passes over the clitoris—the structure, in the female, corresponding to the penis (excluding the urethra) in the male—to form a fold, the prepuce of the clitoris, and a lower portion passes beneath the clitoris to form its frenulum. The two labia minora are joined at the back across the midline by a fold that becomes stretched at childbirth. The labia minora lack hairs but possess sebaceous and sweat glands.

The clitoris is a small erectile structure composed of two corpora cavernosa separated by a partition. Partially concealed beneath the forward ends of the labia minora, it possesses a sensitive tip of spongy erectile tissue, the glans clitoridis. The external opening of the urethra is some 2.5 centimetres (about one inch) behind the clitoris and immediately in front of the vaginal opening.

Clitoris and hymen

The vestibule of the vagina is the cleft between the labia

minora into which the urethra and vagina open. The hymen vaginae lies at the opening of the vagina: it is a thin fold of mucous membrane that varies in shape. After rupture of the hymen, the small rounded elevations that remain are known as the carunculae hymenales. The bulb of the vestibule, corresponding to the bulb of the penis, is two elongated masses of erectile tissue that lie one on each side of the vaginal opening. At their posterior ends lie the greater vestibular glands, small mucous glands that open by a duct in the groove between the hymen and each labium minus. They correspond to the bulbourethral glands of the male.

The blood supply and nerve supply of the female external genital organs are similar to those supplying corresponding structures in the male.

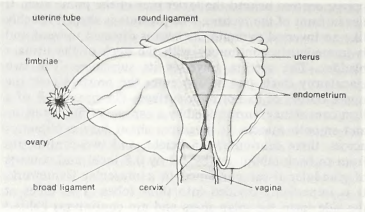


Figure 28: The uterus.
Adapted from H. Gray, *Anatomy of the Human Body*, 28th ed. by C.M. Goss (1966). Lea & Febiger

INTERNAL STRUCTURES

The vagina. The vagina (the word means "a sheath") is the canal that extends from the cervix (outer end) of the uterus within the lesser pelvis down to the vestibule between the labia minora. The orifice of the vagina is guarded by the hymen. The vagina lies behind the bladder and urethra and in front of the rectum and anal canal. Its walls are collapsed; the anterior wall is some 7.5 centimetres (three inches) in length, whereas the posterior wall is about 1.5 centimetres (0.6 inch) longer. The vagina is directed obliquely upward and backward. The axis of the vagina forms an angle of over 90° with that of the uterus. This angle varies considerably, depending on conditions in the bladder, in the rectum, and during pregnancy. The cervix of the uterus projects for a short distance into the vagina and is normally pressed against its posterior wall. There are, therefore, recesses in the vagina at the back, on each side, and at the front of the cervix. These are known as the posterior fornix (behind the cervix and the largest), the lateral fornices (at the sides), and the anterior fornix (at the front of the cervix). The position of the uterus in relation to the vagina is described further in the section on the uterus.

The upper part of the posterior wall of the vagina is covered by peritoneum or membrane that is folded back onto the rectum to form the recto-uterine pouch. The lower part of the posterior vaginal wall is separated from the anal canal by a mass of tissue known as the perineal body.

Mucous membrane and muscle coat of vagina

The vagina has a mucous membrane and an outer smooth muscle coat closely attached to it. The mucous membrane has a longitudinal ridge in the midline of both the anterior and posterior walls. The ridges are known as the columns of the vagina; many rugae or folds extend from them to each side. The furrows between the rugae are more marked on the posterior wall and become especially pronounced before birth of a child. The membrane undergoes little change during the menstrual cycle (except in its content of glycogen, a complex starchlike carbohydrate); this is in contradistinction to the situation in many mammals in which marked exfoliation (shedding of the surface cells) can occur. No glands are present in the vaginal lining, and mucus present has been secreted by the glands in the cervical canal of the uterus. The smooth muscle coat consists of an outer longitudinal layer and a less developed inner circular layer. The lower part of the vagina is surrounded by the bulbospongiosus muscle, a striped muscle attached to the perineal body.

The blood supply to the vagina is derived from several adjacent vessels, there being a vaginal artery from the internal iliac artery and also vaginal branches from the uterine, middle rectal, and internal pudendal arteries, all branches of the internal iliac artery. The nerve supply to the lower part of the vagina is from the pudendal nerve and from the inferior hypogastric and uterovaginal plexuses.

The uterus. The uterus, or womb, is shaped like an inverted pear. It is a hollow, muscular organ with thick walls, and it has a glandular lining called the endometrium. The human uterus is normally a single structure, termed a simplex uterus; in the majority of mammals there are two horns or pouches to the uterus (bicornuate). In an adult virgin the uterus is 7.5 centimetres (three inches) long, five centimetres (two inches) in breadth, and 2.5 centimetres (one inch) thick, but it enlarges to four to five times this size in pregnancy. The narrower, lower end is called the cervix; this projects into the vagina. The cervix is made of fibrous connective tissue and is of a firmer consistency than the body of the uterus. The two uterine tubes enter the uterus at opposite sides, near its top. The part of the uterus above the entrances of the tubes is called the fundus; that below is termed the body. The body narrows towards the cervix, and a slight external constriction marks the juncture between the body and the cervix.

The uterus does not lie in line with the vagina but is usually turned forward (anteverted) to form approximately a right angle with it. The position of the uterus is affected by the amount of distension in the urinary bladder and in the rectum. Enlargement of the uterus in pregnancy causes it to rise up into the abdominal cavity, so that there is closer alignment with the vagina. The nonpregnant uterus also curves gently forward; it is said to be anteflexed. The uterus is supported and held in position by the other pelvic organs, by the muscular floor or diaphragm of the pelvis, by certain fibrous ligaments, and by folds of peritoneum. Among the supporting ligaments are two double-layered broad ligaments, each of which contains a uterine tube along its upper free border and a round ligament, corresponding to the gubernaculum testis of the male, between its layers. Two ligaments—sometimes called Mackenrodt's ligaments—at each side of the cervix are also important in maintaining the position of the uterus.

The cavity of the uterus is remarkably small in comparison with the size of the organ. Except during pregnancy, the cavity is flattened, with front and rear walls touching, and is triangular. The triangle is inverted, with its base at the top, between the openings of the two uterine tubes, and with its apex at the internal os, the opening into the cervix. The canal of the cervix is flattened from front to back and is somewhat larger in its middle part. It is traversed by two longitudinal ridges and has oblique folds stretching from each ridge in an arrangement like the branches of a tree. The cervical canal is 2.5 centimetres (about one inch) in length; its opening into the vagina is called the external os of the uterus. In virgins the external os is small, almost circular, and often depressed. After childbirth, the external os becomes bounded by lips in front and in back and is thus more slitlike. The cervical canal is lined by a mucous membrane containing numerous glands that secrete a clear, alkaline mucus. The upper part of this lining, in the region called the isthmus, undergoes cyclical changes resembling, but not as marked as, those occurring in the body of the uterus. Numerous small cysts (Naboth's follicles) are found in the cervical mucous membrane. It is from this region that cervical smears are taken in order to detect early changes indicative of cancer.

The tissue layers in the uterus

The uterus is composed of three layers of tissue. On the outside is a serous coat of peritoneum (a membrane exuding a fluid like blood minus its cells and the clotting factor fibrinogen) which partially covers the organ. In front it covers only the body of the cervix; behind it covers the body and the part of the cervix that is above the vagina and is prolonged onto the posterior vaginal wall; from there it is folded back to the rectum. At the side the peritoneal layers stretch from the margin of the uterus to each side wall of the pelvis, forming the two broad ligaments of the uterus.

The middle layer of tissue is muscular (the myometrium) and comprises the greater part of the bulk of the organ. It is very firm and consists of densely packed, unstriped, smooth muscle fibres. Blood vessels, lymphatics, and nerves are also present. The muscle is more or less arranged in three layers of fibres running in different directions. The outermost fibres are arranged longitudinally. Those of the middle layer run in all directions without any orderly arrangement; this layer is the thickest. The innermost fibres are longitudinal and circular in their arrangement.

The innermost layer of tissue in the uterus is the mucous membrane, or endometrium. It lines the uterine cavity as far as the internal os, where it becomes continuous with the lining of the cervical canal. The endometrium contains numerous uterine glands that open into the uterine cavity and that are embedded in the cellular framework or stroma of the endometrium. Numerous blood vessels and lymphatic spaces are also present. The appearances of the endometrium vary considerably at the different stages in reproductive life. It begins to reach full development at puberty and thereafter exhibits dramatic changes during each menstrual cycle. It undergoes further changes before, during, and after pregnancy; during the menopause; and in old age. These changes are for the most part hormonally induced and controlled by the activity of the ovaries.

To understand the nature of the changes in the endometrium during each menstrual cycle it is usual to consider the endometrium to be composed of three layers. They blend imperceptibly but are functionally distinct: the inner two layers are shed at menstruation; the outer or basal layer remains in position against the innermost layer of the myometrium. The three layers are called, respectively, the stratum compactum, the stratum spongiosum, and the stratum basale. The stratum compactum is nearest to the uterine cavity and contains the lining cells and the necks of the uterine glands; its stroma is relatively dense. Superficial blood vessels lie beneath the lining cells. The stratum spongiosum is the large middle layer. It contains the main portions of uterine glands and accompanying blood vessels; the stromal cells are more loosely arranged and larger than in the stratum compactum. The stratum basale lies against the uterine muscle; it contains blood vessels and the bases of the uterine glands. Its stroma remains relatively unaltered during the menstrual cycle.

The menstrual cycle

The menstrual cycle extends over a period of about 28 days (normal range 21–34 days), from the first day of one menstrual flow to the first day of the next. It reflects the cycle of changes occurring in the ovary, which is itself under the control of the anterior lobe of the pituitary. The menstrual cycle is divided into four phases; menstrual, postmenstrual, proliferative, and secretory phases. These states are discussed in more detail below; see *The menstrual cycle.*

The secretory phase reaches its climax about a week after ovulation. Ovulation occurs in midcycle, about 14 days before the onset of the next menstrual flow. The endometrium has been prepared and has been stimulated to a state of active secretion for the reception of a fertilized ovum. The stage has been set for the attachment of the blastocyst, derived from a fertilized ovum, to the endometrium, and for its subsequent embedding. This process is called implantation; its success depends on the satisfactory preparation of the endometrium in both the proliferative and secretory phases. When implantation occurs, a hormone from certain cells of the blastocyst causes prolongation of the corpus luteum and its continued activity. This causes suppression of menstruation and results in the maintenance of the endometrium and its further stimulation by progesterone, with consequent increased thickening. The endometrium of early pregnancy is known as the decidua.

In a cycle in which fertilization of the ovum has not taken place, the secretory phase terminates in menstruation.

The phenomenon of menstruation occurs also in the great apes and in the Old World monkeys, but in New World monkeys, uterine changes are not as dramatic and bleeding is slight. It does not occur in other mammals (except perhaps in modified form in the elephant shrew); and the slight bleeding that may occur in some mammals at about the time of ovulation, caused by high levels of estrogen, is distinct from menstruation. Menstruation may seem a wasteful process, with its loss of tissue and of blood and the iron contained in the blood. If not excessive or abnormally frequent, this loss can be readily made good by a healthy woman. The endometrium needs to be in a certain state of preparedness before implantation can occur. When this stage has been passed, menstruation occurs. Repair then reestablishes an endometrium capable of being stimulated again to the critical stage when implantation can occur.

The uterus is supplied with blood by the two uterine arteries, which are branches of the internal iliac arteries, and by ovarian arteries, which connect with the ends of the uterine arteries and send branches to supply the uterus. The nerves to the uterus include the sympathetic nerve fibres, which produce contraction of uterine muscle and constriction of vessels, and parasympathetic (sacral) fibres, which inhibit muscle activity and cause dilation of blood vessels.

The uterine tubes. The uterine tubes, often called the fallopian tubes, carry ova from the ovaries to the cavity of the uterus. Each opens into the abdominal cavity near an ovary, at one end, and into the uterus, at the other. Three sections of the tubes are distinguished: the funnel-shaped outer end, or infundibulum; the expanded and thin-walled intermediate portion, or ampulla; and the cordlike portion, the isthmus, that opens into the uterus. The infundibulum is fringed with irregular projections called fimbriae. One fimbria, somewhat larger than the others, is usually attached to the ovary. The opening into the abdomen is at the bottom of the infundibulum and is small. Fertilization of the ovum usually occurs in the ampulla of the tube. Normally the fertilized egg is transported to the uterus, but occasionally it may adhere to the tube and start developing as an ectopic or tubal pregnancy. The tube is unable to support this pregnancy, and the conceptus may either be extruded through the abdominal opening or may cause rupture of the tube, with ensuing hemorrhage.

The uterine tube is covered by peritoneum except on its border next to the broad ligament. There are inner circular and outer longitudinal layers of smooth muscle fibres continuous with those of the uterus. The inner lining has numerous longitudinal folds that are covered with ciliated columnar and secretory cells. Muscular contraction, movement of the hairlike cilia, and the passage of the watery secretions all probably assist in the transport of spermatozoa to the ampulla and of a fertilized ovum toward the uterus.

The ovaries. The female gonads, or primary sex organs, corresponding to the testes in a male, are the two ovaries. Each is suspended by a mesentery, or fold of membrane, from the back layer of the broad ligament of the uterus. In a woman who has not been pregnant, the almond-shaped ovary lies in a vertical position against a depression, the ovarian fossa, on the side wall of the lesser pelvis. This relationship is altered during and after pregnancy. Each ovary is somewhat over 2.5 centimetres (one inch) in length, 1.25 centimetres (0.5 inch) across and slightly less in thickness, but the size varies much with age and with the state of activity.

Shape, position, structure, of ovaries

The mesentery of the ovary helps to keep it in position, and within this membrane lie the ovarian artery and vein, lymphatics, and nerve fibres. The uterine tube arches over the ovary and curves downward on its inner or medial surface.

Except at its hilum, the point where blood vessels and the nerve enter the ovary and where the mesentery is attached, the surface of the ovary is smooth and is covered by cubical cells. Beneath the surface, the substance of the ovary is divided into an outer portion, the cortex, and an inner portion, or medulla. The outermost part of the cortex, immediately beneath the outer covering, forms a thin connective tissue zone, the tunica albuginea. The rest of the cortex consists of stromal or framework cells, contained in a fine network of fibres, and also the follicles and corpora lutea.

The ovarian follicles, sometimes called graafian folli-

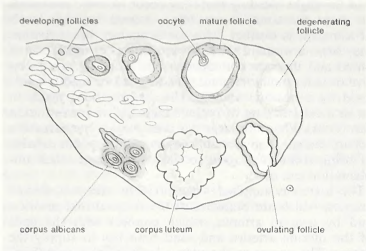


Figure 29: Cross section of the ovary.

cles, are rounded enclosures for the developing ova in the cortex near the surface of the ovary. At birth and in childhood they are present as numerous primary or undeveloped ovarian follicles. Each contains a primitive ovum, or oocyte, and each is covered by a single layer of flattened cells. As many as 700,000 primary follicles are contained in the two ovaries of a young child. Most of these degenerate before or after puberty.

Cyclical development of follicles

During the onset of puberty and thereafter until the menopause (except during pregnancy), there is a cyclic development of one or more follicles each month into a mature follicle. The covering layer of the primary follicle thickens and can be differentiated into an inner membrana granulosa and an outer vascularized theca interna. The cells of these layers (mostly the theca interna) produce estrogenic steroid hormones that exert their effects on the endometrium of the uterus and on other tissues. The maintenance and growth of the follicle to maturity is brought about by a follicle-stimulating hormone (FSH) from the anterior lobe of the pituitary gland. Another hormone, called luteinizing hormone (LH), from the anterior lobe, assists FSH to cause the maturing, now fluid-filled follicle to secrete estrogens. LH also causes a ripe follicle (1.0–1.5 centimetres [0.4–0.6 inch] in diameter) to rupture, causing the liberation of the oocyte into the peritoneal cavity and thence into the uterine tube.

This liberation of the oocyte is called ovulation; it occurs at about the midpoint of the human reproductive cycle, on the 13th–14th day of a 28-day cycle as measured from the first day of the menstrual flow.

Corpus luteum

After ovulation the ruptured follicle collapses because of loss of its follicular fluid and rapidly becomes transformed into a soft, well-vascularized glandular structure known as the corpus luteum. The human corpus luteum ("yellow body") is not yellow but is a creamy gray on section. It develops rapidly, becomes vascularized after about four days, and is fully established by nine days. The gland produces the steroid hormone progesterone and some estrogens. Its activity is both stimulated and maintained by luteinizing hormone. Progesterone stimulates glandular proliferation and secretion in an endometrium primed by estrogens.

While the ovarian follicle matures, the primary oocyte divides into a secondary oocyte and a small rudimentary ovum called the first polar body. This occurs at about the time when the follicle develops its cavity; the oocyte also gains a translucent, acellular covering or envelope, the zona pellucida. The secondary oocyte is liberated at ovulation: it is 120–140 microns in diameter and is surrounded by the zona pellucida and a few layers of cells known as the corona radiata. The final maturation of the oocyte, with the formation of the rudimentary ovum called the second polar body, occurs at the time of fertilization.

If fertilization does not occur, then the life of the corpus luteum is limited to about 14 days. Degeneration of the gland starts toward the end of this period and menstruation occurs. The corpus luteum shrinks, fibrous tissue is formed, and it is converted into a scarlike structure, called a corpus albicans, which persists for a few months. Should fertilization occur and be followed by implantation of the blastocyst, hormones (particularly chorionic gonadotrophin) are produced by cells of the blastocyst to prolong the life of the corpus luteum. It persists in an active state for at least the first two months of pregnancy until the placental tissue has taken over its endocrine (hormone-producing) function. The corpus luteum of pregnancy then also retrogresses and is becoming a fibrous scar by the time of parturition.

The ovarian arteries arise from the front of the aorta, in a manner similar to the testicular arteries, but at the brim of the lesser pelvis they turn down into the pelvic cavity. Passing in the suspensory ligament of the ovary, each artery reaches the broad ligament below the uterine tube and then passes into the mesovarium to divide into branches distributed to the ovary. One branch continues in the broad ligament to anastomose with the uterine artery. The ovarian veins emerge from each ovary as a network that eventually becomes a single vein; the terminations are similar to those of the testicular veins. The nerves are derived from the ovarian nerve network on the ovarian artery. (R.J.Ha.)

THE MENSTRUAL CYCLE

Menstruation is the periodic discharge from the vagina of blood, secretions, and disintegrating mucous membrane that had lined the uterus.

The biological significance of the process can best be explained by reference to the reproductive function in other mammals. In a number of species of wild sheep, for example, there is only one breeding season in the year; during this season a cycle of changes takes place in the reproductive organs, characterized by ripening and release of ova from the ovaries, increased blood supply to the genital tract, growth of the uterus, and proliferation of its lining. There is a discharge of blood and mucus from the uterus and vagina, and this is the time when coition may take place. Pregnancy normally follows, but if the ewe is not served by the ram the changes retrogress until the next breeding season. This cycle of changes is termed the estrous cycle.

The estrous cycle

In many domesticated sheep there is more than one estrous cycle in the breeding season. If the ewe does not become pregnant in the first cycle there is a short resting phase; then ovulation is repeated and another cycle of activity of the reproductive system takes place. After each breeding period, with its succession of estrous cycles, there is a relatively long resting phase.

In most female primates, including women, there is no resting phase; an unbroken series of estrous cycles occurs throughout the year, and pregnancy can occur in any one of them.

In some animals a variety of external stimuli act through the central nervous system on the hypothalamic region of the brain. The hypothalamus controls the release from the pituitary gland of hormones that induce ripening of ovarian follicles—ova and the cellular structures that enclose them. These pituitary hormones, called gonadotropic hormones, are carried to the ovaries by way of the bloodstream. In primates the hypothalamic mechanism normally is independent of external stimuli, and regular discharge of ova into the tubes leading to the uterus occurs even in the absence of coitus. Under the influence of the pituitary gonadotropic hormones, the ovary produces other hormones, which cause growth and increased vascularity of the uterus and vagina. These hormones are estrogens—chiefly 17 beta-estradiol—and progesterone. It is as though the ovary prepares the uterus for the reception of the ovum that is released in the particular cycle.

Phases of the menstrual cycle. The normal human menstrual cycle is 28 days, but no woman is always precisely regular, and cycles as short as 21 days or as long as 35 days are not abnormal. It is customary to call the first day of the menstrual period the first day of the cycle, although menstruation is the end rather than the beginning of a process. On this basis the cycle is described as starting with about five days of menstruation, followed by a proliferative phase that lasts to about the 14th day, and then a secretory phase that lasts until the next menstruation. The external manifestation of menstruation depends upon cyclical change in the lining of the body of the uterus. The lining, called endometrium, consists of tubular glands that

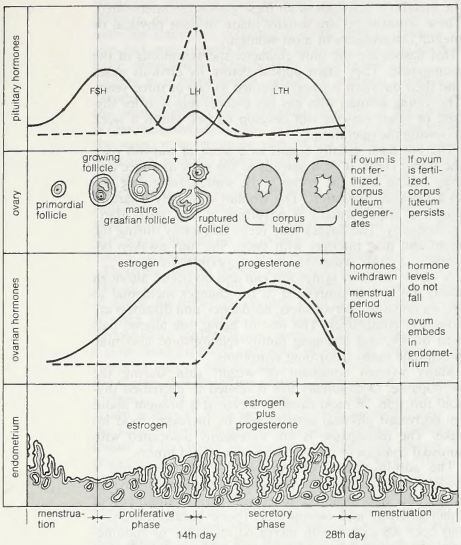


Figure 30: Menstrual cycle.
From Ciba Collection of Medical Illustrations, vol. 4. J. & A. Churchill, Ltd.

open into the uterine cavity. The glands lie in a vascular framework, or stroma, and are separated by it.

The proliferative phase

At the end of menstruation, just at the beginning of the proliferative phase, the endometrium is thin, with short, straight glands, and the ovary is quiescent. Under the influence of the gonadotropic hormones from the pituitary gland an ovarian follicle (occasionally more than one) ripens in one of the ovaries. This ovarian follicle contains the ovum, which is a cell about 0.14 millimetre (0.006 inch) in diameter, surrounded by a group of smaller cells, called granulosa cells. The granulosa cells multiply, with the ovum situated in the wall of the rounded structure that they form, and secrete an estrogenic hormone, estradiol (see BIOCHEMICAL COMPONENTS OF ORGANISMS: *Hormones*). This hormone causes proliferative changes in the endometrium, so that the glands become taller and the whole endometrium becomes thicker and more vascular.

At about mid-cycle ovulation occurs: The ovum is discharged out of the follicle and from the surface of the ovary, to be received into the fallopian tube, down which it is carried to the uterus. After ovulation the granulosa cells lining the follicle from which the ovum has been extruded accumulate yellow lipid and are therefore called lutein cells, from the Latin word *luteus*, "saffron-yellow." The altered follicle is called corpus luteum. The corpus luteum continues to secrete estrogens but now also secretes progesterone; this additional hormone induces the secretory phase in the endometrium. The endometrial glands

The secretory phase

are distended with secretion and become very tortuous, while the stromal cells are swollen. The appearance of the endometrium at the end of the menstrual cycle is indistinguishable from that of early pregnancy, and this endometrial change is a preparation for the reception of the ovum. If it is fertilized, the ovum liberated at mid-cycle reaches the uterine cavity at a time when the endometrium is in the secretory phase, and the ovum embeds itself in the endometrium and starts its growth. If the ovum is not fertilized the endometrium breaks down and menstruation occurs. Menstruation has therefore been described as the outward evidence of the abortive close of one cycle and the hopeful commencement of the next.

When the ovum dies, the corpus luteum degenerates and ceases to produce hormones. On the withdrawal of estrogens and progesterone there is sudden spasm of the endometrial blood vessels, and all but the basal layer of the endometrium dies. The disintegrating endometrium is shed, together with some blood. The endometrium contains plasmin, an enzyme that dissolves blood clots, so that the menstrual discharge is normally fluid. The total blood loss does not ordinarily exceed 50 millilitres (1.69 U.S. fluid ounces).

After menstruation the endometrium regenerates from the residual basal layer during the proliferative phase of the next cycle.

Hormonal control of menstrual cycle. The ovarian hormones circulate in the blood and are excreted in modified forms in the urine. Estimation of the urinary output by chemical methods gives an indication of the blood levels and of the total production of these substances. There are several natural estrogens, and numerous synthetic modifications of these and of progesterone have been devised; many are active when taken by mouth and are used for treatment of hormonal disorders and as oral contraceptives.

The cyclic events in the ovary that have already been mentioned depend on gonadotropic hormones secreted by the anterior lobe of the pituitary gland; this gland is situated in a small recess at the base of the skull. There are two, and possibly three, gonadotropic hormones: follicle-stimulating hormone (FSH), luteinizing hormone (LH), and, possibly, luteotropic hormone (LTH).

FSH is secreted in greatest amount in the first half of the menstrual cycle, and LH has its peak of secretion at mid-cycle. It is believed that the sequential action of FSH and LH causes ripening of the follicle and ovulation. In some animals LTH is necessary for maintenance of the corpus luteum, but in women under treatment for infertility ovulation has been successfully induced with FSH and LH alone. Multiple births, as the result of multiple ovulation, have occurred after excessive doses of FSH have been given.

Inhibitory effect of estrogens and progesterone

The pituitary gland stimulates the ovary to produce estrogens and progesterone, but there is a "negative feedback" by which the estrogens inhibit the output of FSH from the pituitary gland (and probably stimulate the output of LH). In addition, progesterone is believed to inhibit the further output of LH. In this process, in which the pituitary first stimulates the ovary, and the ovary then inhibits the pituitary, the basic rhythm is under the control of the hypothalamus; nevertheless, ovulation can be inhibited by oral contraceptives, which contain estrogens and progestogens—modifications of progesterone.

The anterior lobe of the pituitary gland is connected by its stalk to the hypothalamic region of the brain. The anterior lobe secretes many important hormones, including those that control the activity of the adrenal and thyroid glands, the growth hormone, and the gonadotropic hormones. From the hypothalamus substances are carried in the veins in the pituitary stalk that cause release of hormones from the pituitary, including FSH and LH, but also a factor that inhibits release of LTH. The higher brain centres no doubt affect the hypothalamic function; this explains the temporary disturbances of menstruation that may follow emotional stress.

Ovulation and the fertile phase. Ovulation occurs at about the midpoint of each normal cycle, and the ovum is probably capable of fertilization for only about two days after this. In the majority of women the time of ovulation is fairly constant. In women with cycles of irregular length the date of ovulation is uncertain; in these women the long menstrual cycles are usually due to prolongation of the proliferative phase; the secretory phase tends to remain normal in length. In some animals, ovulation only follows coitus; this mechanism has been used to explain cases in which human pregnancy has apparently followed coitus early or late in the menstrual cycle, but there is no definite evidence for such a mechanism in women.

The rhythm method of contraception is based on the fact that ovulation normally occurs at mid-cycle, but the date of ovulation may vary unexpectedly even in women whose menstrual cycles were previously regular.

The menarche. The first menstruation, or menarche,

usually occurs between 11 and 13 years of age, but in a few otherwise normal children menstruation may begin sooner or may be delayed. If the menstrual periods have not started by the age of 16 gynecological investigation is indicated. The menarche is preceded by other signs of estrogenic activity, such as enlargement of the breasts and the uterus and growth of pubic hair. The ovarian response to gonadotropic hormones may be erratic at first, so that irregular or heavy bleeding sometimes occurs, but this irregularity nearly always disappears spontaneously.

Normal menstruation. Each menstrual period lasts for about five days, but the duration and amount of the flow vary considerably even in perfect health.

Discomfort in normal menstruation

In some women there may be premonitory symptoms such as pelvic discomfort, soreness of the breasts (because of the response of these organs to estrogens), and emotional tension. Ovarian hormones cause retention of sodium and water in the tissue fluids; premenstrual tension, sometimes called Premenstrual Syndrome, may be partly due to this and in some cases can be relieved by diuretics, drugs that increase the production of urine. When the menstrual flow starts the uterus contracts to expel the blood and disintegrating endometrium. These contractions may be painful, especially in young women who have never been pregnant. Menstrual discomforts such as those that have been mentioned vary greatly in degree from woman to woman and from time to time but ordinarily do not interfere with normal activities. (Menstrual disorders are discussed in detail below; see *Reproductive system diseases.*) (S.G.C./Ed.)

MENOPAUSE

"Menopause" (female climacteric) is the final cessation of menstruation and therefore the end of a woman's reproductive life. The popular term "change of life" is neither descriptive nor accurate, for it tends to indicate a physical, mental, and sexual deterioration, whereas deterioration does not occur.

In most women, menopause begins between the ages of 45 and 55 years. Although the average age for onset is about 50 years, menopause may begin as early as age 40 or be delayed to the late 50s. Although the age of onset is probably determined by the hereditary background of the individual, good nutrition and health habits tend to postpone onset.

A premature menopause—*i.e.,* one that takes place spontaneously before the age of 40—occurs in about 8 percent of women. An artificial menopause may be induced by removing the ovaries by surgery or by destroying them with X-rays or radium.

The natural life of the ovaries is about 35 years. The menopause is brought on by a progressive decline in ovarian function. This decline is a normal result of aging and is accelerated as the menopause approaches. During the reproductive years follicles in the ovaries mature and release their ova periodically under hypothalamic-pituitary stimulation. In the years immediately preceding the menopause, however, first some follicles and later all follicles fail to rupture and release their ova. The failure to ovulate results in a disturbed menstrual pattern. A woman may miss a period or two and suspect pregnancy. A medical examination, however, will establish the proper diagnosis. The continued decline in ovarian activity may provoke prolonged intervals between periods or irregular bleeding episodes. The length of the periods may vary, and the flow may become either more scant or profuse. In a fortunate small minority the periods cease abruptly.

Estrogen level

As the ovaries decline in function, they produce smaller and smaller amounts of the hormone estrogen; this decline in estrogen initiates subtle rearrangements in the hormonal activity of the glands that control the reproductive function. The decrease in output of estrogen disturbs the neurovascular mechanism of the hypothalamus and probably initiates the vasomotor changes that may provoke the characteristic "hot flashes" of the menopause. The metabolism of the pituitary gland is altered and increasing amounts of follicle stimulating hormone (FSH) appear in the blood and urine. Rearrangements in the hormonal activity of the adrenal and thyroid glands also take place, for the metabolic activities of all these glands are interrelated. These adjustments are usually made without physical or mental disturbances in most women.

Hot flashes are the only characteristic symptoms of the menopause. They often appear before its obvious onset and their duration is usually limited to two or three years. The young woman who has her ovaries removed for disease or other reasons will develop flashes within a week following the operation.

The hot flash usually begins as a sense of warmth over the upper chest. It then spreads to the neck and face and may extend over the entire body, sometimes giving rise to a prickling sensation. The woman is acutely aware of blushing, which usually is disturbing to her, particularly in company. The flashes may recur frequently during the night and may interfere with sleep. She may awaken because of a chilly sensation and may perspire freely.

A variety of other symptoms can and do occur, although many are entirely unrelated to the changes incidental to the menopause. Nervousness, headaches, and dizziness are common complaints. The fear of aging, the altered pattern of life, and changing family relationships also may precipitate many disturbing symptoms.

Many women complain of weight gain during the menopause. Occasionally this is related to decreased thyroid function. In most cases, however, it is brought about by decreased physical activity and by increased food intake. The menopause is not necessarily associated with unusual changes in physical appearance and fitness.

The administration of estrogenic hormones—once a widely approved remedy for relief of menopausal symptoms and also believed to retard the development of atherosclerosis and osteoporosis (bone decalcification)—has been associated with increased incidence of endometrial cancer and is being reevaluated. (M.E.D./Ed.)

Diseases of the human reproductive system

The human reproductive system may be affected by abnormal hormone production coming from the ovaries or the testes or from other endocrine glands, such as the pituitary, thyroid, and adrenals. Reproductive-system diseases can also be caused by genetic abnormalities, congenital anomalies (abnormalities), infections, tumours, and disorders of unknown cause.

The main divisions of this section are concerned with (1) genetic and congenital abnormalities; (2) functional genital disorders; (3) infections; (4) structural changes of unknown cause; and (5) tumours. Diseases and disorders of pregnancy and delivery are covered below; see *Abnormal changes in pregnancy* and *Accidents during labour.* Endocrine disorders affecting reproductive organs and functions and discussed in the article ENDOCRINE SYSTEMS.

GENETIC AND CONGENITAL ABNORMALITIES

In the male. Congenital anomalies of the prostate and seminal vesicles are rare; they consist of absence, hypoplasia (underdevelopment), or the presence of fluid- or semisolid-filled sacs, called cysts. Cysts of the prostatic utricle (the uterine remnant found in the male) are often found in association with advanced stages of hypospadias (a defect in the urethra, see below) and pseudohermaphroditism (in which sex glands are present but bodily appearance is ambiguous as to sex; *i.e.,* the secondary sexual characteristics are underdeveloped). Cysts may also cause urinary obstructive symptoms through local pressure on the bladder neck.

Abnormalities of the penis

Severe anomalies of the penis are rare and are generally associated with urinary or other systemic defects that are incompatible with life. Anomalies are those of absence, transposition, torsion (twisting), and reduplication of the penis. An abnormally large penis frequently is present in boys affected by precocious puberty, in congenital imbeciles, in dwarfs, in men with overactive pituitaries, and in persons affected by adrenal tumours. A small penis is seen in infantilism and in underdevelopment of the genitals, or undersecretion of the pituitary or pineal gland, and failure of development of the corpora cavernosa.

The only anomaly of the foreskin of grave concern is

congenital phimosis, characterized by a contracture of the foreskin, or prepuce, sufficient to prevent its retraction over the glans; the preputial opening may be pinhole in size and may impede the flow of urine. The condition is easily remedied by circumcision, a permanent cure.

There is a considerable variety of urethral anomalies. Stenosis (contracture) of the external opening (meatus) is the most common, but congenital stricture of the urethra occasionally occurs at other points. Valves (or flaps) across the anterior or posterior part of the urethra may cause congenital urethral obstruction in boys. Posterior urethral valves are more common than anterior valves and consist of deep folds of mucous membrane, often paper-thin and usually attached at one end to the verumontanum, a small prominence in the back wall of that part of the urethra that is surrounded by the prostate gland. If too tight, the valves may obstruct the urethra and destroy the kidneys.

There are various defects associated with incomplete closure of the urethra. One of the commonest is hypospadias, in which the underside (ventral side) of the urethral canal is open for a distance at its outer end. Frequently the hypospadiac meatus is narrowed, and the penis also has a downward (ventral) curvature beyond the meatus. The posterior part of the urethra is never involved; therefore, the muscle that closes the urethra, the sphincter, functions normally, and urinary control exists. Although the condition occurs in both sexes, it is seen predominantly in the male. There is a high incidence of partial or complete failure of the testes to develop, cryptorchism (failure of the testes to descend into the scrotum), and small external and internal genitalia; variable male–female admixtures may be associated with this deficiency. Epispadias, an opening in the upper (dorsal) side of the penis, is considerably less common than hypospadias. Dorsal curvature may also be present, but the disabling aspect is that the defect usually extends through the urinary sphincter and causes urinary incontinence. Other less common urethral anomalies include complete absence of the urethra, double urethra, urethra fistula (an opening in the urethra), urethrorectal fistula (an opening between the urethra and the rectum), and urethral diverticulum (a pouch in the wall of the urethra). Most of the above conditions are correctable by surgery.

Absence or excessive number of testes

Anorchism (absence of one or both testes) is rare; it may be associated with the absence of various other structures of the spermatic tract. Generally, if one testis is absent, the other is found to be within the abdomen rather than in the scrotum. Congenitally small testes may be a primary disorder or may occur because of underactivity of the pituitary. In both disorders, there is a lack of development of secondary sexual characteristics and some deficiency in libido and potency. Supernumerary testicles are extremely rare; when present, one or more of the supernumerary testicles usually shows some disorder such as torsion of the spermatic cord. Synorchism, the fusion of the two testicles into one mass, may occur within the scrotum or in the abdomen. Cryptorchism is the term applied to all forms of imperfectly descended testes, the commonest anomaly of the spermatic tract. The condition is often bilateral, and in the unilateral cases there is no preponderance between the left or right side. Hormonal treatment may be useful in correcting the condition, but usually surgery is necessary for correction.

In the female. The female external genitalia are less complex than those of the male but have anomalies that can at times severely interfere with the functioning of the female urogenital tract. The clitoris, an erectile structure that corresponds to the penis, except that it does not contain the urethra, may be absent but in other cases may be enlarged on either a congenital or a hormonal basis. Fusion of the labia minora (small folds of skin covering the clitoris, the urethral opening, and the opening of the vagina) is a midline "sealing together" of the labia minora; usually a minute unfused area is left just below the clitoris, through which the child urinates and later menstruates. The chief difficulty with this anomaly is concerned with obstruction to the flow of urine and associated urinary-tract infection. An imperforate hymen (the membrane closing off the opening of the vagina) causes distension of the uterus and vagina with fluid other than blood before puberty and with blood after puberty (the two conditions are called hydrometrocolpos and hematocolpometra, respectively). The distended vagina compresses the urethra enough to interfere with urination and commonly may even cause complete retention of urine in the bladder and distension of the entire upper urinary tract. Fusion of the urethra and the hymen is characterized by a dense hymenal ring and a stenosed urethral opening. The consequent urinary obstruction commonly results in persistent urinary infection. Most of the conditions are readily remedied by surgery.

Anomalies of vagina and uterus

Anomalies of the vagina and uterus consist of complete absence, incomplete development, and duplication. The female urethra may have a congenitally narrow opening, or meatus; may be distended; may have an abnormal pouch, or diverticulum, in its wall; or may open abnormally into the vagina. Hypospadias may occur in the female but is far less common than in the male. Epispadias is also present in the female. Reconstructive surgery is the only method of treatment. One of the rarest and most severe of the urogenital-tract anomalies, called persistent cloaca, consists in congenital intercommunication between the rectum and the bladder and vagina or between the rectum and the urethra and vagina.

Intersexuality. Intersexuality (having both male and female characteristics) may be noticeable at birth or may become apparent after puberty. Intersexuality noticeable at birth may be classified as female or male pseudohermaphroditism or true hermaphroditism. Female pseudohermaphroditism, or female intersex, may be of adrenal or nonadrenal type. The adrenal type develops because of an inborn error in the metabolism of the adrenal hormone cortisol that leads to an increased secretion of corticotropin (ACTH) and consequent excessive secretion of androgens (male sex hormones). The newborn female with this condition is a chromosomal female and resembles a normal female, but an excess male hormone has a masculinizing effect on the external genitalia; the vagina tends to be connected to the urethra; the clitoris is enlarged, as are the labia (the labia majora are prominent folds of skin, corresponding to the scrotum in the male). Effective treatment can be achieved by administration of adrenal hormones (*e.g.*, cortisone, hydrocortisone), which suppress the pituitary so that its stimulus to adrenal production of androgenic hormones is minimized. The nonadrenal type of intersex is seen in infants whose mothers have been administered synthetic androgens or progestational compounds (substances that stimulate changes in the uterus that further the implantation and growth of the fertilized ovum) during pregnancy. Rarely, the condition is associated with the presence in the mother of a masculinizing tumour of the ovary or the adrenal gland. The newborn infant is a female with varying degrees of ambiguous genitalia; no treatment is necessary, and normal female development occurs at puberty.

Male pseudohermaphrodites are males with varying deficiencies of internal and external virilization. Most commonly, the male intersex has a markedly hypospadiac penis, undescended testes, a cleft scrotum, and an enlarged prostatic utricle; a complete uterus and fallopian tubes may be found, with the vagina opening into the posterior wall of the urethra. (Such persons are pseudohermaphrodites in that they do not have ovaries.)

True hermaphroditism

True hermaphrodites have recognizable ovarian and testicular tissue. A uterus is always present, but the internal genitalia otherwise vary greatly, often including both male and female structures. The external genitalia are usually ambiguous, but in 75 percent of the reported cases the children have been raised as males. At puberty, over 80 percent of them develop enlarged breasts, and approximately half menstruate. Most hermaphrodites are chromatin positive—that is, they have, within and near the periphery of the nuclei of their cells, a substance, chromatin, that is normally found in the cells of females but not in those of males—and over half have a characteristically female set of chromosomes in their peripheral blood cells.

Surgical and hormonal therapy directed to producing ei-

ther a male or a female configuration of the body is based on the existing physical and psychological findings.

Inter-sexuality syndromes associated with puberty

Klinefelter's, Turner's, and testicular feminizing syndromes are intersexuality syndromes that become apparent prior to or after puberty. Klinefelter's syndrome is a genetic disorder of phenotypic males (persons who have a male body configuration) who do not produce sperm, have small testes, and have varying degrees of eunuchoidism. Patients with this syndrome have various associated medical problems, such as chronic disease of the lungs, varicose veins, thrombophlebitis (inflammation of the blood vessels), obesity, diabetes mellitus, hyperlipemia (abnormally high blood levels of fats), and enlarged breasts at the time of puberty. Mental retardation and antisocial behaviour are also associated with this syndrome.

Turner's syndrome is a disorder of phenotypic females—persons who are female in physical configuration. Characteristically, such persons are short, do not menstruate, and show estrogen (a female sex hormone) deficiency; there is a distinctive cluster of congenital anomalies.

The disorder known as the testicular feminizing syndrome is inherited. Affected persons seem to be of normally developed females but have a chromosomal sex that is that of the normal male. The gonads are well-developed testes, and evidence indicates that there is a normal production of testosterone (male hormone), but there is cellular resistance to the action of this hormone, and therefore the affected person becomes female in appearance. Because these gonads are apt to form malignant tumours, they are usually removed surgically. Female sexual characteristics are then maintained by the administration of estrogenic hormones.

FUNCTIONAL GENITAL DISORDERS

Affecting both male and female systems. *Delayed puberty.* The term delayed puberty may be a misnomer, because puberty delayed beyond age 19 is in fact a permanent failure of sexual development because of an abnormally low secretion by the pituitary gland of gonadotropic hormone, the hormone that stimulates growth and activity of the sex glands; this condition is called hypogonadotropic eunuchoidism. The term delayed puberty is usually applied to boys who develop more slowly than the average but who still eventually undergo full sexual development. Only in retrospect—*i.e.*, after the affected person reaches the age of 20—can one clearly differentiate these cases from the classic or incomplete forms of hypogonadotropic eunuchoidism. If there are social and psychological problems related to the sexual underdevelopment, therapy may consist of a course of chorionic gonadotropin, a hormone produced by the placenta and secured from the urine of pregnant women. If puberty is merely delayed, it will usually progress normally after this treatment. If it fails to progress, the patient does not have delayed puberty but rather has hypogonadotropic eunuchoidism.

Precocious puberty. In healthy girls living in a temperate climate, the earliest sign of puberty occurs at a mean age of 10.6 years (standard deviation of 1.2 years), whereas, in boys, testicular growth begins at a mean age of 11.8, with a standard deviation of one year. The average age of menstruation is 13.5 years (range, 9–17 years). What is called true precocious puberty is a condition in which normal pituitary–gonadal function is activated at an abnormally early age. It is always isosexual with the sex gonads (*i.e.*, it is always in keeping with the sex of the gonads) and with development of the secondary sexual characteristics and production of spermatozoa or ova. Pseudoprecocious puberty includes development of secondary sexual characteristics but not production of spermatozoa or ova and may be isosexual or may be heterosexual (*i.e.*, it may involve virilization in the female or feminization in the male).

Pseudo-precocious puberty

The causes of true precocious puberty are several—including brain lesions and hypothyroidism (abnormally low secretion by the thyroid glands); the largest proportion of cases are of unknown cause. Precocious pseudopuberty in females may be caused by ovarian tumours, a cyst of the ovary, a tumour of the adrenal cortex (outer substance of the adrenal gland), or congenital overdevelopment of the adrenal gland. In males, the causes include congenital overdevelopment of the adrenal glands, tumour of the adrenal cortex, tumour involving the Leydig cells of the testes, and teratoma (a tumour containing numerous types of tissue; in these circumstances it includes adrenal-cortical tissue).

Infertility. At least 10 percent of marriages are barren, and deficiencies of sperm production in the male are the causal factor in 40 percent of these. The common causes of male infertility are deficiencies in maturation of germ cells (sperm); orchitis (mumps), with destruction of the testes; obstruction of the passageways for sperm; abnormally low thyroid or high adrenal secretion; varicocele (enlargement of the veins of the spermatic cord); or formation of antibodies to sperm by the male or the female. The most important steps in the evaluation of male infertility are examination of the semen and of a specimen of the tissue of the testes. Evaluation also includes chromatin analysis and observation of thyroid, adrenal, and pituitary function. The results of treatment of infertility in the male are usually unsatisfactory, except when a varicocele or obstruction in the sperm passageways is the cause, in which case surgical correction may be beneficial.

Causes of infertility in females

Infertility in the female is related to the faulty production of ova or to interferences with their union with spermatozoa. Vaginal causes are usually uncommon, but obstruction may be due to an unruptured hymen or may be functional and arise from enlargement and contraction of the levator ani muscles (these muscles form a supporting sheet under the pelvic cavity, with openings for structures such as the anus and the vagina). Abnormalities of the cervix are among the most important causes obstructing the passage of sperm. (The sperm normally enter the uterus through the cervix and, from the uterus, move into a uterine or fallopian tube, where fertilization of an ovum takes place.) During the few days prior to ovulation—release of an ovum from the ovary—the glands within the cervix normally secrete a thin, watery mucus that is beneficial to sperm survival and migration. Various factors, such as infection or estrogen deficiency, may decrease the quality of the mucus.

Uterine anomalies such as a bicornuate (double) uterus may play a role in infertility. Total or partial blocking of the uterine tubes can result from inflammation due to infection (*e.g.*, gonorrhea) or from endometriosis, a condition involving the presence of tissue resembling that which lines the uterus elsewhere in the pelvic cavity. Thyroid, pituitary, adrenal, or ovarian disease may interfere with ovulation, as may the presence of large numbers of cysts in the ovaries (the condition known as polycystic ovaries).

Finally, emotional factors may play a role in causing infertility. Treatment consists of the use of various hormones, surgical correction of tubal blockage, and psychotherapy. With the advent of new hormone preparations, the results in achieving pregnancy have been vastly improved.

(N.A.Ro./J.K.La.)

Affecting female system. Abnormalities of menstrual function include absence of menstruation, called amenorrhea; excessive blood loss at each period, known as menorrhagia; irregular cycles, or metrorrhagia; and painful menstruation, or dysmenorrhea. In addition, there may be premenstrual tension, and in a few women, pain at the time of ovulation. As a sexual and reproductive function, menstruation has deep emotional significance, but the popular belief that regular menstrual flow is necessary for health is unfounded. The belief arises in part from the fact that in any severe illness, and during emotional disturbances and psychiatric illness, the cycles may be disturbed. This is particularly true of diseases of the endocrine system, not only of the pituitary gland but of other glands, such as the thyroid and the adrenal, as well.

Menstrual abnormality may also be a symptom of local disease of the pelvic organs. Irregular bleeding, bleeding after intercourse, and bleeding at or after the menopause may be early signs of uterine malignant disease.

Ovulatory pain, premenstrual tension, dysmenorrhea

A few women have transient abdominal discomfort at the time of ovulation because of slight bleeding from the follicle into the peritoneal cavity. Oral administration of estrogens and progestogens will remedy the condition by suppression of ovulation, but the discomfort seldom recurs

in every cycle or is severe enough to merit such treatment.

Premenstrual tension has already been mentioned. When it is due to fluid retention there is increase in weight before menstruation, and diuretics such as chlorothiazide give relief. In most instances emotional tension is the main complaint, and relief is attained by the use of mild sedatives or tranquillizers. Objective studies of women who do fine work show a reduction in accuracy and in concentration at this time, and outbursts of emotion may occur. The claim that relief is obtained by administration of progestogens is not generally accepted.

Painful menstruation in young women who have not borne children is a common complaint; the pain is sometimes so severe as to interfere with daily occupations. There is, as a rule, no endocrine or anatomical abnormality, and fertility is normal. The pain occurs during the active menstrual flow and is due to colicky contractions of the uterus. Minor degrees of pain are adequately controlled with analgesic drugs. In the past severe cases were often treated by surgical dilation of the cervix, and occasionally by severing the nerves that carry the pain impulses from the uterus, but now hormonal treatment is more usual. This type of dysmenorrhea usually does not persist after birth of a child.

Another type of dysmenorrhea results from pelvic disease such as inflammation of the tubes and ovaries, or from endometriosis. In endometriosis, deposits of endometrium, which undergo cyclic response to the ovarian hormones, are found in the ovaries and in other sites outside their normal location; these deposits form blood-filled cysts, and pain and excessive bleeding result. In painful menstruation secondary to pelvic disease there is, before menstruation, pain associated with a feeling of congestion, and the menstrual bleeding is often excessive. Surgical treatment may be required, although the symptoms of endometriosis can be suppressed by use of progestogens.

Excessive or irregular menstrual bleeding; intermenstrual bleeding

Excessive or irregular menstrual bleeding may be due to an imbalance of the pituitary and ovarian hormones, but it may also be the result of local disease of the pelvic organs. This local disease may be inflammation due to infection; it may be a benign tumour such as a fibromyoma, or fibroid; it may be a polyp, or projecting mass of endometrium; or it may be a cancer, especially after age 35. Some types of local pelvic disease may require removal of the uterus or treatment by irradiation but polyps and some fibroids can be removed without loss of the uterus.

Irregular or excessive bleeding often results from emotional disturbance; this type of abnormality tends to disappear spontaneously.

As the menopause approaches, extremely heavy bleeding may occur, causing anemia, tiredness, and ill health. Menorrhagia in this instance is due to overdevelopment of the endometrium as a result of excessive or unbalanced action of estrogens. Younger or childless women can be treated with progestogens; for others removal of the uterus may be necessary.

Bleeding between periods or after intercourse is frequently due to some abnormality of the cervix; in women who have borne children the possibility of cancer must be borne in mind. Such bleeding may also come from a simple polyp on the cervix or a cervical erosion. The latter is a rather common condition in which the glandular lining of the canal of the cervix extends out onto its surface. Treatment is often unnecessary, but erosions are easily dealt with by cauterization. Polyps require removal.

Irregular bleeding also may occur during pregnancy when there is danger of abortion; if any menstrual periods have been missed this possibility must be considered.

Amenorrhea

Amenorrhea, or absence of menstruation, is normal during pregnancy and for a variable time after delivery. If the mother is breast-feeding her baby, as much as six months may pass before return of menstruation; earlier return of menstruation is not abnormal and is to be expected if the mother is not producing milk. Pregnancy is the commonest cause of amenorrhea during the reproductive years.

The term primary amenorrhea refers to the absence of menstruation in a woman who has never previously menstruated. In rare cases, primary amenorrhea is due to gonadal dysgenesis, the failure of the ovaries to develop normally, and may be associated with chromosomal abnormalities. Instead of the normal female complement of 46 chromosomes in each cell, including two X-chromosomes, a patient may have only one X-chromosome, or even a male pattern of an X- and a Y-chromosome. In such persons the uterus and tubes often are absent, although the general physique may be female.

Even with normal ovaries, absence of the uterus occasionally occurs. A less rare abnormality is vaginal atresia, or closure, an obstruction of the vagina by a membrane just above the level of the hymen; menstruation occurs, but the discharge cannot escape and distends the vagina. This condition, called false amenorrhea or cryptomenorrhea, is easily corrected by an incision in the membrane.

Cessation of periods after menstruation has been established, but before the normal time for the menopause, is usually the result of some general illness, some emotional disturbance, or a psychiatric illness. It may also be due to disease of the endocrine system, not only of the pituitary gland but of other endocrine glands as well. Secondary amenorrhea results if the ovaries are removed or are irradiated but is unlikely to be caused by ovarian disease, as both ovaries would have to be totally destroyed to stop all function. There is a functional disorder of the ovaries in which production of estrogens is disturbed. Symptoms of this disorder include abnormal growth of facial hair because of abnormal androgenic—that is, masculinizing—activity. An ovarian tumour, arrhenoblastoma, that secretes androgenic hormone, is another extremely rare cause of amenorrhea and abnormal growth of hair.

Most cases of secondary amenorrhea are temporary, and spontaneous improvement is to be expected, especially when the cause is some general illness or emotional disturbance. Treatment of amenorrhea serves no purpose unless pregnancy is ultimately desired and is possible. The feasibility of treatment with hormones is determined by a general medical examination and a complete pelvic examination, including various hormonal assays and inspection of a specimen of endometrium.

Menstruation without discharge of an ovum

In some women who complain of infertility but apparently have normal periods, it has been found that the cycles are anovular. Estrogens are produced in each cycle in amounts sufficient to cause endometrial proliferation, but ovulation does not occur and the corpus luteum is not formed, so that there is no secretion of progesterone. The endometrium breaks down and bleeds in each cycle as the estrogens are withdrawn. Cycles of this type occur in women who are using oral contraceptives.

A simple way to determine whether ovulation is occurring is to record the woman's early morning temperature daily. In a normal cycle the temperature is about 0.5° C lower in the first than in the second half of the cycle, and the rise in temperature occurs at the time of ovulation. No rise occurs in anovular cycles. Another test is to remove and inspect a fragment of endometrium in the late part of the cycle; if microscopic examination shows that normal secretory changes are present ovulation must have taken place (see above *Infertility*). (S.G.C./Ed.)

Affecting the male system. *Impotence.* Impotence is inability of the male to have satisfactory sexual intercourse and varies in form from the inability to gain an erection to weak erections, premature ejaculation, or loss of normal sensation with ejaculation. Almost all of these complaints are psychogenic in origin, but impotence may be caused by subnormal functioning of the testes, by arteriosclerosis (hardening of the arteries), diabetes mellitus (a metabolic disease in which there is inadequate secretion or utilization of insulin), or by some disease of the nervous system. Certain medications prescribed for the treatment of such diseases as peptic ulcer, hypertension, or psychiatric illnesses may adversely affect sexual ability. Therapy, usually limited in its success, includes administration of sex hormones and psychotherapy.

Priapism. Priapism is prolonged penile erection that is painful and unassociated with sexual stimulation. The blood in the spaces of the corpora cavernosa becomes sludgelike and may remain for hours or even for days. About 25 percent of the cases are associated with leukemia (a disease of the blood-forming tissues that results in

extremely high numbers of white blood cells), sickle-cell anemia (an inherited disease in which red blood cells are abnormal in shape and function and the hemoglobin is of a particular type), metastatic carcinoma (cancerous development at a distance from the primary site), and diseases of the nervous system, but in the majority of cases the causation is not clear. There have been many forms of therapy, but prompt surgical treatment with evacuation of the blood from the corpora appears to be the best. Regardless of treatment, impotence is common after an episode of priapism and even more common after repeated episodes of priapism.

(N.A.Ro./J.K.La.)

INFECTIOUS DISEASES SPREAD BY SEXUAL CONTACT

Sexually transmitted (or transmissible) diseases, formerly called venereal diseases, are usually contracted during sexual intercourse with an infected partner. The principal disorders commonly transmitted in this manner include syphilis, gonorrhea, genital herpes, nongonococcal urethritis, and chancroid.

In addition, various intestinal disorders, among them amebic dysentery, shigellosis, and giardiasis, and type B hepatitis may be transmitted during sexual intercourse, more frequently in homosexual than in heterosexual relations. An often fatal disease complex designated acquired immune deficiency syndrome (AIDS), first identified by U.S. public authorities in 1980, has also been associated with homosexual contact—as well as with intravenous drug use and blood transfusions. The lowered immunity of AIDS victims renders them particularly susceptible to rare forms of cancer and pneumonia. Cause, cure, and methods of prevention are unknown. Because the above mentioned diseases are also transmitted by other than sexual mechanisms they are not regarded as exclusively sexually transmissible.

Acquired immune deficiency syndrome

Syphilis. Syphilis is caused by the bacterial spirochete *Treponema pallidum*. Although known in Europe since the 15th century, syphilis was not recognized as a venereal disease until some 200 years ago. It first appears as a painless lump on the skin or mucous membranes of the genitals two to four weeks after sexual exposure, although the initial symptoms may appear in other areas in unusual cases. Syphilis is considered a systemic disease from its onset and can have serious consequences in the nervous system and other organs. The initial lump breaks down to form a hard ulcer called a chancre; at this point, diagnosis can be made by observations of spirochetes in material taken from the open sore and viewed under the microscope. The infection induces antibodies against *T. pallidum* that can be identified in the bloodstream by various tests some weeks after the initial infection. If left untreated, the chancre disappears, and the patient develops flat, raised nodules on the genitals (secondary syphilis). Subsurface nodules, called gumma, appear in the tertiary stage of the disease.

The organism invades the nervous system at an early stage, but neurologic symptoms, including behavioral aberrations, often do not occur until the infection has been present for several years. Massive doses of penicillin are used to treat all stages of syphilis but are most effective during the primary stage; penicillin can also prevent transmission of the infection from a pregnant woman to the fetus, which could result in miscarriage or severe congenital defects. Other antibiotics may be effective but generally are used only in patients allergic to penicillin. Syphilis exhibited a marked decline in incidence beginning in 1946, three years after the introduction of penicillin, but has since shown a renewed increase.

Gonorrhea. The most common venereal disease, gonorrhea, was known in ancient times. It is caused by *Neisseria gonorrhoeae*, a bacterium identified in 1879. The organism has an extremely short incubation period, making it difficult to interrupt the chain of transmission. Infection, almost invariably due to sexual intercourse, can be prevented by the use of a condom; some attempts have been made to develop prophylactic vaccines but have not met with much success. The chief symptom of gonorrhea in the male is pain or burning during urination. Some 50 percent of infected females are asymptomatic; in symptomatic cases the signs of infection are similar to those seen in the male. Gonorrhea spreads locally along mucosal surfaces, ascending the urethra in the male and either the vagina or the urethra in the female. The advancing infection causes a purulent discharge into the urine. The bacteria may also be disseminated through the blood to more distant sites; systemic manifestations include headache, and if untreated, arthritis or heart disease.

The usual treatment is penicillin, although *N. gonorrhoeae* has developed resistance requiring a steady increase in the recommended dosage since the drug was introduced. In the 1970s strains of gonorrhea resistant to penicillin at any dosage were identified, chiefly in the Far East; in such resistant infections, more toxic antibiotics may be used. Despite antibiotic therapy, gonorrhea increased steadily in the United States and Britain beginning in the mid-1950s. Testing of an experimental vaccine against gonorrhea began in the early 1980s.

Resistance to penicillin

Nongonococcal urethritis. Although recognized only recently by public health officials, nongonococcal, or nonspecific, urethritis (NGU) is one of the most common sexually transmitted infections. While caused by a variety of microorganisms, it is most commonly attributed to *Chlamydia* species, which also cause lymphogranuloma venereum (see below). In about half the cases, although no bacteria can be identified, an infectious transmission is strongly implicated. The symptoms are those of low-grade urethritis, chiefly pain and burning on urination, but are generally milder than those of gonorrhea. Treatment varies depending on the causative microorganism.

Herpes. Herpes genitalis became a major problem in the 1970s and '80s. The disease may be caused by the herpes simplex viruses identified as type 1 (HSV-1; the cause of cold sores of the lips and mouth) and type 2 (HSV-2). Another common herpes virus, cytomegalovirus (CMV), present in numerous healthy persons and widespread among male homosexuals, is associated with high mortality in patients taking immunosuppressive drugs. Genital herpes first appears as groups of small blisters on the surface of the penis in men and the vulva in women. The initial infection clears spontaneously within a few days, but herpes commonly recurs with varying frequency thereafter, burning or itching at the infection site containing the lesions. Herpes is generally transmitted only when an active lesion is present; it can be prevented by avoidance of intercourse during the active phase. The risk of transmission is diminished by the use of a condom. At present, there are no satisfactory treatments or effective vaccines against the herpes virus. Active herpes can be fatal to infants during delivery; in a large percentage of cases it causes blindness or brain damage in newborns. In women, genital herpes has also been associated with cancer of the cervix although no causal mechanism is evident.

Chancroid. Chancroid, also called "soft sore," is caused by the microorganism *Haemophilus ducreyi* and occurs chiefly in the tropics and in Asia. The bacteria has a short incubation period, producing small red pustules generally within fewer than five days after exposure but occasionally in as many as 30. The pustules burst to form painful ulcers; chancroid can be diagnosed by culturing bacteria from these ulcers. Unlike syphilis, which it may resemble, chancroid is a purely localized disease of the genitals. Treatment is with sulfonamides, streptomycin, or tetracycline.

Lymphogranuloma venereum. This infection, common in the tropics but very rare in temperate regions, is caused by *Chlamydia*. It is usually transmitted through intercourse but may be contracted in other ways. Typically, a transient genital blister is followed by regional inflammation of the lymph nodes. If untreated, this condition may progress to genital elephantiasis and has in some cases been linked to malignancy. The most effective remedy is sulfonamides, but no treatment is totally reliable. Surgical removal of diseased tissue may be necessary.

Genital elephantiasis

Genital warts (condyloma acuminata). Genital warts—caused by the same papilloma virus that produces common skin warts—are almost always transmitted through sexual intercourse. The wart begins as a pinhead-sized

swelling that enlarges and becomes pedunculated; the mature wart is often composed of many smaller swellings and may resemble the genital lesions of secondary syphilis.

Granuloma inguinale. Although it may be spread through sexual contact, granuloma inguinale has very low infectivity and has not been shown to be consistently transmitted between sexual partners. Granuloma inguinale is caused by infection with *Donovania granulomatis* and occurs primarily in tropical and subtropical climates, including the southern United States. Initial symptoms are painless papules that become ulcerated, ultimately forming granulomatous masses that tend to bleed easily. These lesions occur on the genitals, thighs, and groin of infected persons and may resemble syphilis lesions, a reason for additional concern. Malignancy has also been associated with granuloma inguinale. Treatment is chiefly with tetracyclines or penicillin.

Genital candidiasis (moniliasis). Local infections with the yeast *Candida albicans* in men almost always are acquired through sexual contacts, but in women, in whom candidiasis is much more common, the infection can be acquired in a variety of ways. In men, condidiasis involves the surface of the glans penis, causing intense burning or itching. In women, candidiasis frequently produces no symptoms but can cause vaginal and vulval irritation (sometimes with production of a thick, white discharge) or pain during urination. The diagnosis is made by culturing yeast from the involved area; treatment is by local antifungal agents.

Trichomoniasis. Infection with the flagellate protozoan *Trichomonas vaginalis* is usually, but not exclusively, spread by sexual contact. The condition is commonly asymptomatic in males. In females trichomoniasis has a variety of manifestations, including vaginal discharge and inflammation of the vulva, perineum, and thighs. Both sexes may experience complications, cystitis and urethritis; males may also contract prostatitis and epididymitis. Treatment with metronidazole, an antibacterial, antiprotozoal agent, is standard.

OTHER INFECTIONS AFFECTING THE REPRODUCTIVE SYSTEM

Puerperal infection. A common cause of death during childbirth, especially before the adoption of modern sanitary practices, puerperal infections occur when bacteria, usually *Streptococcus,* invade wounds in the birth canal. The infection may cause abscess formation and can involve all of the genital organs and adjacent blood vessels, reproductive structures, and other abdominal tissues. Treatment consists of antibiotics, supportive therapy, and occasionally surgical drainage of abscesses.

Tuberculosis. Primary tuberculosis of the reproductive system is rare and is usually brought from elsewhere in the body through the bloodstream. Nodular or pustular lesions on the penis or scrotum of men or the vulva of women, resembling the gumma (nodules) of tertiary syphilis, may appear one week after tubercular infection. The nodules can become ulcerated, resembling the primary chancre of syphilis. Tubercular abscesses can also develop in most of the internal reproductive organs. Treatment consists of administration of antitubercular drugs for up to two years. As the incidence of tuberculosis has declined in the developed countries, tuberculosis of the reproductive system has became exceedingly rare.

Inflammatory conditions. Balanitis, or inflammation of the glans penis, and posthitis, or infection of the prepuce, result from the retention of secretions and bacteria beneath the foreskin and can be prevented with good sexual hygiene. Balanitis can also develop as a complication of certain sexually transmitted diseases. Acute prostatitis, inflammation of the prostate gland, may be caused by any of a variety of microorganisms, including those which cause venereal diseases; chronic prostatitis, the most common reproductive system infection in men older than 50, often follows the acute infection. Epididymitis, inflammation of the epididymis (a duct of the sperm canal), can result in sterility. All of these are nonspecific infections that must be treated with antibiotics appropriate for the causative organisms.

Acute prostatitis

In women, other infections of the reproductive system include bartholinitis, an inflammation of Bartholin's duct near the opening of the vagina, and vaginitis, generalized inflammation of the vagina caused by various yeasts and bacteria. The most common symptoms of such ailments are vaginal discomfort or itching and pain during urination or intercourse. Again, treatment of these conditions depends largely on the causative organism. (Ed.)

STRUCTURAL CHANGES OF UNKNOWN CAUSES

In the female: endometriosis. Endometriosis, a disease occurring only during a woman's menstrual life, is the growth of endometrial tissue in an abnormal location. This may occur in the uterus or elsewhere. The most common location of the implants of endometrial tissue are the ovaries; other areas and organs affected (in order of incidence) are uterosacral ligaments (thickened portions of the sheet of connective tissue covering the pelvic organs), the rectovaginal septum (the membrane dividing the rectum from the vagina), the sigmoid colon (that portion of the large intestine that leads into the rectum), the lower genital tract, the round ligaments of the uterus, and the peritoneum (membrane) lining the pelvis. Although endometriosis is a progressive disease in most instances, pain relief following conservative surgery (surgery that preserves ability to bear children) has occurred in an estimated 80 percent of patients, and 40–50 percent were able to become pregnant.

In the male: benign hypertrophy of the prostate. Benign prostatic hypertrophy, an overgrowth of normal glandular and muscular elements of the prostate gland, arises in the immediate vicinity of the urethra and is the most frequent cause of urinary obstruction. The enlarged prostate usually causes symptoms after the age of 50. If undetected, the obstruction may cause bladder and kidney damage. The diagnosis is made by rectal examination, excretory urography (X-raying the urinary tract while an opaque substance is being excreted in the urine), and cystourethroscopy (direct viewing of the bladder and urethra). Treatment is by surgical removal of the excess tissue. The prognosis is good if detection is early and treatment is given before the kidneys are damaged.

TUMOURS

In the male. *External genitalia.* Tumours of the penis are almost all of epithelial (covering or lining) origin and usually involve the foreskin or glans. Cancer of the penis (epithelioma) is rarely found in men who have been circumcised during infancy. The growth arises on the glans or inner surfaces of the prepuce, and metastases (secondary growths at distant parts) occur through lymph channels that lead to the inguinal (groin) and iliac nodes (nodes along the aorta and iliac arteries). The diagnosis is made by examination of a specimen of the lesion. Treatment for small lesions consists of surgical removal of a part of the penis or by X-ray therapy, while spread to inguinal nodes may be treated by removal of the node. The outlook is good if the cancer is small and there has been no metastasis.

Tumours of the scrotal skin are rare; most are thought to arise from occupational exposure to various carcinogens (cancer-causing substances), such as the soot in chimney sweeps' clothing. Primary tumours of the epididymis are also uncommon, and most are benign.

Testicular tumours. Testicular tumours are usually malignant; the peak incidence is between the ages of 20 and 40 years. This type of cancer accounts for about 0.5 percent of all malignant growths in men and about 4 percent of all tumours affecting the genitourinary tract. The great majority of testis tumours (greater than 95 percent) are of types that do not reproduce cells resembling those of the tissue of origin. The major route of metastases is via the lymphatics. The lymph nodes in the loins and the mediastinum—the region between the lungs—are most commonly involved, but the lungs and liver are also frequent sites of tumour spread. The remaining 5 percent of the testicular tumours, which usually resemble the cells from which they arise, include the hormone-secreting tumours. In general, these tumours have been described in all age groups, have usually been benign in behaviour, and have

been most frequent in poorly developed or undescended testes.

The most common symptom first observed in all groups is painless enlargement of the testis. If, after careful examination, tumour cannot be ruled out, the testicle is removed for microscopic examination. Further treatment may consist of removal of the retroperitoneal lymph nodes (the lymph nodes in the region behind the peritoneum, the membrane lining the abdomen), X-ray therapy, or chemotherapy.

Carcinoma of the prostate. Carcinoma (cancer) of the prostate is rare before the age of 60 but increases in frequency every decade thereafter. It is the second most common cause of death from cancer in the male, second to cancer of the lung. In men over 60, it is the commonest cause of cancer deaths. Like most tumours, prostatic cancer has no known cause, but it is clear that its growth is strikingly influenced by sex hormones or their withdrawal. Viruses may also play a role. The progress of the cancer is so slow that, by the time it produces symptoms of urinary obstruction, metastasis has occurred in many cases, most frequently to the spine, the pelvic bones, or the upper portions of the thigh bones. The diagnosis is made by finding cancer cells in a specimen of tissue taken from the prostate. Elevated levels of acid phosphatase (an enzyme of the prostate) are found in the blood (in 75 percent of cases) when the cancer has extended outside the prostate capsule and metastases are present.

Slow growth of prostate cancer

If the tumour is discovered before it has extended beyond the prostate, the gland is removed. If spread has occurred, various palliative measures offer the affected person much relief.

In the female. *Carcinoma of the vulva.* Primary carcinoma of the vulva (the external female genital organs) usually occurs in women over 50 and usually arises from the labia majora or labia minora. Most patients first notice a lump on the vulva or perineum; the diagnosis is made by examination of a specimen of tissues. Treatment consists of surgical removal of the vulva and of regional lymph nodes.

Cancer of the cervix of the uterus. Cancer of the cervix is the most common malignant tumour of the female genital tract; it is second only to cancer of the breast as a cause of death from cancer in women. The average age of occurrence for cancer of the cervix is the 45th year. The initial diagnosis is made by screening with such tests as those developed by Papanicolaou and Traut. (These consist of staining smears from vaginal and other secretions and examining them for cancer cells.) The final diagnosis rests on examining specimens of tissue from the cervix. Treatment now is usually irradiation instead of surgery because of the uncertainties of total surgical excision and the illness associated with extreme surgery. The prospect of five-year survival is as good as 85 percent if the cancers have not spread beyond the cervix.

Uterine fibromyomas. Uterine fibromyomas (fibroids) are the most frequent cause of enlargement of the uterus. They are most common in blacks and in persons who have not borne children and are most often identified in women aged 30–45 years. New tumours do not originate after the menopause, and existing ones usually regress at that time but do not disappear. The tumours, which are benign, originate from the smooth muscle cells of the uterus wall and may be single but usually are multiple, pseudoencapsulated nodules. The symptoms are quite variable and depend largely on the location and size of the tumour. Excessive menstrual bleeding is often caused by fibroids. The diagnosis is tentatively made by pelvic examination and confirmed at surgery. Small asymptomatic fibromyomas need not be treated; the larger ones are dealt with by total or partial removal of the uterus or by irradiation.

Diagnosis of fibromas

Carcinoma of the body of the uterus. Cancer of the endometrium (the lining) of the body of the uterus is the second most common malignant tumour of the uterus and the female genital tract. The peak incidence is in the mid-50s, and there is also a strikingly high incidence in women who have not borne children. The chief symptom of the cancer is postmenopausal uterine bleeding. The Papanicolaou smear is not a reliable screening test, and an examination of a specimen of endometrial tissue must be performed. The treatment is primarily surgical but is often supplemented with preoperative intrauterine radium application or preliminary deep X-ray therapy to the pelvis. The survival rate from this disease is relatively good if the tumour is confined to the uterine body.

Ovarian tumours. No other organ in the body develops such a variety of tumours as does the ovary. The symptoms and signs may be due to the hormones secreted or may be only those of an enlarging mass in the pelvis. The final diagnosis is usually made at abdominal exploration. The treatment consists of surgery, X-ray therapy, or chemotherapy. The prognosis is variable and depends on the type of tumour that is present as well as the extent of metastatic spread. (N.A.Ro./J.K.La./Ed.)

HUMAN REPRODUCTION FROM CONCEPTION TO BIRTH

The normal events of pregnancy

INITIATION OF PREGNANCY

Union of the ovum and the sperm

A new individual is created when the elements of a potent sperm merge with those of a fertile ovum, or egg. Before this union both the spermatozoon (sperm) and the ovum have migrated for considerable distances in order to achieve their union. A number of actively motile spermatozoa are deposited in the vagina, pass through the uterus, and invade the uterine (fallopian) tube, where they surround the ovum. The ovum has arrived there after extrusion from its follicle, or capsule, in the ovary. After it enters the tube, the ovum loses its outer layer of cells as a result of action by substances in the spermatozoa and from the lining of the tubal wall. Loss of the outer layer of the ovum allows a number of spermatozoa to penetrate the egg's surface. Only one spermatozoon, however, normally becomes the fertilizing organism. Once it has entered the substance of the ovum, the nuclear head of this spermatozoon separates from its tail. The tail gradually disappears, but the head with its nucleus survives. As it travels toward the nucleus of the ovum (at this stage called the female pronucleus), the head enlarges and becomes the male pronucleus. The two pronuclei meet in the centre of the ovum, where their threadlike chromatin material organizes into chromosomes.

Originally the female nucleus has 44 autosomes (chromosomes other than sex chromosomes) and two (X, X) sex chromosomes. Before fertilization a type of cell division called a reduction division brings the number of chromosomes in the female pronucleus down to 23, including one X-chromosome. The male gamete, or sex cell, also has 44 autosomes and two (X, Y) sex chromosomes. As a result of a reducing division occurring before fertilization, it, too, has 23 chromosomes, including either an X or a Y sex chromosome at the time that it merges with the female pronucleus.

Determination of sex

After the chromosomes merge and divide in a process termed mitosis, the fertilized ovum, or zygote, as it is now called, divides into two equal-sized daughter cells. The mitotic division gives each daughter cell 44 autosomes, half of which are of maternal and half of paternal origin. Each daughter cell also has either two X-chromosomes, making the new individual a female, or an X- and a Y-chromosome, making it a male. The sex of the daughter cells is determined, therefore, by the sex chromosome from the male parent.

Fertilization occurs in the uterine tube. How long the zygote remains in the tube is unknown, but it probably reaches the uterine cavity about 72 hours after fertilization. It is nourished during its passage by the secretions from the mucous membrane lining the tube. By the time it reaches the uterus, it has become a mulberry-like solid mass called a morula. A morula is composed of 60 or

more cells. As the number of cells in a morula increases, the zygote forms a hollow bubblelike structure, the blastocyst. The blastocyst, nurtured by the uterine secretions, floats free in the uterine cavity for a short time and then is implanted in the uterine lining. Normally, the implantation of the blastocyst occurs in the upper portion of the uterine lining. (The mechanism of implantation is described below.)

DIAGNOSIS OF PREGNANCY

Symptoms and signs; biological tests. Outward early indications of pregnancy are missed menstrual periods, morning nausea, and fullness and tenderness of the breasts; but the positive and certain signs of gestation are the sounds of the fetal heartbeat, which are audible with a stethoscope between the 16th and the 20th week of pregnancy; ultrasound images of the growing fetus, which can be observed throughout pregnancy; and fetal movements, which usually occur by the 18th to the 20th week of pregnancy.

Persons who note their body temperature upon awakening, as many women do who wish to know when they are ovulating, may observe continued elevation of the temperature curve well beyond the time of the missed period; this is strongly suggestive of pregnancy. During the early months of pregnancy, women may notice that they urinate frequently, because of pressure of the enlarging uterus on the bladder; feel tired and drowsy; dislike foods that were previously palatable; have a sense of pelvic heaviness; and are subject to vomiting (which can be severe) and to pulling pains in the sides of the abdomen, as the growing uterus stretches the round ligaments that help support it, singly or together. Most of these symptoms subside as pregnancy progresses. The signs and symptoms of pregnancy are so definite by the 12th week that the diagnosis is seldom a problem.

Biological tests for pregnancy

Biological tests for pregnancy depend upon the production by the placenta (the temporary organ that develops in the womb for the nourishing of the embryo and the elimination of its wastes) of chorionic gonadotropin, an ovary-stimulating hormone. In practice, the tests have an accuracy of about 95 percent, although false-negative tests may run as high as 20 percent in a series of cases. False-negative reports are frequently obtained during late pregnancy when the secretion of chorionic gonadotropin normally decreases. The possibility not only of false-negative but also of false-positive tests makes the tests, at best, probable rather than absolute evidence of the presence or absence of pregnancy. Chorionic gonadotropin in a woman's blood or urine indicates only that she is harbouring living placental tissue. It does not tell anything about the condition of the fetus. In fact, the greatest production of chorionic gonadotropin occurs in certain placental abnormalities and disorders that can develop in the absence of a fetus.

Tests using immature mice (the Aschheim-Zondek test) and immature rats have been found to be extremely accurate. Tests using rabbits (the Friedman test) have been largely replaced by the more rapid and less expensive frog and toad tests.

The use of the female South African claw-toed tree toad, *Xenopus laevis,* is based on the discovery that this animal will ovulate and extrude visible eggs within a few hours after it has received an injection of a few millilitres of urine from a pregnant woman. The male common frog, *Rana pipiens,* will extrude spermatozoa when treated in the same way. Both of these tests are considered somewhat unsatisfactory because false-positive reactions are not uncommon.

Several immunological reaction tests in common use are based upon the inhibition of hemagglutination (clotting of red cells). A positive test is obtained when human chorionic gonadotropin (HCG) in the woman's urine or blood is added to human chorionic gonadotropin antiserum (rabbit blood serum containing antibodies to HCG) in the presence of particles (or red blood cells) coated with human chorionic gonadotropin. The hormone from the woman will inhibit the combination of coated particles and antibody, and agglutination does not occur. If there is no chorionic gonadotropin in her urine, agglutination will occur and the test is negative.

Several "signs" noted by the physician during an examination will suggest that a patient may be in the early months of pregnancy. Darkening of the areola of the breast (the small, coloured ring around the nipple) and prominence of the sebaceous glands around the nipple (Montgomery's glands); purplish-red discoloration of the vulvar, vaginal, and cervical tissues; softening of the cervix and of the lower part of the uterus and, of course, enlargement and softening of the uterus itself are suggestive but not necessarily proof of pregnancy.

Conditions that may be mistaken for pregnancy. Other conditions may confuse the diagnosis of pregnancy. Absence of menstruation can be caused by chronic illness, by emotional or endocrine disturbances, by fear of pregnancy, or by a desire to be pregnant. Nausea and vomiting may be of gastrointestinal or psychic origin. Tenderness of the breasts can be due to a hormonal disturbance.

Absence of menstruation; false-positive tests

Any condition that causes pelvic congestion, such as a pelvic tumour, may cause duskiness of the genital tissues. At times a soft tumour of the uterus may simulate a pregnancy. The question of pregnancy may be raised if the woman does not menstruate regularly; the absence of other symptoms and signs of gestation indicates that she is not pregnant. There are rare ovarian and uterine tumours that produce false-positive pregnancy tests. It may be difficult for the physician to exclude pregnancy on the basis of an examination if the uterus is tipped back and difficult to feel, or if it is enlarged by a tumour within it. If other signs of pregnancy are absent, however, and the tests for pregnancy are negative, pregnancy can most likely be ruled out.

Childless women who greatly desire a baby sometimes suffer from false or spurious pregnancy (pseudocyesis). They stop menstruating, have morning nausea, "feel life," and have abdominal enlargement caused by fat and intestinal gas. At "term" they may have "labour pains." Signs of pregnancy are absent. Treatment is by psychotherapy.

Menopausal women often fear pregnancy when their periods stop; information that they show no signs of pregnancy usually reassures them. Retained uterine secretions of bloody or watery fluid, caught above a blocked mouth of the uterus (cervix), prevent menstruation, cause softening and enlargement of the uterus, and may cause the patient to wonder whether she is pregnant. There are no other signs of pregnancy, and the hard cervix, closed by scar tissue, explains the problem.

Duration of pregnancy. There are, as a rule, 266 to 270 days between ovulation and childbirth, with extremes of 250 and 285 days. Physicians usually determine the date of the estimated time for delivery by adding seven days to the first day of the last menstrual period and counting forward nine calendar months; *i.e.,* if the last period began on January 10, the date of delivery is October 17. Courts of law, in determining the legitimacy of a child, may accept much shorter or much longer periods of gestation as being within the periods of possible duration of a pregnancy. One court in the state of New York has accepted a pregnancy of 355 days as legitimate. British courts have recognized 331 and 346 days as legitimate with the approval of medical consultants. Fully developed infants have been born as early as 221 days after the first day of the mother's last menstrual period.

Because the exact date of ovulation is usually not known, it is seldom possible to make an accurate estimate of the date of delivery. There is a 5 percent chance that a baby will be born on the exact date estimated from the above rule. There is a 25 percent chance that it will be born within four days before or after the estimated date. There is a 50 percent chance that delivery will occur on the estimated date plus or minus seven days. There is a 95 percent chance that the baby will be born within plus or minus 14 days of the estimated date of delivery.

ANATOMIC AND PHYSIOLOGIC CHANGES OF NORMAL PREGNANCY

Changes in organs and tissues directly associated with childbearing. *Ovaries.* The ovaries of a nonpregnant

young woman who is in good health go through cyclic changes each month. These changes centre about a follicle, or "egg sac." A new follicle develops after each menstrual period, casts off an egg (ovulation), and, after ovulation, forms a new structure (the corpus luteum).

Ovulation and formation of corpus luteum; progesterone and estrogen

If the egg is fertilized, it is sustained for a short time by the hormones produced by the corpus luteum. Progesterone and estrogen, secreted by the corpus luteum, are essential for the preservation of the pregnancy during its early months. If pregnancy does not occur, the egg disintegrates and the corpus luteum shrinks. As it shrinks, the stimulating effect of its hormones, progesterone and estrogen, is withdrawn from the endometrium (the lining of the uterus), and menstruation occurs. The cycle then begins again.

Pregnancy, if it occurs, maintains the corpus luteum by means of the hormones produced by the young placenta. The corpus luteum is not essential in human pregnancy after the first few weeks because of the takeover of its functions by the placenta. In fact, human pregnancies have gone on undisturbed when the corpus luteum has been removed as early as the 41st day after conception. Gradually the placenta, or afterbirth, begins to elaborate progesterone and estrogen itself. By the 70th day of pregnancy the placenta is unquestionably able to replace the corpus luteum without endangering the pregnancy during the transfer of function. At the end of pregnancy the corpus luteum has usually regressed until it is no longer a prominent feature of the ovary.

During the first few months of pregnancy the ovary that contains the functioning corpus luteum is considerably larger than the other ovary. During pregnancy, both ovaries usually are studded with fluid-filled egg sacs as a result of chorionic gonadotropin stimulation; by the end of pregnancy, most of these follicles have gradually regressed and disappeared.

The blood supply to both ovaries is increased during pregnancy. Both glands frequently reveal plaques of bright red fleshy material on their surfaces, which, if examined microscopically, demonstrate the typical cellular change of pregnancy, called a decidual reaction. In this reaction, cells develop that look like the cells in the lining of the pregnant uterus. They result from the high hormone levels that occur during pregnancy and disappear after the pregnancy terminates.

The uterus before and during pregnancy

The uterus and the development of the placenta. The uterus is a thick-walled, pear-shaped organ measuring seven centimetres (about 2.75 inches) in length and weighing 30 grams (about one ounce) in an unpregnant woman in her later teens. It has a buttonlike lower end, the cervix, that merges with the bulbous larger portion, called the corpus. The corpus comprises approximately three-fourths of the uterus. There is a flat, triangular-shaped cavity within the uterus. At term, the uterus is a large, thin-walled, hollow, elastic, fluid-filled cylinder measuring approximately 30 centimetres (about 12 inches) in length, weighing approximately 1,200 grams (2.6 pounds), and having a capacity of 4,000 to 5,000 millilitres (4.2 to 5.3 quarts).

The greater size of the uterus as a result of pregnancy is due to a marked increase in the number of muscle fibres, blood vessels, nerves, and lymphatic vessels in the uterine wall. There is also a five- to tenfold increase in the size of the individual muscle fibre and marked enlargement in the diameters of the blood and lymph vessels.

During the first few weeks of pregnancy, the shape of the uterus is unchanged, but the organ becomes gradually softer. By the 14th week it forms a flattened or oblate spheroid. The fibrous cervix becomes remarkably softer and acquires a protective mucus plug within its cavity, but otherwise it changes little before labour. The lower part of the corpus, the isthmus, first becomes elongated and then, as the uterine contents demand more space, stretches and unfolds to form a bowl-shaped formation called the lower uterine segment. The fibrous nature of the cervix causes it to resist this unfolding action.

The uterine wall is stretched and thinned during pregnancy by the growing conceptus, as the whole product of conception is called, and by the fluid that surrounds it. By term, this process converts the uterus into an elastic, fluid-filled cylinder. It is only late in pregnancy that the cervix gradually thins out and softens; during labour it dilates for passage of the infant.

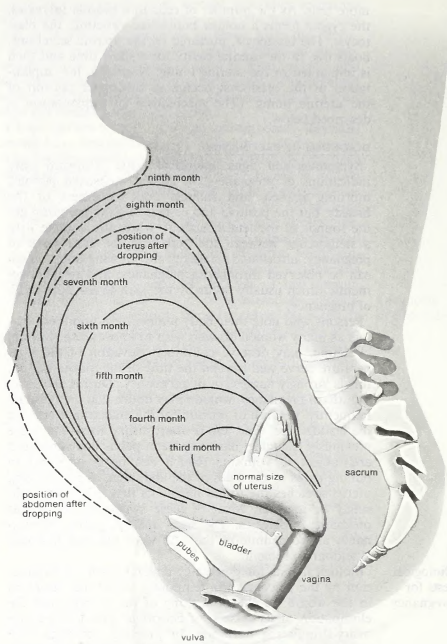


Figure 31: Levels of the uterus with advancing months of pregnancy.
From *Dorland's Illustrated Medical Dictionary*, 24th ed. (1965). W.B. Saunders Co., Philadelphia

As pregnancy progresses, the uterus rises out of the pelvis and fills the abdominal cavity. It is top-heavy near term so that it falls forward and, because of the large bowel on the left side, rotates to the right. It presses on the diaphragm and pushes the other organs aside. The uterus may sink downward in the pelvis several weeks before term in a process that is known as lightening or dropping. This occurs as the fetal head descends into the pelvis. In some women, particularly those who have borne children, lightening does not occur until the onset of labour. Lightening may be impossible in women who have an abnormally small pelvis, an oversized fetus, or a fetus lying in an abnormal position.

For a short time after fertilization, the conceptus, a minute bubblelike structure called a blastocyst, lies unattached in the uterine cavity. The cells that will become the embryo (the embryonic disk) form a thickened layer on one side of the bubble. Elsewhere, the walls of the bubble consist of a single layer of cells; these cells are the trophoblast, which has a special ability to attach to and invade the uterine wall. The trophoblast plays an important role later in the development of the placenta or afterbirth. The conceptus makes contact with the uterine lining about the fifth or sixth day after conception. After contact the blastocyst collapses to form a rounded disk with the embryonic mass on the surface and the trophoblast against the endometrium (uterine lining). The part of the trophoblast

Implantation

that is in contact with the endometrium grows into and invades the maternal tissue. Concomitant disintegration of the endometrium allows the conceptus to sink into the uterine lining.

The chorionic cavity

Soon the entire blastocyst is buried in the endometrium. Proliferation of the trophoblast over the part of the collapsed bubble that is opposite the embryo is part of the implantation procedure that helps to cover the blastocyst. After a few days, a cavity forms that bears the same relation to the embryonic disk that the blastocyst cavity did before; this cavity will become the fluid-filled chorionic cavity containing the embryo. Ultimately it will contain the amniotic fluid that surrounds the fetus, the fetus itself, and the umbilical cord.

The body stalk, which will become the umbilical cord, then begins to separate the embryo from the syncytiotrophoblast, the outer layer of the trophoblast lying against the endometrium; the inner lining of the trophoblast is called cytotrophoblast. As the syncytiotrophoblast advances into the endometrium, it surrounds minute branches of the uterine arteries that contain maternal blood. Erosion of the endometrium about these blood sinuses allows them to open into the small cavities in the trophoblast. The cytotrophoblast, which lines the cavity, forms fingers of

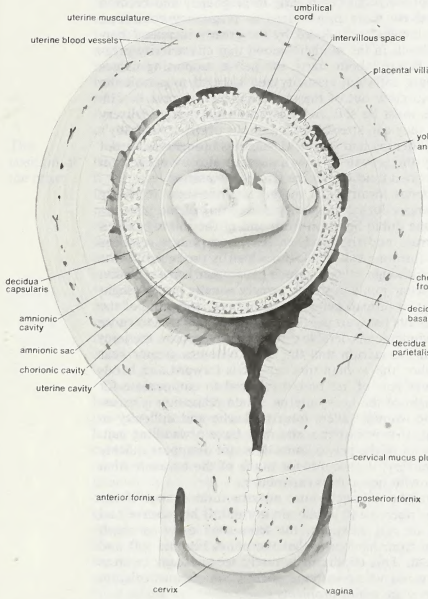


Figure 32: Sagittal section of the pregnant uterus at the seventh week of pregnancy showing the placenta and fetal membranes.

proliferating cells extending into the syncytiotrophoblast. After the placenta is developed, these fingers will be the cores of the root-like placental villi, structures that will draw nutrients and oxygen from the maternal blood that bathes them. This is the first step in uteroplacental circulation, which supplies the fetus with all of the sustenance necessary for life and growth and removes waste products from it. During the third week of pregnancy, the syncytiotrophoblast forms a single layer of cells covering the growing villi and lining the syncytial lacunae or small cavities between the villi. The conceptus is buried in the endometrium, and its whole surface is covered at this time by developing villi. The greater part of the chorionic wall is now cytotrophoblast. Fingers of cytotrophoblast in the form of cell masses extend into the syncytial layer. Soon thereafter, a layer of connective tissue, or mesoderm, grows into the villi, which now form branches as they spread out into the blood-filled spaces in the endometrium adjacent to the conceptus.

By the end of the third week, the chorionic villi that form the outer surface of the chorionic sac are covered by a thick layer of cytotrophoblast and have a connective tissue core within which embryonic blood vessels are beginning to develop. The vessels, which arise from the yolk sac, connect with the primitive vascular system in the embryo. As growth progresses the layer of cytotrophoblast begins to regress. It disappears by the fifth month of pregnancy.

The layer of endometrium closest to the encroaching conceptus forms, with remnants of the invading syncytiotrophoblast, a thin plate of cells known as the decidua basalis, the maternal component of the mature placenta; it is cast off when the placenta is expelled. The fetal part of the placenta—the villi and their contained blood vessels—is separated from the decidua basalis by a lakelike body of fluid blood. This pool was created by coalescence of the intervillous spaces. The intervillous spaces in turn were formed from the syncytial lacunae in the young conceptus. Maternal blood enters this blood mass from the branches of the uterine arteries. The pool is drained by the uterine veins. It is so choked by intermingling villi and their branches that its continuity is lost on gross inspection.

The decidua capsularis, the chorion laeve, and the chorion frondosum

The chorionic cavity contains the fluid in which the embryo floats. As its shell or outer surface becomes larger, the decidua capsularis, which is that part of the endometrium that has grown over the side of the conceptus away from the embryo (*i.e.*, the abembryonic side) after implantation, becomes thinner. After 12 weeks or so, the villi on this side, which is the side directed toward the uterine cavity, disappear, leaving the smooth chorion, now called the chorion laeve. The chorion frondosum is that part of the conceptus that forms as the villi grow larger on the side of the chorionic shell next to the uterine wall. The discus-shaped placenta develops from the chorion frondosum and the decidua basalis.

At term, the normal placenta is a disk-shaped structure approximately 16 to 20 centimetres (about six to seven inches) in diameter, three or four centimetres (about 1.2–1.6 inches) in thickness at its thickest part, and weighing between 500 and 1,000 grams (1.1 and 2.2 pounds). It is thinner at its margins, where it is joined to the membrane-like chorion which spreads out over the whole inner surface of the uterus and contains the fetus and the amniotic fluid. The amnion, a thinner membrane, is adherent to and covers the inner surface of the chorion. The inner or fetal surface of the placenta is shiny, smooth, and traversed by a number of branching fetal blood vessels that come together at the point—usually the centre of the placenta—where the umbilical cord attaches. The maternal or uterine side of the placenta, covered by the thin, flaky decidua basalis, a cast-off part of the uterine lining, is rough and purplish-red, and has a raw appearance. When the placenta is cut across, its interior is seen to be made of a soft, crepelike or spongy matrix from which semisolid or clotted blood, caught when it is separated from the uterine wall to which it was attached, can be squeezed. Detailed examination shows that the villi and their branches form an arborescent (treelike) mass within the huge blood lake of the intervillous space. Anchoring villi extend outward from the fetal side and fuse with the decidua basalis to hold the organ's shape. Others, algae-like, float freely in the blood lake. Dividing partitions, formed from the trophoblast shell, project into the intervillous space from the decidual side. They divide the placenta into 15 or 20 compartments, which are called cotyledons.

Maternal blood flows from the uterine vessels into the trophoblast-lined intervillous blood lake. Within each villus is a blood vessel network that is part of the fetal circulatory system. Blood within the villous vessel is circulated by the fetal heart. The blood vessel wall, the connective tissue of the villous core, and the syncytiotrophoblast covering the villus lie between the fetal and the maternal bloodstreams. This is known as the placental

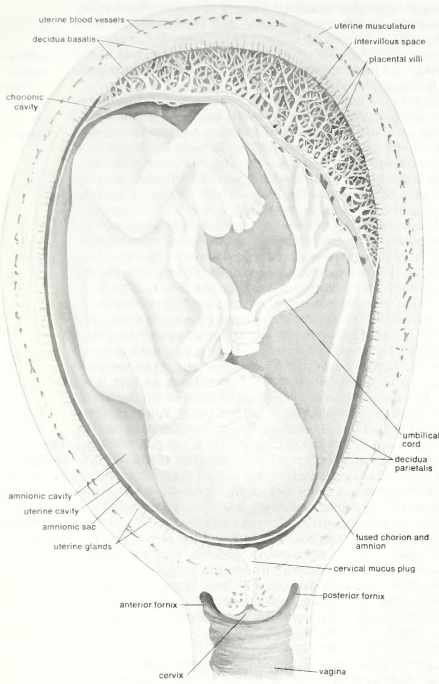


Figure 33. Sagittal section of the uterus in the fourth month of pregnancy.

barrier. As pregnancy progresses, the fetal blood vessels become larger, the connective tissue stretches over them, and the syncytiotrophoblastic layer becomes fragmentary. As a result, the placental barrier becomes much thinner. Normally, blood cells and bacteria do not pass through it, but nutrients, water, salt, viruses, hormones, and many other substances, including many drugs, can filter across it.

Uterine tube as site of fertilization

Uterine tubes. One of the two uterine tubes is the pathway down which the ripe ovum travels on its way to the uterus or womb. The spermatozoa from the male migrate up the tube, and it is there that they meet the ovum and fertilization occurs. During the first few days after fertilization the zygote, or fertilized egg, moves downward in the tube toward the uterus. While it is lying free in the tubal canal, the young conceptus is nourished by secretions from the tube. After the fertilized egg (or conceptus) passes into the uterus, the tube ceases to play any part in the pregnancy; in fact, the only function the tube has is carried out during those few days before, during, and after conception. As pregnancy goes on, the tube gradually enlarges, however, and contains more blood, as do all the pelvic organs; some of its cells may show a reaction, called a decidual reaction, to the hormones of pregnancy. As the uterus increases in size, the tubes stretch upward with it until they become two greatly enlarged elongated strands, one on each side of the uterus.

Vagina. The pinkish tan colour of the lining of the vagina gradually takes on a bluish cast during the early months of pregnancy as a result of the dilation of the blood vessels in the vaginal wall; later the vaginal wall tends to become a purplish red colour as the blood vessels become further engorged. The cells of the vaginal mucosa increase in size. Added numbers of these cells peel off the surface of the mucosa and mix with the increased vaginal fluid. This produces a profuse vaginal secretion. Thickening, softening, and relaxation of the loosely folded, succulent lining of the vagina and the sodden tissues beneath it greatly increase distensibility and capacity of the vaginal cavity; this is a process that partially prepares the birth canal for the passage through it of the large fetal mass.

External genital structures. Changes in the external genitalia are similar to those in the vagina. The tissues become first softened and more succulent and later extremely fragile, as an increasing amount of blood and fluid collects in them. They take on a purplish red colour because of increased blood supply. Darkening of the vulvar skin, frequently seen during pregnancy, is particularly common among women of Mediterranean ethnic groups.

The blood vessels, muscles, and ligaments

Other pelvic tissues. The pelvic blood vessels and lymph channels become larger and longer. They develop new branches adequate to transport the greatly increased amounts of blood and tissue fluid that accumulate in the uterus and the other pelvic organs during pregnancy. Congestion and engorgement of blood in the pelvis, both within and without the uterus, are characteristic of pregnancy.

Changes in the muscles, ligaments, and other supporting tissues of the pelvis begin early in pregnancy and become progressively more pronounced as pregnancy continues. These changes are induced by the greatly increased hormonal levels in the mother's blood that characterize pregnancy. Before labour starts, the pelvic supporting tissues must have sufficient elasticity and strength to permit the uterus to grow out of the pelvis and yet support it. The muscles must be soft and elastic enough during delivery so that they can stretch apart and not obstruct the baby's birth. Softening and greater elasticity is brought about not only by the growth of new tissue but also by congestion and retained fluid within the tissues themselves.

The bones forming the mother's pelvis show relatively few changes during pregnancy. Loosening of the joint between the pubic bones in front and of the joints between the sacrum and the pelvis in back occurs as a response to the hormone called relaxin, which is produced by the ovary. Although relaxin, which causes marked separation of the pelvic joints in some animals, usually has too slight an effect in human beings to be noticed, softening of the attachments between the bones may be sufficient to cause a few women considerable distress. The strain on the joint between the sacrum and the spine becomes greater near term when the woman tilts her pelvis forward and bends the upper part of her body backward to compensate for the weight of the heavy uterus. When relaxation is excessive, the woman suffers from backache and difficulty in walking. If it is extreme, she may have a waddling gait. Relaxation of the pelvic joints does not disappear quickly after delivery; it accounts for much of the backache that women with new babies experience.

The mother's bones show no structural change if her calcium reserve and intake are normal. If her reserve and intake are not adequate, the fetus may draw so much calcium from her bones that the bones become soft and deformed. This condition is rarely seen, except in areas of the world where extreme poverty and serious calcium deficiency are major problems.

Changes in the breasts before and after delivery

Breasts. The earliest changes in the breasts during pregnancy are an exaggeration of the frequently experienced premenstrual discomfort and fullness. The sensation is so specific for pregnancy that many women who have been pregnant before are made aware of their condition by the feeling that they have in their breasts. As pregnancy progresses the breasts become larger, the lightly pigmented area (areola) around each nipple becomes first florid or dusky in colour and then appreciably darker; during the later months the areola takes on a hue that is deep bronze or brownish black, depending on the woman's natural pigmentation. The veins beneath the skin over the breast become enlarged and more prominent. The small oily or sebaceous glands (glands of Montgomery) about the nipple become prominent.

These changes are due to the greatly increased levels of estrogen and progesterone in the woman's blood. These

ovarian hormones also prepare the breast tissue for the action of the lactogenic (milk-causing) hormone, prolactin, produced by the pituitary gland. During the later part of pregnancy a milky fluid, colostrum, exudes from the ducts or can be expressed from them.

After delivery the decrease in estrogen and progesterone levels presumably permits the pituitary gland to release prolactin, which causes the breast to secrete milk. It is thought that the high hormonal levels inhibit the action or secretion of prolactin before delivery. Prolactin continues to be produced, and lactation usually continues, as long as the mother feeds her baby at the breast.

Anatomic and physiologic changes in other organs and tissues. *Cardiovascular and lymphatic systems.* During pregnancy the increasing needs of the growing fetus and of her own tissues throw an added burden on the mother's heart. The work that the heart does is measured by the amount of blood it expels per minute (the cardiac output). Rapid increase in the cardiac output occurs between the 9th and the 14th week of gestation. During the period from the 28th to the 30th week, when the load is heaviest, the heart of a pregnant woman is doing 25 to 30 percent more work than it was doing before pregnancy. As the time of delivery approaches, the heart's workload diminishes to some extent; when the baby is born, the load is approximately equal to what it was when the mother was in the nonpregnant state. This decrease in cardiac output and cardiac work, which occurs in spite of the continued needs of the fetus and of the maternal tissues for blood-borne oxygen and nutriments, is explained by the more efficient way that the tissues draw on the mother's blood for oxygen and nourishment during the terminal weeks of pregnancy.

The position of the heart

The position of the heart is changed to a greater or lesser degree during pregnancy. As the uterus enlarges, it elevates the diaphragm. This in turn pushes the heart upward, to the left, and somewhat forward, so that it is nearer the chest wall beneath the breast. Near the end of gestation the large uterus may raise the heart until the latter lies almost at a right angle to the long axis of the woman's body. These changes, which also bring some rotation of the heart, vary considerably in different individuals. When present to a marked degree, they may give an examining physician the erroneous impression that a normal heart is considerably enlarged. Actually, in spite of its greater workload, a healthy heart enlarges little or not at all even during the midportion of pregnancy, when the load is greatest.

Changes in the position of the heart, the greater workload, the increased volume of blood that the heart expels per beat, the decreased viscosity of the blood, and the larger amount of blood in the woman's blood vessels (discussed below) will, in many women, cause some distortion of the sounds that the physician hears when listening to a patient's heart with a stethoscope. Such distorted sounds, called "functional" murmurs (as distinguished from "organic" murmurs, which may be present when the heart is diseased), do not indicate that anything is amiss, although they may be sufficiently atypical to cause the obstetrician to refer the patient to a cardiologist for evaluation. Pregnancy sometimes produces minor changes in the electrocardiogram, but these changes are within normal limits.

Such is the ability of the heart to respond to an increased workload that even the pregnant woman with serious heart disease, given proper care and without an unexpected complication, will usually go through her pregnancy and delivery without a catastrophe. She may, however, encounter difficulty when she tries to cope with the stress of caring for her family after the baby is born.

Normal pregnancy does not increase the mother's blood pressure. Indeed, a slight lowering of the blood pressure is commonly noted during the course of the pregnancy. Any notable rise in a pregnant woman's blood pressure is reason for alertness on the part of her physician, and, if it continues to rise, for concern; it usually foretells the onset of preeclampsia (see below).

The pulse rate is a trifle more rapid during pregnancy, reflecting the more rapid heartbeat that is necessary in order to move the larger volume of blood present. The rate at which blood flows through the myriad of small blood vessels in the skin (the peripheral circulation) is accelerated during pregnancy, leading to the elevated skin temperature, the tendency to perspire, and, in part, to the redness of the palms and the tiny dilated blood vessels in some women as their pregnancies progress.

The most notable change in the circulatory system during pregnancy, other than those described in the heart, is a slowing of the blood flow in the lower extremities. With this decrease in the rate of flow there is an increase in the pressure within the veins and some stasis—stagnation—of the blood in the legs. These changes, which are believed to be caused primarily by the pressure of the uterus on the large blood vessels in the pelvis, are progressive during pregnancy and disappear after delivery. They also are thought to be caused in part by the marked increase in the amounts of the hormones estrogen and progesterone in the circulating blood. Increased venous pressure, slowing of the rate of venous flow, and partial stasis of the blood in the veins are major factors in causing the swelling of the legs and the varicose (abnormally dilated) veins of the lower legs that are commonly present near the end of pregnancy.

The lymphatic vessels of the pregnant woman's pelvis become enlarged in response to the increased amount of tissue fluid in the engorged pelvic organs. As the uterus grows in size, it presses on these channels, causing impairment of the lymphatic drainage from the woman's legs, with resultant swelling and distention of her feet and legs.

Edema in legs

Although some fluid almost invariably collects in the feet, ankles, and legs near the time of delivery, sudden swelling of the feet and legs or a notable increase in swelling may be an early signal of impending preeclampsia, a serious disorder of pregnancy that is discussed below. Generalized swelling—*i.e.,* swelling of the hands, face, and other parts of the body—is a cause for serious concern.

Respiratory tract. One would expect that, as the uterus grows larger and pushes the diaphragm up, it would interfere with breathing, but the lungs actually work as efficiently as they do in the nonpregnant state. This is due to a change in the shape of the chest cavity during pregnancy; the chest diameter increases as its height decreases, so that there is actually a slight increase in the space that the lungs occupy.

The amount of air drawn in and expelled per minute by the lungs increases progressively during pregnancy. Immediately before delivery the number of breaths per minute is approximately twice what it is after the baby is born. This, like so many of the other changes in the mother's body, is an adaptation of one of her vital functions that is necessary to supply her tissues and those of the growing fetus with increasing amounts of oxygen.

Gastrointestinal tract. A number of alterations, often causing more or less distress, occur in the physical condition and functions of the gastrointestinal tract during pregnancy.

Disturbances of the sensations of taste and smell, relatively common during early months of gestation, are often accompanied by a dislike of odours and a distaste of foods formerly found to be agreeable. The inflammation of the mouth and gums that some pregnant women complain of is more often caused by poor oral hygiene, by vitamin deficiencies, or by anemia than by the pregnancy itself.

Hydrochloric acid and pepsin, adequate amounts of which are necessary for satisfactory digestion, are produced by the stomach in decreased amounts during pregnancy. This decrease in the amount of acid in the stomach may explain some of the otherwise inexplicable anemias that occasionally occur during the course of an otherwise seemingly normal pregnancy.

During pregnancy the stomach muscles lose some of their tone and become more flabby, and the contractility of the stomach is reduced. As a result, the time it takes for the stomach to empty its contents into the intestinal tract is prolonged. As pregnancy progresses, the stomach is pushed upward; near term it lies like a flabby pouch across the top of the uterus instead of hanging downward, as it normally does, in a semivertical position. The loss of tone of the stomach muscles, the decrease in stomach acidity,

and the change in position of the stomach are conducive to the flow of intestinal contents back into the stomach.

These disturbances in gastric function are responsible, in part at least, for the intolerance for fatty foods, the indigestion, the discomfort felt in the upper part of the abdomen, and the heartburn experienced by most pregnant women at some time during their pregnancies.

The musculature not only of the stomach but also of the entire intestinal tract loses much of its tonicity. As a result, peristalsis, the series of wavelike movements of the intestines, is slowed, the length of time it takes food to pass through the intestinal tract is prolonged, and there is more or less stagnation of the intestinal contents.

Intestinal problems

Constipation and hemorrhoids that cause rectal pain and bleeding are common complaints during pregnancy. The constipation is caused by lack of tone of the intestinal tract and stagnation of the bowel contents. Pregnant women may also lose the urge to defecate because of the pressure of the uterus on the lower bowel and inhibition of a reflex stimulus, known as the gastrocolic reflex, from the stomach to the rectum. The latter mechanism, which depends on normal stomach function, is responsible for the increased activity of the lower bowel that follows increased stomach activity, such as that induced by eating. It is this reflex that causes many persons to feel a desire to defecate within an hour or so after eating a full meal. Hemorrhoids—greatly enlarged or varicose veins in the lower rectum—that appear during pregnancy are due to constipation, to stasis of blood in the pelvic veins, and to pressure by the enlarging uterus on the blood vessels in the pelvis.

Liver. The liver, which plays an essential role in many of the vital processes—processes as diverse as participating in the metabolism of nutriments and vitamins and the elimination of the waste products of metabolism—changes anatomically and functionally during pregnancy to meet the added load placed on it by the maternal organism, the enlarging uterus, and, to a lesser extent, the growing fetus.

The liver's ability to synthesize proteins and to supply minerals and nutriments is augmented in response to the increased requirements of the mother's tissue and the fetus. The liver adjusts to the greatly augmented amounts of hormones circulating in the mother's blood during pregnancy. It helps to dispose of or detoxify the larger amounts of waste material produced by the metabolic processes in the growing fetus, the enlarging uterus, and the mother's tissues. Furthermore, the blood vessels in the liver enlarge to accommodate the larger amount of blood in the mother's blood vessels. At the same time, the liver must compensate for the larger number of circulating red blood cells.

In response to these demands, the liver increases in size and weight, and its blood vessels become larger, but otherwise its anatomic structure changes relatively little during pregnancy.

The hormones produced by the placenta and the metabolic changes in the maternal organism, rather than the fetus, are the factors responsible not only for the increased work the liver does but also for many of the physical and functional alterations that appear during gestation.

Changes in bladder function

Urinary tract. Changes that take place in the bladder and the urethra during pregnancy are attributable to relaxation of the muscles supporting these structures, to change in position, and to pressure.

The uterus lies over the bladder and presses upon it during early pregnancy. Later the uterus rises out of the pelvis. As the uterus grows larger and moves upward, the bladder is pushed forward and pulled upward. The urethra, the tube through which urine is discharged from the bladder, is stretched and distorted. As these distortions take place, the wall of the bladder becomes thickened, the blood vessels become enlarged, and fluid collects in the tissues forming the wall of the bladder. The results are swelling, stasis of blood in the blood vessels, and some mechanical inflammation of the bladder wall.

The woman is likely to urinate frequently during the early months of pregnancy when the heavy uterus presses on the bladder. Frequent urination is less common during midpregnancy, but it recurs after the baby descends into the pelvis near the time of delivery. As the bladder and urethra are pulled upward and distorted by the growing uterus, the stretched muscles that control urination are less efficient, and the woman may lose some urine involuntarily when she coughs, sneezes, or laughs; this is known as stress incontinence.

The swelling, mechanical inflammation, and stasis of blood in the blood vessels of the bladder near the end of pregnancy are conducive to bladder infection, a symptom of which is pain on urination. A microscopic examination of the urine is necessary to differentiate between the effect of pregnancy on bladder function and the symptoms caused by a bladder infection. An untreated bladder infection may lead to serious urinary tract troubles later.

Structure and functioning of ureters

Changes in the structure and function of the ureters, the two rubbery, spaghetti-like tubes that carry urine from the kidneys to the bladder, are present in 80 percent of all pregnancies. As pregnancy progresses, each ureter becomes larger, so that it lies in multiple broad curves rather than forming an almost straight line downward from the kidney. In addition, both ureters, but particularly the right one, become greatly dilated, so that the urine flows very slowly or collects in them.

The funnellike part of the kidney, called the kidney pelvis, also becomes dilated. With this dilation of the kidney pelvis and the ureters there is also a loss of tonicity or contractility in the pelvis of the kidney and the ureters. This loss of tonicity during pregnancy is similar to that mentioned in the description of the changes in the intestinal tract. Since it is the contractility of peristalsis within the ureter that propels urine downward from the kidney into the bladder, stasis of urine in the ureter is accentuated during the pregnancy. In the nonpregnant state the hydrostatic pressure in the kidney is greater than that in the bladder; during pregnancy the situation is reversed. This change of pressure further increases the stasis of urine in the ureter and kidney pelvis. As a result, bladder infections are more serious during pregnancy, because they are more likely to involve the kidney.

After delivery the ureters rapidly return to their normal condition.

Ordinary functioning of kidney; changes in pregnancy

The kidney of a healthy person selectively filters and secretes water, sodium, potassium, chlorides, protein, and other substances from the blood. It then reabsorbs water and essential elements in amounts that are needed to maintain the fluid, electrolytic, and other chemical balances in the body. It also filters waste products of metabolism from the blood and excretes them in the urine. During pregnancy the kidney continues to carry on these functions. The workload placed on it, however, is greater because of the increase in the amount of water and blood and in the rate of metabolism during gestation.

In early pregnancy, secretion of large amounts of dilute urine of decreased acidity, together with pressure of the uterus on the bladder, causes frequency of urination and nocturnal voiding. Less urine is excreted toward the end of pregnancy. The storage of large amounts of nitrogen, as part of the metabolism of proteins, causes a decrease in the urinary excretion of urea and of total nitrogen during gestation.

Although many healthy pregnant women occasionally show a trace of protein (albumin) in their urine, the detection of even small amounts of protein in the urine is a cause for alertness on the part of a physician, because anything more than an extremely small amount may be the first signal of impending preeclampsia or kidney disease, both of which are serious complications.

The kidney's ability to reabsorb sugar (glucose) is lower during pregnancy, and for this reason many pregnant women have transient periods during which their urine contains small amounts of glucose; such women have unimpaired ability to metabolize carbohydrates and have normal sugar levels in the blood. Glucose in the urine also may be the first sign that a person has diabetes mellitus, however; consequently, a pregnant woman whose urine contains traces of glucose is tested to make sure that she can metabolize sugar normally.

The preceding discussion of kidney function illustrates the need for a pregnant woman to be under a health-

care provider's care, an essential part of which is periodic examination of her urine for protein, sugar, pus, bacteria, and other abnormal constituents.

Increase in blood volume

Blood. The total amount of blood in a pregnant woman's body has increased by approximately 25 percent by the time of delivery. The increase is accounted for by the augmented volume of blood plasma (the liquid part of the blood), which is caused by fluid retention, plus an increase in the total number of red blood cells. Additional blood is needed to fill the large vessels of the uterus. Also, more blood is required to carry the oxygen and nutriments needed by the fetus and the maternal tissues and to carry away waste products. Furthermore, it is a protective reserve in case of hemorrhage during delivery.

During pregnancy the blood-forming organs, such as the bone marrow, make more erythrocytes, or red blood cells, which carry iron and oxygen. Despite this, there is usually a decrease in a pregnant woman's blood cell count—the number of red cells per cubic millimetre of blood—because the amount of blood plasma increases approximately 30 percent, while the total number of red blood cells increases by only about 20 percent. This results in apparent anemia. With these changes, the viscosity of the blood decreases and the hematocrit, which measures the relative amounts of liquid and solid constituents in the blood, is lower. Usually there is a moderate increase in the number of white blood cells per cubic millimetre during early pregnancy; this increase disappears during the latter part of pregnancy.

If a pregnant woman is otherwise healthy and receives adequate available iron for the production of hemoglobin, her red blood cell count does not ordinarily fall below 3,750,000 cells per cubic millimetre, her hemoglobin below 13.5 grams per 100 cubic millimetres of blood, and her hematocrit below 35. (Normal values for nonpregnant women are 4,200,000–5,400,000 cells, 13.8–14.2 grams hemoglobin, and 37–47 hematocrit.) Physicians usually make blood counts for their pregnant patients every two months because of the need for repeated evaluation.

Endocrine system. Most of the endocrine glands become larger, and some display alterations in function, during pregnancy; they all revert to a normal state after delivery.

Pituitary, thyroid, pancreas, and adrenals

The anterior lobe of the pituitary gland increases in size during pregnancy, but the production of pituitary gonadotropins, the gonad-stimulating hormones, ceases soon after the placenta begins to produce chorionic gonadotropins. The pituitary continues to secrete the hormones that stimulate the other endocrine glands. Near term, as the mother's estrogen level drops, a milk-stimulating hormone, prolactin, is produced by the pituitary. The posterior lobe of the pituitary gland does not change in size or weight during pregnancy.

The thyroid gland enlarges moderately, but there is no true increase in thyroid function during gestation. The parathyroid glands also increase in size during pregnancy but presumably are not otherwise affected by it.

The part of the pancreas that secretes insulin, the islets of Langerhans, becomes larger. Whatever increase in function is displayed may be assumed to be a balanced response to the body's demand for the products of carbohydrate metabolism. The level of plasma insulin or of insulin-like substances in the plasma is higher during pregnancy, and the destruction of insulin is also more rapid.

The blood and urinary levels of 17-hydroxycorticosteroids, hormones that affect protein, fat, and carbohydrate metabolism and that are produced by the adrenal glands, rise during pregnancy; but there is no increased effect from the hormones, because their higher level is more than offset by the increased levels of transcortin, a protein that inactivates them.

As gestation progresses, there is an elevation in the secretion of aldosterone, an adrenal hormone that plays a role in the retention of salt and water in the body. It has been suggested that this is a protective mechanism to counterbalance the tendency for progesterone to cause the excretion of sodium ions in the urine.

Skin. Pregnancy usually causes an increase in the secretion of the oil and sweat glands in the skin. Body odours may become more pronounced. Many women notice that their hair becomes thinner and drier and their nails more brittle. Others may develop an increased amount of facial and body hair. The "mask of pregnancy" seen particularly in brunettes is a deposit of brownish pigment in the skin of the forehead, the cheeks, and the nose. Puffiness and thickening of her skin may cause the pregnant woman's face to appear coarse and almost masculine. Increased pigmentation, particularly of the smooth skin about the nipples (the areolas of the breasts) and the vulva, is almost universal.

Changes in sweat glands and skin colour

Bright red discoloration of the palms of the hands and tiny spiderweb-like red blood vessels in the skin of the arms or face are not unusual during pregnancy. Many of these changes are thought to be associated with the greatly increased levels of estrogen in the mother's bloodstream. Most of the changes disappear after delivery.

"Stretch marks," which appear on the breasts and abdomen during pregnancy, are due to the tearing of the elastic tissues in the skin that accompanies enlargement of the breasts, distention of the abdomen, and the deposition of subcutaneous fat. They are pink or purplish red lines during pregnancy. The lines become permanent scarlike marks after delivery. Some women never develop stretch marks despite bearing several children; others lose most of the tone in their skin after one pregnancy. Stretch marks cannot be considered evidence that a woman has borne a child, however, because they sometimes are seen in women who have not been pregnant.

Metabolic changes. Metabolic changes during pregnancy are among the many adjustments that the mother's organs make to meet the requirements created by the increase in her own breast and genital tissues and the growth of the conceptus (the fetus and afterbirth). In addition, reserves must be established to meet the demands that will be put on her body during pregnancy, delivery, and the postdelivery period.

The basal metabolic rate. The amount of oxygen consumed is an index of the pregnant woman's metabolism when she is at rest—her basal metabolism. The rate begins to rise during the third month of pregnancy and may double the normal rate (+10 percent) by the time of delivery. The rate rises in specific proportion to the size of the fetus and represents the effects of the mother's activities plus those of the fetus and the uterine structures. An elevation of the basal metabolic rate (BMR) to 20 or 25 percent during pregnancy is not an indication of an overly active thyroid gland.

Weight. The early part of pregnancy usually is accompanied by moderate weight loss caused by the woman's lack of appetite and in some cases nausea and vomiting. Between the third and the ninth month of pregnancy most women gain about 9 kilograms (20 pounds) or more. Ideally, during pregnancy, body weight is gained at the rate of about 0.5 kilogram (1 pound) per week for a total of not more than 9 to 11.5 kilograms (20 to 25 pounds). In an average pregnancy the infant, the afterbirth, and the fluid in the uterus weigh about 4.5 kilograms (10 pounds). The uterus and the breasts together weigh approximately 2.25 kilograms (5 pounds). The remaining 2.25 kilograms consist of stored fluids and fat. Weight gain exceeding 11.5 kilograms usually represents fat and fluids that are in excess of the reserve requirements for a normal pregnancy. A woman loses approximately 7 kilograms (15 pounds) at delivery, and another 2.25 kilograms of stored fluid are eliminated as the uterus shrinks. She does not lose many additional kilograms during the weeks following the delivery of the baby unless she limits her caloric intake. Fat stored during pregnancy is lost more slowly than stored fluids, proteins, and carbohydrates.

Weight gain

Excessive weight gain during pregnancy is a matter of concern for both the patient and the doctor. Although it may be only the result of overeating, it may be caused by a disturbance in metabolism and by an abnormal retention of fluids and salts. In the latter instance it may be the first sign of preeclampsia.

Protein. During pregnancy, nitrogen, derived from the metabolism of ingested protein, is needed for growth of the fetus, the placenta, the uterus, and the mother's breasts

Changes in protein, carbohydrate, and fat metabolism

and other tissues. A considerable amount of nitrogen also is required for the increase in the mother's red cell volume and blood plasma. The fetus's demand for nitrogen is slight at first, but during the last month of pregnancy it acquires almost half of its total protein. In the process of accumulating this store and of building a reserve for the period after delivery, the woman who is on an adequate diet retains between two and three grams of nitrogen daily during her pregnancy; by term she and the fetus will have acquired approximately 500 grams (about 1.1 pounds) of nitrogen.

Carbohydrates. During pregnancy greater quantities of blood are being processed through the kidneys, but the kidneys are incapable of reabsorbing increased amounts of sugar. Consequently, a lower level of sugar in the blood is tolerated, and slight amounts of sugar are excreted in the urine. During pregnancy the level of sugar in the blood after fasting is slightly lower, probably because there is less usable insulin in the blood to regulate the sugar metabolism. Oral glucose-tolerance tests show a prolonged elevation of blood sugar after ingestion of glucose; this may be an indication that carbohydrate use is less rapid or that the absorption of glucose from the gastrointestinal tract is slower. Glucose-tolerance tests that depend on injection of the sugar solution into the veins show no difference between nonpregnant and pregnant nondiabetic women. A few women demonstrate diabetes for the first time when they are pregnant, a condition referred to as gestational diabetes. This occurs because pregnancy taxes insulin productivity in women with a marginal pancreatic islet reserve, so that diabetes may first become evident during gestation.

Fat. The total blood lipids average 600 to 700 milligrams per hundred millilitres of blood in the nonpregnant woman. They increase to approximately 900 to 1,000 milligrams per hundred millilitres of blood during the latter part of pregnancy. This increase, which involves all the lipid fractions, has not been explained, but it is worthy of notice that the gain in fat reaches its acme during the period that the fetus acquires most of its adipose (fatty) tissue.

Increases in body fluids

Water. Pregnancy is characterized by increases in the amount of body water and in the total volume of body fluid. During pregnancy between 3,500 and 4,000 millilitres of fluid (about 3.2 to 3.6 quarts) will be added to that already present in the tissues of a healthy woman. The uterus, the placenta, the amniotic fluid, and the fetus each account for approximately equal amounts. In addition to the water that increases blood volume, there is also added fluid in the mother's muscles, her pelvic soft tissues, her breasts, and her other tissues.

Toward the end of pregnancy a considerable amount of retained fluid accumulates in the woman's lower extremities. It is this fluid that produces the pitting and swelling of the legs that many normally pregnant women display during the month or two before delivery.

Retention of large amounts of electrolytes, particularly sodium, accompanies the increase in the amount of body fluids. Approximately 12 grams of sodium are retained monthly. In addition to a positive sodium balance, there is a positive chloride and potassium balance during pregnancy. As a result, additional water is required to maintain the balance of the solution of sodium, chloride, and potassium in the blood, in the fluid of the spaces between the tissue cells, and within the cells themselves. Not all of the sodium, however, goes into fluid. Some of it is stored, and some replaces potassium in the cells.

A number of factors contribute to a positive sodium balance, which in turn leads to retention of fluid; these include alterations in the kidneys' excretion of sodium and water; increased retention of water in the pregnant woman's legs; the large amounts of hormones, particularly estrogen, that the placenta secretes; and the secretion of adrenal hormones, especially aldosterone. The latter, in particular, reduces the kidneys' secretion of sodium. Because sodium and water interact with each other, whatever contributes to the retention of one leads to the retention of the other. Generalized swelling appears when the accumulation of sodium and water becomes too great.

Minerals. The pregnant woman's reserves and intake of iron and calcium must be enough not only for her own needs but also for those of the fetus. An increase in serum copper levels occurs during pregnancy. The mother has some phosphorus reserve but must acquire enough from her diet to supply her own tissues and those of the fetus. The use of phosphorus and that of calcium are interdependent, so that the use of phosphorus depends on the calcium intake. (J.W.Hu.)

PRENATAL CARE AND TESTING

An adequate maternal diet is necessary to ensure proper fetal development as well as to maintain the health of the mother. As discussed above, the physiological adjustments of a pregnant woman's body are significant, and nutritional requirements increase as a result. In addition to an awareness of the substances that are of benefit during pregnancy, a knowledge of which substances are harmful and should be avoided is equally important. Alcohol has been found to be teratogenic (causing developmental malformations in the fetus). Intake of large to moderate quantities of alcohol during pregnancy is responsible for fetal alcohol syndrome, which is characterized by impaired growth and development, facial abnormalities, cardiac defects, and skeletal and joint malformations. The effects of limited intake of alcohol are not as well known, but avoidance of any amount of alcohol throughout pregnancy is recommended. Smoking of tobacco during pregnancy is believed to lower the birth weight of the fetus and is also associated with placenta praevia, abruptio placentae, and elevated maternal blood pressure. Sudden infant death syndrome, delayed mental development in childhood, and spontaneous abortion also have been linked to smoking. Limiting the use of caffeine also is encouraged. While not believed to have teratogenic effects, excessive caffeine intake may account for low birth weight in infants. Over-the-counter medications as well as prescription drugs can adversely affect fetal development and should not be taken unless a health-care provider is consulted.

Ultrasound. The use of high-frequency sound waves to produce a graphic image of the growing fetus—ultrasonography—is becoming a ubiquitous tool in prenatal medicine, furnishing information on the morphological and functional status of the fetus. It is commonly used to estimate the gestational age of the fetus, identify fetal number, assess growth, determine fetal heart activity, and provide a general survey of fetal anatomy. The presentation of the fetus and placenta and the volume of amniotic fluid also can be determined using ultrasound. In most European countries an ultrasound scan is routinely included in obstetric examinations, but, although it is widely used in the United States and Canada, its inclusion in standard prenatal evaluations has not been recommended. This reluctance is based on the lack of clear evidence that this procedure has no negative effects. Theoretical risks are involved because of the invasive nature of this technique (*i.e.,* sound waves are reflected off tissues). Studies to date, however, have revealed no evidence of tissue damage when diagnostic ultrasound is used, and the benefits of this procedure seem to outweigh the risks.

Amniocentesis. In the procedure of amniocentesis, amniotic fluid is aspirated (withdrawn) from the uterus by a needle inserted through a woman's abdomen, using ultrasound to circumnavigate the fetus and placenta. Spinal cord defects and a host of genetic abnormalities such as Down syndrome and autosomal recessive diseases such as Tay-Sachs disease and cystic fibrosis can be screened for by amniocentesis. It can also be used to determine the sex of the fetus and identify sex-linked diseases. Not all birth defects, however, can be detected by this procedure. This test is generally performed about the 16th week of pregnancy, and results take several weeks to obtain. Of the potential risks associated with this procedure, the most significant one is that of fetal loss, which may result from disruption of the placenta.

Chorionic villi sampling. The technique of retrieving a sample of villi from the chorion (outer embryonic membrane) within the uterus is similar to amniocentesis but can be carried out much earlier in pregnancy, between

the 8th and 12th week of gestation. The test can be performed through either the abdomen or the vagina and cervix. The latter technique is carried out using ultrasonic visualization, and a thin catheter is inserted through the vagina into the uterus; a sample of villi from the chorion is then extracted and examined. If unfavourable results are received, termination of the pregnancy can be accomplished at an earlier stage than would be possible with amniocentesis. This procedure does carry a slightly higher risk of fetal loss than does amniocentesis, possibly because it is carried out at an earlier stage in fetal development. With this technique there is also concern that fetal limb reduction or malformation may result, but reports are inconclusive.

Alpha-fetoprotein screening. Shed by the yolk sac and fetal liver, alpha-fetoprotein can be used to screen for neural tube defects such as anencephaly and spina bifida (developmental abnormality in which spinal cord is not fully enclosed). The measurement of elevated levels of alpha-fetoprotein in a woman's blood between the 16th and 18th weeks of pregnancy are associated with this abnormality. Because other circumstances such as multiple pregnancies, underestimation of gestational age, and fetal death are associated with high levels of alpha-fetoprotein, ultrasound should be used to help rule out these different causes. Abnormally low levels of alpha-fetoprotein have been linked to a significant incidence of Down syndrome. A high rate of false-positive results is associated with this test, and so it is not recommended routinely. This procedure has been reserved primarily for those women with a family history of neural tube defects. (Ed.)

Abnormal changes in pregnancy

ECTOPIC PREGNANCY

An ectopic pregnancy is one in which the conceptus (the products of conception—*i.e.*, the placenta, the membranes, and the embryo) implants or attaches itself in a place other than the normal location in the lining of the upper uterine cavity. The site of implantation may be either at an abnormal location within the uterus itself or in an area outside the uterus. Ectopic pregnancies outside the uterine cavity occur about once in every 300 pregnancies. They are one of the major causes of maternal deaths.

Types of ectopic pregnancy

Normally an ovum or egg passes from the ovary into the tube, is fertilized in the tube, and moves downward into the uterus. It buries itself in the lining of the upper part of the uterine cavity. It may pass farther down and attach itself to the lining of the mouth of the uterus (the cervix), creating a cervical pregnancy. These are rare and cause severe vaginal bleeding; the conceptus is expelled or discovered within a few months after implantation.

If a conceptus attaches itself to the lower part of the uterine cavity, it is a low implantation. When a low implantation occurs, the placenta grows over the cervical opening, in a formation called a placenta praevia. This causes the woman to bleed, often profusely, through the vagina, because the placenta tears as the cervix begins to open during the latter part of pregnancy.

When the fertilized egg implants in the narrow space or angle of the uterine cavity near the connection of the uterus with the fallopian tube, it is called an angular pregnancy; many angular pregnancies terminate in abortions; others go to term but are complicated because the placenta does not separate properly from the uterine wall after the birth of the baby. An angular pregnancy differs from a cornual pregnancy, which develops in the side of a bilobed or bicornate uterus.

Implantation in the narrow part of the fallopian, or uterine, tube, which lies within the uterine wall, produces what is called an interstitial pregnancy. This occurs in approximately 4 percent of ectopic pregnancies. An interstitial pregnancy gradually stretches the wall of the uterus until—usually between the 8th and 16th week of gestation—the wall ruptures in an explosive manner and there is profuse bleeding into the abdomen.

Most persons associate ectopic pregnancies with tubal pregnancies, because most ectopic pregnancies occur in the uterine tubes. The tube beyond the uterus has three parts: the isthmus, a narrow section near the uterus; the ampulla, which is wider and more dilatable; and the infundibulum, the flaring, trumpetlike portion of the tube nearest the ovary. A tubal ectopic pregnancy is designated by the area of the tube in which it is implanted. An isthmic pregnancy differs from one in the ampulla or infundibulum because the narrow tube cannot expand. Rupture of the affected tube with profuse intra-abdominal hemorrhage occurs early, usually within eight weeks after conception.

Ampullar pregnancies, which are by far the most common, usually terminate either in a tubal abortion, in which the embryo and the developing afterbirth are expelled through the open end of the tube into the abdomen; by a tubal rupture; or, less commonly, by absorption of the conceptus.

Sometimes the tube ruptures into the tissues attaching it to the wall of the pelvis, producing an intraligamentous pregnancy. Rarely, the embryo is expelled into the abdomen and the afterbirth remains attached to the tube; the embryo lives and grows. Such a condition is referred to as a secondary abdominal pregnancy. Primary abdominal pregnancies, in which the fertilized egg attaches to an abdominal organ, and ovarian pregnancies are rarer still.

Causes of tubal pregnancies

It is generally believed, but not proved, that most tubal

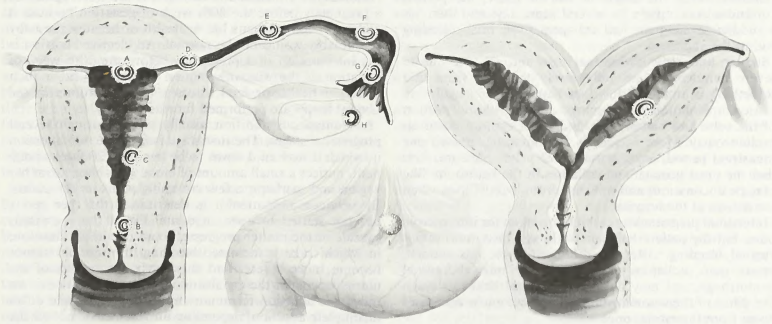


Figure 34. *Sites of implantation.*
(A) Normally in the upper part of the corpus. (B) In the lower corpus with later development of a placenta praevia. (C) Low implantation. (D) In the interstitial portion of the uterine tube. (E) In the isthmic part of the tube. (F) In the ampulla of the tube. (G) In the infundibulum of the tube. (H) In the ovary. (I) On the peritoneum with development of an abdominal pregnancy. (J) In the cornu of a double or rudimentary horn of a uterus.

pregnancies are caused by scars, pockets, kinks, or adhesions in the tubal lining resulting from tubal infections. The infection may have been gonorrhea; it may have occurred after an abortion, after the delivery of a baby, or after a pelvic surgical operation; or it may have been caused by appendicitis. Kinking, scarring, and partial adhesions of the outside of the tube may be the result of inflammation following a pelvic operation or of an abdominal inflammation. Tubes, defective from birth, may be too small for the passage of the conceptus or may be pocketed or doubled with one tubal half forming a blind pocket. There may be areas in the tubal lining that behave like the lining of the uterus (they show a decidual reaction that is conducive to implantation) so that they offer a favourable spot for the fertilized egg to implant. Pelvic tumours may distort the tube and obstruct it so that the conceptus cannot move downward. Theoretically, endocrine disturbances may delay tubal motility.

Whatever the cause, when a tubal implantation occurs, it may be assumed that either migration of the fertilized egg within the tube was delayed by an extrinsic factor so that the egg grew to the point where it should implant or that the mechanism for implantation within the egg itself was prematurely activated in the tube. One or the other of these causative factors can sometimes be seen when a woman is operated upon for an ectopic pregnancy. In a great number of cases, however, no tube abnormality can be found. There is no satisfactory explanation for most abnormal implantations in the uterus, although defective uterine structure has been noted in some cases.

Primary abdominal and ovarian pregnancies can best be explained by a mechanism in which the fertilized ovum is swept out of the tube by a reverse peristalsis of the tube, but it is quite possible that, in rare instances, the ovum and spermatozoa meet and fertilization and implantation take place within the abdomen.

Signs of ectopic pregnancy

Ectopic pregnancy is frequently mistaken for other disorders. Typically, but not invariably, the woman who has an ectopic pregnancy in the ampullar part of the tube will have missed one or two menstrual periods. She need not have other symptoms of pregnancy. She has felt enough discomfort in the lower part of her abdomen to lead her to consult a physician. She has had recurrent episodes of rather light, irregular bleeding from the vagina. She has felt weak or faint at times. The signs of pregnancy are not likely to be present, and results of a pregnancy test are more often negative than positive. The physician, on pelvic examination, feels a tender, soft mass in one side of the pelvis. At this stage the differentiation must be made between an ectopic pregnancy and an intrauterine pregnancy with abortion, acute appendicitis, intestinal colic, inflammation of a fallopian tube, and a twisted ovarian tumour. Unless the diagnosis can be made, the patient continues to complain for several more days and then has a sudden severe pain and collapses from brisk bleeding within the abdomen.

Sudden and acute abdominal pain and collapse due to severe hemorrhage are only rarely the first signs that something is amiss. If this does happen, it is usually because implantation has occurred in the isthmic portion of the tube and hemorrhage and tubal rupture occur simultaneously. More frequently, a woman has missed one menstrual period, has a sensation of pelvic pressure, feels that she must urinate, and collapses in the bathroom. She may be unconscious and pulseless from loss of blood when she arrives at the hospital.

Interstitial pregnancies

Interstitial pregnancies are often mistaken for intrauterine ones, but the patient has pain and may have intermittent vaginal bleeding. After several months she has sudden, severe pain, collapses from a massive intra-abdominal hemorrhage, and may die before surgical help can reach her. Most of the women who die from ectopic pregnancies do so from interstitial ones.

Combined pregnancies, in which there is an ectopic pregnancy and a normal one in the uterus, or a fetus in each tube, have occurred and have compounded the difficulty in making a diagnosis. In a number of instances, the ectopic conceptus has been removed without complications, and the uterine fetus has progressed to term.

Not all ectopic pregnancies end with a catastrophic hemorrhage and collapse. In a few instances tubal, abdominal, and broad ligament pregnancies have gone on until a living baby was obtained at the time of operation. In other cases the fetus died and, if very young, was resorbed; in others, when the fetus was larger, death was followed by absorption of the fluid in the sac, and the fetus was gradually converted into a more or less mummified mass. Some ectopic pregnancies of this type have caused no symptoms and have been carried by women for years. Undoubtedly many ectopic pregnancies that are in an early stage when they are expelled emerge through the open end of the uterine tube, are resorbed, and are never recognized.

Once diagnosed, the treatment of ectopic pregnancies outside the uterine cavity is almost always a matter of prompt surgical intervention with proper attention to replacement of blood and fluid.

ABORTION

Abortion is the termination of a pregnancy before the infant can survive outside the uterus. The age at which a fetus is considered viable has not been completely agreed upon. Many obstetricians use either 21 weeks or 400–500 grams (0.9–1.1 pounds) birth weight as the baseline between abortion and premature delivery, because few infants have survived when they weighed less than 500 grams at birth or when the pregnancy was of less than 21 weeks' duration. Generally speaking, the fetus has almost no chance of living if it weighs less than 1,000 grams (2.2 pounds) and if the pregnancy is of less than 24 weeks' duration. In one effort to resolve the matter, the American College of Obstetricians and Gynecologists has defined abortion as the expulsion or extraction of all (complete) or any part (incomplete) of the placenta or membranes, with or without an abortus, before the 20th week (before 134 days) of gestation. Early abortion is an abortion that occurs before the 12th completed week of gestation (84 days); late abortion is an abortion that occurs after the 12th completed week but before the beginning of the 20th week of gestation (85–134 days).

In the past the word abortion usually meant to nonmedical persons the elective interruption of a pregnancy, whereas "miscarriage" indicated a spontaneous expulsion of the uterine contents. The term miscarriage is seldom used medically.

Spontaneous abortion is the expulsion of the products of conception before the 20th week of gestation without deliberate interference. As a general rule, natural causes are responsible for loss of the pregnancy. An induced abortion is the deliberate interruption of a pregnancy by any means before the 20th week of gestation. In medical terminology an abortion may be therapeutic or elective (voluntary). A therapeutic abortion is the interruption of a pregnancy before the 20th week of gestation because it endangers the mother's life or health or because the baby presumably would not be normal. An elective abortion is the interruption of a pregnancy before the 20th week of gestation at the woman's request for reasons other than maternal health or fetal disease. Most abortions in the United States are performed for this reason.

Stages of spontaneous abortion

A spontaneous abortion usually passes through several progressive stages. The first stage is a threatened abortion in which a woman, known to be less than 20 weeks pregnant, notices a small amount of bloody discharge from her vagina and, perhaps, a few cramping pains in her uterus. By pelvic examination it is determined that her cervix has not started to open or dilate. Either the symptoms subside or the matter progresses to an inevitable abortion, in which there is increased bleeding, the uterine cramps become more severe, and the cervix, or mouth of the uterus, opens for the expulsion of the uterine contents. An inevitable abortion terminates either as a complete or an incomplete abortion, depending on whether or not all the products of gestation are expelled. The process may start abruptly with pain and profuse bleeding and be over in a few hours, or it may go on for days with only a modest loss of blood. Spontaneous abortions early in pregnancy tend to be complete. When the pregnancy is further advanced, it is more likely to be incomplete. Usually the physician

removes the retained tissue in the uterus surgically when there is an incomplete abortion.

If the fetus dies and is retained in the uterus for eight weeks or longer, the condition is referred to as a missed abortion. Women who lose three or more consecutive pregnancies of less than 20 weeks' duration are said to suffer from recurrent abortion. An infected abortion is an abortion associated with infection of the genital organs.

Approximately 15 percent of all clinically evident pregnancies terminate in spontaneous abortion. A much higher rate of early pregnancy loss—more than 40 percent—is believed to occur. Some are lost so early that the woman and her physician are not sure whether she aborted or had a menstrual period that was slightly delayed, particularly heavy, and more painful than usual. The majority occur between the 6th and the 12th week after conception. Modifications in the abortion laws in several countries, including the United States, have greatly increased the number of requested abortions; it is believed that in some areas the number of abortions exceeds that of babies delivered alive.

Causes of abortion

At least half of all spontaneous first-trimester abortions have been found by karyotyping (examination of chromosome characteristics) to have a chromosomal abnormality. Some of these genetic mistakes are caused by abnormal characteristics carried in the egg or sperm or by the failure of normal rearrangement of the chromosomes to occur after the egg and sperm unite. It has been shown in animals that disturbances in the transportation of the fertilized egg to the uterus may cause premature or delayed implantation of the conceptus; fertilized eggs that are too young or too old tend to abort. Inadequate secretion of the ovarian hormones estrogen and progesterone, needed for the development of the newly fertilized egg, may cause failure of the lining of the uterus and its secretions to sustain the young embryo. Later, failure of the placenta to take over the hormone-producing function of the ovary may adversely affect the growth of the uterus and its contractility. X rays in large doses, radium, and certain drugs may cause abortion because they damage embryonic tissues. Abnormal development of the mother's uterus may make it impossible for it to retain the pregnancy.

Late abortion is sometimes caused by the weakness of the cervix or by fetal death following knotting of the umbilical cord. Uterine tumours may cause abortion because they increase uterine irritability or create an unfavourable environment for embryonic growth. In most instances in which psychological factors allegedly caused an abortion, examination of the baby and of the afterbirth have shown defects in one or both that had occurred before the mother had suffered her emotional disturbance. Physical injury to the mother is a causative factor in only one in a thousand abortions. Abortions thought to be caused by automobile accidents, falls, kicks, and so forth are often the result of deleterious changes in the fetus and sac that occurred before the injury. Systemic diseases may play a role in causing an abortion. This is particularly true of acute infectious diseases with high fever and bacteria in the bloodstream, or of diseases such as pneumonia, in which there is a marked reduction in the supply of oxygen to the fetus. Heart disease, kidney disease, diabetes, high blood pressure, and other chronic diseases may be associated with premature birth and fetal death after the 21st week but do not ordinarily cause abortions.

Perhaps 3 percent of threatened abortions are prevented by rest and hormonal therapy. Most abortions are inevitable because the fertilized egg is abnormal; these cannot be controlled medically. Many women who suffer from recurrent abortion respond well to treatment; in some of these cases corrective surgery is necessary. An early spontaneous abortion without infection is rarely followed by ill health when the affected person receives proper medical treatment. Infected abortions, many the result of elective interruptions of pregnancy, have caused chronic pelvic distress and, in some cases, sterility.

Effects of abortion

SYSTEMIC DISEASES AND PREGNANCY

Heart disease. Heart disease occurs in approximately 1 percent of pregnant women. It is first as a cause of maternal deaths among the disorders incidental to pregnancy and fourth, after hemorrhage, preeclampsia, and infection, as a cause of all maternal deaths. Rheumatic heart disease is the most prevalent type. Congenital heart disease accounts for approximately one-fourth of the cases.

A number of factors, including her response to physical activity, a history of heart failure, the type of heart disease that she has, and her age, are used in estimating how a particular woman will fare during pregnancy and labour. A person whose heart disease causes no limitation of normal physical activity will usually go through a normal pregnancy and delivery without notable difficulty, provided that she avoids undue physical activity, has sufficient rest, avoids infections, and is under the care of an obstetrician and a cardiologist who are on the alert for signs of early heart failure. Such a person will not face an appreciably increased risk, and her heart disease will not be affected by the pregnancy.

The woman whose physical activity is limited to some extent because it causes undue fatigue, shortness of breath, heart palpitation, or heart pain, but who has never experienced heart failure, will seldom suffer heart failure if she follows a strict regimen outlined by her physicians throughout the pregnancy, the labour, and the puerperium (the period immediately after childbirth) and if she does not experience a complication of pregnancy or of her heart disease. A diseased heart, although able to carry the load put on it by pregnancy, may not be able to stand up under an additional burden. This is particularly true if the pregnant woman gains an excessive amount of weight; if she develops preeclampsia, kidney disease, pulmonary disease, or an infection; or if she overworks physically, is subjected to sudden severe emotional stress, or becomes anemic. The possibility that a woman with serious heart disease will have heart failure is greater if she is over 35 years of age.

Management of pregnancy preceded by heart failure

More than half of the women who have suffered from heart failure before they became pregnant do so again during pregnancy, usually between the fifth and the ninth month, when pregnancy throws the greatest workload on the heart. Because so many women with a history of previous heart failure have difficulties during pregnancy, many obstetricians and cardiologists restrict the physical activity of such women and try to keep them in the hospital and under close medical surveillance. Some women with serious heart disease are kept in bed in the hospital throughout the course of the pregnancy and thus avoid heart failure. Cardiac surgery during the first few months of pregnancy, although a hazardous procedure, has lessened the necessity for prolonged bed rest in some cases and materially improved the prognosis in others.

Women with serious heart disease often deliver prematurely, and their labours are often short and their deliveries easy. There is an increase in fetal mortality because many pregnancies are interrupted and because many of the babies of women with heart disease are born prematurely. Babies who are not born prematurely are not notably different from those of normal mothers.

Endocrine diseases. *Diabetes.* Before insulin was available, most diabetic women were sterile, or, if they became pregnant, aborted. Half of the babies and one-fourth of the mothers died if they went to term. Today, if they are adequately supervised, less than 1 percent of pregnant diabetic women die of diabetes during pregnancy or the puerperium. Diabetic women do suffer from an increased incidence of preeclampsia, infections, and hydramnios (excessive amniotic fluid). Abnormalities of labour are increased because the babies tend to be unusually large, and congenital abnormalities of the fetus are more common, as is hydramnios; hydramnios is a problem in 25 percent or more of diabetic women.

Untreated diabetes is associated with a high incidence of fetal defects, abortion, stillbirths, premature labour, and excessively large babies. Even with diet and insulin, more than 50 percent of the babies delivered by diabetic women weigh over eight pounds at birth. Even though they appear healthy at birth, many of them are not as strong as smaller babies whose mothers are not diabetic. Fetal loss is greater if the mother became diabetic in childhood, if she

has been diabetic for a long time, or if she has vascular or kidney disease.

Pregnancy frequently has an adverse effect on diabetes, and diabetes may first become evident during pregnancy. There is a tendency for the carbohydrate metabolism of the diabetic patient to be upset. Most diabetics need more insulin during gestation; a few, for reasons not understood, need less. The changing condition from day to day makes some diabetics, who have no problem maintaining a balance when they are not pregnant, difficult to treat. Even so, adequate medical supervision can bring most diabetics and their babies safely through pregnancy.

Thyroid disease. Simple goitres that are not associated with a change in the amount of thyroid hormone in the mother's blood do not affect pregnancy, nor does pregnancy affect the thyroid in such a case. An inactive or too active thyroid gland, if not adequately treated during pregnancy, may be associated with an increased incidence of abortion. In the few cases in which persons with untreated myxedema, a severe form of hypothyroidism (deficiency of thyroid hormone), have conceived and gone to term, there has been an increased incidence of congenital anomalies of the fetus. Pregnancy and hyperthyroidism (overabundance of thyroid hormone) seem to have no adverse effects on each other.

Pituitary disorders. Most persons with pituitary hypofunction fail to ovulate because their pituitary glands do not produce the gonadotropic hormones necessary for stimulation of the ovaries. Most of these persons also suffer from a lack of hormones from their other endocrine glands because these, too, lack stimulation by the pituitary. A few persons with hypopituitarism have, nevertheless, become pregnant. Their condition is better when they are pregnant because their placentas produce many of the hormones that their endocrine glands, lacking pituitary stimulation, do not ordinarily secrete.

Adrenal glands. Women suffering from adrenal gland insufficiency are not likely to become pregnant. If they do so, they have a greater tendency to suffer from circulatory disturbances and carbohydrate, electrolyte, and fluid imbalances because of the important role the adrenal glands play in the metabolism of water, sodium, potassium, chlorides, and glucose. Such patients and their babies do well if they receive hormonal therapy during gestation.

The increased secretion of adrenal hormones that occurs with hyperplasia of the adrenal cortex (enlargement of the outer layer of the adrenal gland, also called Cushing's syndrome) usually inhibits ovulation. A number of women with this disorder, after treatment with cortisone, have conceived, gone to term, and delivered normal children. Cushing's syndrome complicated by pregnancy is rare; the few cases reported have been associated with a high incidence of severe high blood pressure.

The maternal death rate is approximately 50 percent, and the death rate of the child immediately before or after birth is approximately 40 percent when pheochromocytoma (a type of adrenal tumour associated with, among other things, high blood pressure) complicates pregnancy.

Urinary tract diseases. Infections of the urinary tract are more frequent during pregnancy, and women who have acute infections of the bladder and kidneys while pregnant have a higher incidence of premature labour. This is in accord with the fact that pregnant women with any type of acute infection tend to deliver prematurely. Many women with pyelonephritis (infection of the kidney) in one pregnancy will enter a second pregnancy with bacteria already in the urinary tract, although they are asymptomatic. These women have a greatly increased chance of developing acute urinary tract infections during their prepartum course and have some risk of eventually developing serious kidney disease. Glomerulonephritis, a kidney disease that affects the clusters of capillaries in the nephrons, the functioning kidney units, usually is preceded by infection with streptococcus organisms. The incidence of abortion and of premature delivery is increased among women in whom the condition develops during pregnancy. If the glomerulonephritis has become chronic, the fetus may not survive and the mother's life may be endangered by kidney failure.

Kidney and bladder infections

Healed tuberculosis of the kidney is not a contraindication to pregnancy if the disease has been quiescent for three years or longer and kidney function is normal. If tuberculosis of the kidneys is present but without symptoms, pregnancy may cause it to become active. If this happens, and if the infection is limited to one kidney, there is an increased danger that the opposite kidney will become infected in some way. The interference with the flow of urine that is characteristic of pregnancy is an important factor in the development of such infections. The accepted treatment when tuberculosis was present in one kidney during pregnancy formerly was therapeutic abortion followed by removal of the tuberculous kidney. This procedure is now avoided in some instances because of the effectiveness of the antituberculotic drugs that have been developed.

It is sometimes necessary to remove a person's kidney because of an infection, a stone, a tumour, or tuberculosis. The remaining normal kidney has a reserve that is greatly in excess of the demands that will be made by gestation, provided that it does not become infected. Infections, impaired kidney function, congenital defects, and preeclampsia, however, are more serious for a woman with a solitary kidney than they are for the patient with a normal urinary tract.

Pulmonary disease. Pulmonary disorders have an adverse effect on pregnancy if they seriously decrease the amount of oxygen supplied to the fetus, if they make the mother desperately sick, or if they create a blood infection that is transmitted to the placenta.

An infection of the upper respiratory tract—the nose and throat—does not ordinarily disturb the course of gestation. It may be serious when it occurs in late pregnancy because of the danger that the mother will transmit disease-causing bacteria to her own genitalia or will carry virulent bacteria from her own nose and throat into the labour room and develop a blood infection after the delivery.

Respiratory infections

Epidemic influenza is associated with an increased incidence of maternal deaths. Many women who suffer from it abort or deliver prematurely. The infection may pass through the placenta and cause infection in the fetus. Pregnant women who acquire epidemic influenza are more likely to develop pneumonia than are persons who are not pregnant.

Pregnancy may increase or decrease the severity of asthma or may fail to affect it. A severe attack of asthma may be followed by abortion, but otherwise asthma does not affect pregnancy.

Pneumonia occurring during pregnancy is associated with a high rate of maternal and fetal death unless the pulmonary infection is susceptible to antibiotics or chemotherapy. The mother's cardiovascular system, already carrying the load placed on it by pregnancy, cannot sustain the added stress produced by pneumonia. The fetus often dies from oxygen starvation or from intrauterine infection.

Severe bronchitis and bronchiectasis—abnormal dilation of bronchi with some destruction of bronchial walls—may so interfere with the mother's respiration that the extra strain put on her cardiorespiratory system by pregnancy may put her life in jeopardy. If the disorders are severe enough to cause impaired pulmonary ventilation, the fetus may suffer from a lack of oxygen and may be either stillborn or delivered prematurely. Pregnancy does not adversely affect the course of these pulmonary diseases.

Pulmonary tuberculosis is not, as a rule, affected by pregnancy. This is particularly true if the patient's infection has been quiescent for several years before she becomes pregnant. Even women with active tuberculosis, if given adequate care, usually go through pregnancy without any deterioration in their pulmonary condition. This is not universally true, however, because there is a small group with active disease whose disease becomes worse during pregnancy. For that reason individual evaluation of each person is necessary.

Although there have been a few cases of infection transmitted to the fetus prenatally, the great majority of babies born of tuberculous mothers are healthy at birth.

Pregnant women who have had portions of their lungs

removed for tuberculosis, tumours, or other reasons do well provided that, before becoming pregnant, they are not short of breath with ordinary exertion. The added load of an additional pulmonary infection may not leave such persons with enough pulmonary reserve for the added burden of pregnancy; they may therefore experience difficulties if they contract pneumonia, severe influenza, or acute bronchitis during pregnancy.

Peptic ulcer; colitis; appendicitis

Gastrointestinal diseases. Women may already suffer from a gastrointestinal disease such as gastric or peptic ulcer, gallbladder disease, or ulcerative colitis when they become pregnant; or they may develop some type of gastrointestinal disturbance during the course of the pregnancy. In either event, pregnancy complicates their problems because the gastrointestinal disturbances that often accompany pregnancy may confuse the diagnosis in an individual case.

Gastrointestinal diseases have little or no effect on pregnancy. Pregnancy, on the other hand, tends to aggravate gastrointestinal disorders; the exception is gastric ulcer, which often improves because the concentration of acid in the stomach is decreased with pregnancy. Women with chronic ulcerative colitis are generally advised to avoid pregnancy until their bowel disease has been quiescent for two years; actually, since the woman's psychological reaction to pregnancy is what affects the bowel, the colitis may be made either better or worse by gestation.

Acute appendicitis, occurring during pregnancy, is often confused with other gastrointestinal complaints, and many patients' lives have been jeopardized either because they ignored the symptoms or because the diagnosis was confused by pregnancy. A diagnosis of acute appendicitis calls for immediate surgery regardless of the duration of the pregnancy or the hazard to the fetus.

Nervous system disorders. Neurological disorders and pregnancy most often are coincidental and have no effect on each other, but there are a few neurological diseases that develop during pregnancy, have a deleterious effect on it, or are adversely affected by it.

Epilepsy; poliomyelitis; polyneuritis; neuralgia

Epilepsy of unknown cause does not affect the course of pregnancy but may occur for the first time during gestation. An epileptic person may find her condition improved, aggravated, or unchanged by pregnancy; the effect of gestation cannot be foretold. There is some evidence that excessive fluid and salt retention induces epileptic seizures.

Pregnant women are more susceptible to poliomyelitis (infantile paralysis), but pregnancy does not affect the severity or the course of the disease, nor does poliomyelitis affect the course of pregnancy. If the muscles of respiration are paralyzed, the patient will have difficulty during the latter part of pregnancy, when the uterus presses upward on the diaphragm. There have been a few instances in which babies have acquired infections from the mother before birth.

Polyneuritis, a disorder of the nerves usually resulting from vitamin B deficiency, may complicate pregnancy; this is particularly likely if the patient has suffered from severe and prolonged vomiting. Polyneuritis does not affect the gestation.

Neuralgia (pain that radiates along the nerve) occurs frequently near term. It affects especially the sciatic nerve, which is compressed between the pelvic wall and the head of the fetus.

Brain injury; psychoses

Brain injury, including hemorrhage into the substance of the brain, sometimes occurs as part of the clinical picture of severe preeclampsia or eclampsia. Some types of brain tumours appear to be adversely affected by pregnancy, but, for the most part, brain tumours are not altered by pregnancy and do not disturb gestation.

Latent psychiatric disorders in unstable persons may be aggravated by pregnancy, but major psychiatric problems seldom appear for the first time during the period before delivery. There are a number of mild emotional disturbances, such as increased anxiety, emotional irritability, and fear of labour or for the normality of the fetus, that are likely to be most intense during the early months of gestation. Such disturbances seem to be most prevalent in women who did not anticipate becoming pregnant or who are unduly worried about the baby. Psychiatric disorders rarely influence pregnancy. Emotional disturbances have been said to be a factor in some spontaneous abortions, but satisfactory proof of the relationship is lacking.

(J.W.Hu.)

DISEASES OF PREGNANCY

Hypertensive disorders of pregnancy. The increase of blood pressure (hypertension) during pregnancy, which is often accompanied by accumulation of fluid in the tissues that causes swelling (edema) and proteinuria (protein in the urine), poses a serious threat to both the woman and the fetus. The maternal hazards include seizures, organ disturbances, and death; the fetal risks, premature delivery and death.

For many years, hypertensive disorders of pregnancy were referred to as toxemias of pregnancy. The basis for this terminology was the belief that circulatory toxins were responsible for the symptoms. These toxins were never discovered, and this terminology has been generally abandoned; however, in its wake, many different terms have been used to describe hypertensive disorders of pregnancy. To help clarify this situation, the Committee on Terminology of the American College of Obstetricians and Gynecologists prepared the following classification system, which has been recommended by the National Institutes of Health Working Group on High Blood Pressure in Pregnancy: (1) chronic hypertension, (2) preeclampsia and eclampsia, (3) preeclampsia superimposed on chronic hypertension, (4) transient hypertension, and (5) unclassified.

Chronic hypertension is defined as a systolic blood pressure of 140 millimetres of mercury (mm Hg) or higher and a diastolic blood pressure of 90 mm Hg or higher, which antedate pregnancy. (The systolic is the highest blood pressure after the heart has contracted; the diastolic, the lowest after the heart has expanded.) An elevated blood pressure that first develops during pregnancy and persists beyond the 42nd day postpartum also is classified as chronic hypertension.

Preeclampsia is diagnosed after 20 weeks' gestation and is categorized as either mild or severe, although both forms must be considered dangerous to the mother and fetus. Mild preeclampsia is typically characterized by the following symptoms: a rise in blood pressure from that prior to 20 weeks' gestation of at least 30 mm Hg systolic or 15 mm Hg diastolic (or, if the earlier blood pressure is unknown, a level of 140/90 mm Hg after 20 weeks' gestation) on two occasions at least six hours apart; excretion of 0.3 gram (0.01 ounce) of protein or more in the urine during a 24-hour period; and evident swelling or rapid weight gain resulting from fluid retention. Coagulation and disturbances of liver functions are less common but are extremely serious.

Severe preeclampsia is defined by any of the following symptoms occurring after the 20th week of pregnancy: a systolic blood pressure of 160 mm Hg or higher or a diastolic pressure of 110 mm Hg or higher on two or more occasions at least six hours apart, excretion of five grams or more of protein in the urine during a 24-hour period, a reduction in the amount of urine normally excreted (500 millilitres or less in 24 hours), cerebral or visual disturbances, epigastric pain, and pulmonary edema or cyanosis (bluish or purplish colour of the skin).

Eclampsia

A patient with preeclampsia is always in danger of rapidly developing eclampsia, which is distinguished by convulsions that may lead to coma. Headache, epigastric pain, and facial twitching usually precede these seizures, although occasionally eclampsia can arise with no warning, sometimes developing in a woman who has only mild hypertension. Another type of preeclampsia that includes small variations in blood pressure, minor decreases in blood platelet count, and small elevations in liver enzymes can progress quickly from a benign state to a syndrome of life-threatening proportions. This condition is known as the HELLP syndrome and is denoted by hemolysis, elevated liver enzymes, and low platelet count. In this situation, delivery of the fetus must be induced, or pregnancy must be immediately terminated.

Preeclampsia can occur in women who have had hyper-

tension prior to becoming pregnant, in which case the prognosis is much more serious for the mother and fetus than when either preeclampsia or chronic hypertension occurs alone. The diagnosis of preeclampsia superimposed on chronic hypertension is made based on increases of blood pressure of 30 mm Hg systolic and 15 mm Hg diastolic that are accompanied by the appearance of proteinuria or edema. A rise in blood pressure either during pregnancy or 24 hours postpartum, unaccompanied by other symptoms of preeclampsia or eclampsia, is categorized as transient hypertension.

Approximately 7 percent of women whose pregnancies progress beyond the first trimester will develop preeclampsia. It is most common in women who are pregnant for the first time. A higher incidence of this condition occurs in family members of women who have a history of preeclampsia, which provides evidence for involvement of a single maternal gene. Medical disorders such as diabetes mellitus can predispose women to preeclampsia, and conditions such as twin pregnancies increase the risk of preeclampsia. A hydatidiform mole (an abnormal pregnancy caused by an abnormal ovum) is often responsible for a preeclamptic condition that develops before 24 weeks' gestation.

Poor tissue perfusion

Although hypertension is an important diagnostic sign of preeclampsia, the disease is actually one of poor perfusion to the tissues, including the fetal-placental unit. This inadequate delivery of fluids to virtually all organs is attributed to the profound vasospasm (constriction of blood vessels that reduces calibre and blood flow) characteristic of preeclampsia that is also responsible for the associated blood-pressure elevation.

Certain organ systems are characteristically involved in preeclampsia, and their resultant abnormalities—alterations in renal function, endothelial cell injury, and cardiovascular changes—have provided insight into the mechanism of this complex disease. The normal immunologic changes that occur as a result of fetal-maternal interactions also have been postulated as having a pathogenetic role in the development of preeclampsia. The cause of the underlying vasoconstriction, however, still remains undefined.

Various approaches have been attempted to prevent preeclampsia in women at high risk for developing the condition. Dietary and sodium restrictions have been unsuccessful, but there is interest in the use of low-dose aspirin therapy and calcium supplementation as preventive measures. These methods are still under investigation.

Treatment of preeclampsia involves slowing the condition's progression to a more severe form to allow fetal growth to continue as long as possible. Bed rest is recommended in cases of mild preeclampsia, but, when more serious symptoms are involved, hospitalization is best. In cases of severe preeclampsia in which the fetus is beyond 30 weeks' gestation, delivery of the infant is thought to be the best course. Antihypertensive drugs are not used, because they mask the clinical signs by which worsening of the condition is recognized.

In the United States, magnesium sulfate is the drug of choice for preventing and treating eclamptic convulsions. European treatment differs—a variety of narcotics, barbiturates, and benzodiazepine derivatives are used. Because the preeclamptic process often accelerates during labour and the postpartum period, magnesium therapy is used during this time as well.

Gestational diabetes. Diabetes mellitus that has been diagnosed for the first time during pregnancy and resolves immediately after delivery is referred to as gestational diabetes. It occurs in between 1 and 4 percent of the total pregnant population, usually in the second or third trimester. Approximately 50 percent of women who develop gestational diabetes will, over the course of their lifetime, develop adult onset (type II) diabetes.

Effects that gestational diabetes can have on the fetus include high birth weight for gestational age, neonatal hypoglycemia, premature delivery with respiratory distress syndrome, difficult delivery, and a higher incidence of fetal-neonatal mortality.

Previously only women with recognizable risk factors for gestational diabetes were screened for glucose intolerance; these included obese women, women who had a family history of diabetes, and those older than 35 years. Because a significant proportion of cases of gestational diabetes—up to 50 percent—were missed in this way, it is now recommended that all women between the 24th and 28th week of gestation be screened for glucose intolerance; those at high risk should be screened during their first prenatal visit. Controversy exists concerning the best glucose-tolerance screening procedure to use.

Treatment of gestational diabetes

Treatment of gestational diabetes varies according to the individual case. Controlling diet is the first, conservative approach; insulin therapy is instituted when glucose levels cannot be managed in this way. Fetal monitoring of growth development is necessary to measure the effectiveness of treatment and to anticipate and prevent complications. An early delivery by cesarean section (incision through the abdominal and uterine walls for fetal delivery) was frequently recommended in the past. Today the procedure, which has its own risks, is selected less often, as long as the disease has been well controlled and fetal development is normal. (Ed.)

Diseases of the placenta. *Placenta praevia.* Implantation takes place in the lower half of the uterus in approximately 1 in 500 pregnant patients. The condition is known as placenta praevia when the placenta lies over all or a portion of the internal opening of the cervix. A total placenta praevia is present when the cervical opening is completely covered. When there is a low implantation of the placenta, the latter lies close to but not over any part of the cervical opening.

Recurrent painless bleeding from the vagina without other symptoms after the sixth month of pregnancy is the typical manifestation of placenta praevia. It is caused by disruption of the placenta as the cervix and lower uterine segment are pulled upward. Each bleeding episode tends to become heavier. Without proper treatment, the baby is likely to die and the mother may do so as well. Unremitting watchfulness of the woman with placenta praevia until the fetus has a chance of survival, with preparation for immediate delivery if hemorrhage becomes brisk, a practice accepted in many clinics, has resulted in a decreased infant mortality without an increase in maternal deaths.

Characteristics, course, and outcome of abruptio placentae

Abruptio placentae. Abruptio placentae is separation, during the latter half of pregnancy, of the normally implanted placenta from its attachment to the uterus before birth of the baby. It also is correctly referred to as "premature separation of the normally implanted placenta" and is called "accidental hemorrhage" in Great Britain. It occurs in approximately 1 in 100 pregnancies. The cause is unknown. It is more common in women who have borne several children.

When a small portion of the placenta separates from the uterus, a condition called partial abruptio placentae, blood either collects in a pool between the uterus and the placenta (concealed hemorrhage) or seeps out of the uterus into the vagina (external hemorrhage). When the entire placenta separates from the uterus, there is massive hemorrhage into the uterine cavity and sometimes into the wall of the uterus. Massive hemorrhage is associated with uterine tenderness, abdominal pain, shock, and loss of fetal movement and fetal heart tones. The baby usually dies. If hemorrhage is severe, the mother's life is in danger.

From J. Huffmann, *Gynecology and Obstetrics* (1962). W.B. Saunders Co.

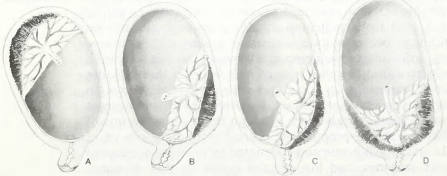


Figure 35: (A) Normal implantation of the placenta in the top or fundus of the uterus. (B) Low implantation. (C) Partial placenta praevia. (D) Complete placenta praevia.

Defective blood clotting occurs in at least 35 percent of patients with abruptio placentae. Kidney failure develops in approximately 1 percent of the cases; it is seen most often in those instances in which treatment has been delayed. Blood replacement, the treatment of shock, the administration of fibrinogen if the patient's clotting mechanism is defective, the administration of oxytocics, and early delivery are the basic essentials of the treatment of abruptio placentae. Delivery is usually by cesarean section.

Placental infarction. Infarction is degeneration and death of a tissue and its replacement with scar tissue. Small yellowish-white deposits of fibrin (a fibrous protein), caused by interference with the maternal circulation, occur normally in the placenta as pregnancy progresses. The fetus usually is not affected by infarction of the placenta unless the process is extensive.

Placenta accreta. Placenta accreta is an abnormal adherence of the placenta to the uterine wall. The chorionic villi attach themselves directly to the uterine muscle in areas where the decidua is poorly developed or absent. All or part of the placenta may be affected. As a result of this abnormality of implantation, the placenta does not separate normally at the time of delivery. Attempts to remove it manually by the physician are frequently followed by severe hemorrhage. Removal of the uterus may be required to save the mother's life.

Benign growths of the placenta

Placental cysts and benign tumours. Placental cysts and benign tumours are relatively rare. Chorionic cysts of small size are disk-shaped, grayish white structures filled with a yellowish fluid and located on the fetal side of the placenta. Decidual cysts are smoothly lined small cavities in the centre of the placenta; they are the result of decidual degeneration and are not true tumours. Angiomas, hemangiomas, fibromas, myxofibromas, and the like are benign growths arising from the placental blood vessels and connective tissue. Solid or semisolid tumours, usually creating small nodular elevations on the fetal side of the placenta, are rarely of clinical significance.

Inflammation of the placenta. Inflammation of the placenta is usually secondary to infection of the membranes. Most often such infections follow the introduction of pus-forming bacteria into the uterus by instrumentation through the vagina; they are the aftermath of prolonged labour or of prolonged rupture of the membranes. If labour is prolonged, bacteria penetrate the fetal side of the placenta, enter the fetal circulation, and often cause death of the infant after delivery.

The placenta may become infected from organisms in the maternal blood. Maternal syphilis, toxoplasmosis, tuberculosis, and malaria may affect the placenta. The viruses of chickenpox and smallpox may cause placental lesions. A number of pathogenic bacteria and viruses cross the placenta and sometimes kill the fetus without causing any specific changes that have been noted in the placenta.

Placental anomalies. Abnormalities in the structure of the placenta are relatively common. It may be partially divided into two or more lobes; there may be extra lobes; or the placenta may be divided into two or more separate structures. Abnormal placentas result from shallow and from deep implantation. The former type, called placenta circumvallata, is associated with several maternal and fetal complications; the latter type, called placenta membranacea, may cause problems at delivery—*e.g.*, bleeding, failure of the membrane to separate.

Anomalies of the umbilical cord. "False knots," which are simply enlarged blood vessels in the cord, are not significant. Actual knots in the cord may become tightened and kill the fetus by cutting off the blood to it. Twisting of the cord also may kill the fetus in the same manner. Spontaneous rupture of the cord interferes with the fetal blood supply and causes fetal death. Extreme shortness of the umbilical cord may interfere with delivery, cause premature separation of the placenta, or tear and cause fetal death from hemorrhage. Another abnormality, called velamentous insertion of the cord, in which multiple blood vessels spread out over the membranes and cervix rather than forming one single cord, is dangerous for the baby because the vessels may tear or be compressed during labour and delivery.

Abnormalities of the amniotic fluid. *Hydramnios.* Hydramnios, sometimes called polyhydramnios, is the presence of an excessive amount of amniotic fluid. Normally the uterus contains approximately 1,000 millilitres (slightly more than one quart) of amniotic fluid; anything over 2,000 millilitres is abnormal. Accumulations of more than 3,000 millilitres occur in approximately one pregnancy in a thousand. Lesser degrees of hydramnios probably occur in about 1 in 150 deliveries. The appearance of large amounts of fluid within the space of a few days is rare; such a condition is met with in fewer than 1 in 4,000.

Excessive amniotic fluid

Hydramnios occurs most often in association with fetal abnormalities, particularly those of the nervous, digestive, and renal systems; when the fetus has erythroblastosis, a disease resulting from incompatibility between the infant's and the mother's blood; when there is more than one fetus; or when the mother has diabetes or preeclampsia. Almost all pregnancies in which the fetus suffers from obstruction of the esophagus and half of those in which there are severe brain anomalies are accompanied by excessive amniotic fluid.

Acute hydramnios causes rapid overdistention and enlargement of the uterus. The woman experiences abdominal pain, nausea and vomiting, and difficulty in breathing. Her heart and blood vessels are put under severe stress; she may show signs of heart failure. Swelling of the feet and legs develops. These manifestations are all caused by the pressure of the rapidly enlarging uterus upon the other viscera.

Chronic hydramnios usually causes enough pressure from the abnormally enlarged uterus to make the affected person uncomfortable.

The cause of hydramnios is unknown. The most tenable theory is that there is a reduction in the amount of fluid that passes from the fetus to the mother and an increase in the amount that passes from the fetus to the amniotic sac. This would explain the relationship between fetal anomalies and hydramnios.

Many pregnancies complicated by an abnormal amount of amniotic fluid terminate prematurely. The fetus has a greatly increased chance of suffering from congenital anomalies. Roughly half of the babies in this group have been lost in the series of cases that have been reported. The greater the amount of fluid, the higher the fetal mortality. Women with hydramnios also are faced with a somewhat higher risk. Premature separation of the placenta and postpartum hemorrhage are the two most significant maternal complications associated with it.

Minor degrees of hydramnios require no treatment. Removal of the excess fluid is the only effective management if symptoms from uterine distention become too distressing. This may be done either by perforating the membranes through the cervix or, preferably, by inserting a needle through the abdominal wall and the wall of the uterus; care is taken to avoid injury to the woman's bowel or the placenta. Either procedure is likely to start labour.

Oligohydramnios. True oligohydramnios, a deficiency in amniotic fluid, is a rare condition of unknown cause. It is seen more often in pregnancies that have extended beyond the projected time of delivery. If it occurs early in pregnancy, there are usually firm adhesions between the membranes and the embryo, with distortion of the fetus. A decrease in the amount of fluid later in pregnancy allows the membranes and uterine wall to press on the baby. The baby's position is distorted, and as a result it may be born with a clubfoot or wryneck. Its skin is dry and thickened. Defective development of the kidneys is common with oligohydramnios. As a rule, the condition causes the mother no distress, but the infant has a greatly increased chance of being born with major anomalies.

Too little amniotic fluid

Trophoblastic disease. *Hydatidiform mole.* A hydatidiform mole is an abnormality of the conceptus in which changes that began early in embryonic life convert the placental villi into a mass of thin-walled, grapelike, translucent vesicles, or blisters, filled with a gelatinous or watery fluid. In a typical case, the uterus is distended by a spongy mass of these vesicles. The primary cause of molar changes is unknown; however, it has been correctly described as a "temporary missed abortion of a blighted ovum." The

embryo is either absent or dead. The immediate condition that causes hydatidiform swelling is disappearance of the blood vessels in the villi, with continued growth and often overgrowth of the trophoblast. Distention of the villi by fluid is due to continued activity of the trophoblast in the absence of a functioning villous circulation.

In the ova there are many degrees of hydatidiform change; many of the changes, usually in younger specimens, are not marked enough to warrant being called hydatidiform moles. True moles—characterized by hyperplasia, or overgrowth, of the trophoblast, edema of the villous connective tissue framework, and defective growth of the villous blood vessels—occur perhaps once in 2,000 pregnancies. They are not tumours and are not the aftermath of a former pregnancy. They are themselves an abnormality of a current pregnancy. Occasionally in a twin pregnancy one fetus is normal and the other a mole. Eighty percent of the moles are expelled about the 20th week of pregnancy and bring the patient no more trouble. Approximately 16 percent of hydatidiform moles invade the uterine muscle, causing bleeding. This type of mole, referred to as an invasive mole or chorioadenoma destruens, may in rare instances perforate the uterus and cause death from hemorrhage. Molar villi rarely are carried to the lung or brain. When they are, the patient may suffer from hemorrhage into the lung or die from hemorrhage within the brain.

Signs of hydatidiform mole; complications

The woman who develops a hydatidiform mole has the symptoms of pregnancy; her uterus usually enlarges more rapidly than it should, she is more likely to suffer from preeclampsia, and she begins to bleed vaginally, usually by the 20th week of gestation. The molar pregnancy is expelled vaginally, or, if hemorrhage is severe, the obstetrician may remove it by surgery.

In approximately 2.5 percent of patients, hydatidiform moles change into choriocarcinoma, a highly malignant tumour of the trophoblast. For that reason, patients who have hydatidiform moles are observed carefully. Continued bleeding or a rising quantity of chorionic gonadotropin in her urine or blood after passage of a mole suggests that a patient has either an invasive mole or a choriocarcinoma. Chemotherapy has been effective treatment for many cases of this type. Removal of the uterus may be necessary. The complexities of diagnosis and the differences in situations require that therapy be keyed to the individual.

Choriocarcinoma. Choriocarcinoma is a rare, extremely malignant type of tumour arising from the trophoblast. The reasons that normal chorionic cells undergo cancerous change, with exaggeration of their natural and potent tendency to invade the uterine muscle and break down blood vessels, are unknown. Choriocarcinoma occurs approximately once in 160,000 normal pregnancies. In approximately 50 percent of the cases the tumour develops from a hydatidiform mole, in another 25 percent after an abortion, and in 25 percent after a normal pregnancy. Occasionally it appears after a tubal pregnancy. It has been known to coexist with pregnancy. It is, for some unknown reason, more common in Asia. Choriocarcinoma developing as a teratoid tumour of the ovary (a tumour made up of a number of different tissues) is a rare entity not related to pregnancy and is not to be confused with the tumour being discussed here.

Signs of choriocarcinoma

As a rule, in the development of a choriocarcinoma there has been a normal pregnancy, an abortion, or the delivery of a mole, and the uterus has not returned to its normal size. The woman begins to bleed from the vagina. Blood loss may be modest or excessive in amount. Tissues obtained by a curettage (scraping) may be, but are not always, indicative of choriocarcinoma.

The tumour begins in the uterus, where it forms a spongy, bleeding mass of easily torn tissue or a shaggy ulcer. When examined microscopically, it is found to consist of both cytotrophoblast and syncytiotrophoblast. The cells spread rapidly by way of the bloodstream, producing secondary tumours in the lung, the brain, the liver, or elsewhere.

Choriocarcinoma formerly was almost invariably fatal. Today an impressive (two out of three in some case series) number of patients have survived for many months after the administration of chemotherapeutic agents. Most workers in this field at this time are using methotrexate. The rapidly growing embryonic cells of the trophoblast need nucleic acids for growth and division; for the synthesis of nucleic acids, folinic acid is essential, and methotrexate, by preventing the conversion of folic acid to folinic acid, cuts off the supply of the latter. A number of other cytotoxic drugs (drugs destructive to cells) also are being used in the treatment of choriocarcinoma, and other chemotherapeutic agents are being tested for effect on this type of tumour; actinomycin D has been used successfully. Removal of the uterus is frequently, but not always, a part of the treatment of choriocarcinoma. (J.W.Hu.)

Parturition: the process of birth

Parturition, or labour, is the process of bringing forth a child from the uterus, or womb. (The prior development of the child in the womb is described in the article GROWTH AND DEVELOPMENT: *Human embryology*.)

THE STAGES OF LABOUR

First stage: dilatation. Early in labour the uterine contractions come on at intervals of 20 to 30 minutes and last about 40 seconds. They are then accompanied by slight pain, which usually is felt in the small of the back. The term labour pains is often used as a synonym for uterine contractions.

As labour progresses, these uterine contractions become more intense and also progressively increase in frequency until, at the end of the first stage, when dilatation is complete, they recur about every three minutes and are quite severe. With each contraction a twofold effect is produced to facilitate the opening up of the cervix. Because the uterus, or womb, is a muscular sac containing a bag of waters (the sac containing the amniotic fluid) that more or less surrounds the child, contraction of the musculature of its walls should diminish its cavity and compress its contents. Because its contents are quite incompressible, however, they are forced in the direction of least resistance, which is in the direction of the internal os, or upper opening of the neck of the womb, and are driven, like a wedge, farther and farther into this opening. In addition to forcing the uterine contents in the direction of the cervix, shortening of the muscle fibres that are attached to the neck of the womb tends to pull these tissues upward and away from the opening and, thus, adds to its enlargement. By this combined action each contraction of the uterus not only forces the bag of waters and fetus downward against the dilating neck of the womb but also pulls the resisting walls of the latter upward over the advancing bag of waters, presenting part of the child.

Duration of first stage

In spite of this seemingly efficacious mechanism, the duration of the first stage of labour is rather prolonged, especially in women who are in labour for the first time. In them the average time required for the completion of the stage of dilatation is between 13 and 14 hours, while in women who have previously given birth to children the average is eight to nine hours. Not only does a previous labour tend to shorten this stage but this tendency often increases with succeeding pregnancies, with the result that a woman who has given birth to three or four children may have a first stage of one hour or less in her next labour.

The first stage of labour is often somewhat prolonged, also, in women who become pregnant for the first time after they have passed the age of 35, because the cervix dilates less readily. A similar delay is to be anticipated in cases in which the cervix is extensively scarred as a result of previous labours, amputation, deep cauterization, or any other operation on the cervix. Even a woman who has borne several children and whose cervix, accordingly, should dilate readily may have a prolonged first stage if the uterine contractions are weak and infrequent or if the child lies in an inconvenient position for delivery and, as a direct consequence, cannot be forced into the mother's pelvis.

On the other hand, the early rupturing of the bag of waters often increases the strength and frequency of the labour pains and thereby shortens the stage of dilatation; occasionally, premature loss of the waters leads to molding

of the uterus about the child and thereby delays dilatation by preventing the child's normal descent into the pelvis. Just as an abnormal position of the child and molding of the uterus may prevent the normal descent of the child, an abnormally large child or an abnormally small pelvis may interfere with the descent of the child and prolong the first stage of labour.

Second stage: expulsion. About the time that the cervix becomes fully dilated, the bag of waters breaks, and the force of the involuntary uterine contractions is augmented by voluntary bearing-down efforts of the mother. With each labour pain she takes a deep breath and then contracts her abdominal muscles. The increased intra-abdominal pressure thus produced may equal or exceed the force of the uterine contractions. When properly used, accordingly, these bearing-down efforts may double the effectiveness of the labour pains. As the child descends into and passes through the birth passages, the sensation of pain is often increased. This condition is especially true in the terminal phase of the stage of expulsion, when the child's head distends and dilates the maternal tissues as it is being born.

Transverse position of the head. The manner in which the child passes through the birth canal in the second stage of labour depends upon the position in which it is lying and the type of the mother's pelvis. The sequence of events described in the following numbered paragraphs is that which frequently occurs when the mother's pelvis is of the usual type and the child is lying with the top of its head lowermost and transversely placed and the back of its head (occiput) directed toward the left side of the mother (A in Figure 36). The top of the head, accordingly, is leading and its long axis lies transversely.

Flexion and internal rotation of head

1. The force derived from the uterine contractions and the bearing-down efforts exerts pressure on the child's buttocks and is transmitted along the vertebral column to drive the head into and through the pelvis. Because of the eccentric attachment of the spine to the base of the skull, the back of the head is made to advance more rapidly than the brow with the result that the head becomes flexed (*i.e.,* the neck is bent) until the chin comes to lie against the breastbone (B in Figure 36). As a consequence of this flexion mechanism, the top of the head (occiput) becomes the leading pole and the ovoid head circumference that entered the birth canal is succeeded by a smaller, almost circular circumference, the long diameter of which is about two centimetres (three-fourths inch) shorter than that of the earlier circumference.

2. As the head descends more deeply into the birth canal, it meets the resistance of the bony pelvis and of the sling-like pelvic floor, or diaphragm, which slopes downward, forward, and inward. When the back of the head, the leading part of the child, is forced against this sloping wall on the left side, it naturally is shunted forward and to the right as it advances, just as a ferryboat is shunted into its wharf by meeting the sloping resistance of its ways (C in Figure 36). This internal rotation of the head brings its longest diameter into relation with the longest diameter of the pelvic outlet and thus greatly assists in the adaptation of the advancing head to the configuration of the cavity through which it is to pass.

3. Further descent of the head directly downward in the direction in which it has been traveling is opposed by the lower portion of the mother's bony pelvis, behind, and the resisting soft parts that are interposed between it and the opening of the vagina (C in Figure 36). Less resistance, on the other hand, is offered by the soft and dilatable walls of the lower birth canal, which is directed forward and upward. The back of the child's head accordingly advances along the lower birth canal, distending its walls and dilating its cavity while the head progresses. Soon the back of the child's neck becomes impinged against the bones of the pelvis, in front, and the chin is forced farther and farther away from the breastbone. Thus, as extension (bending of the head backward) takes the place of flexion, the occiput, brow, eye sockets, nose, mouth, and chin pass successively through the external opening of the lower birth canal and are born (D in Figure 36).

4. The neck, which was twisted during internal rotation of the head, untwists as soon as the head is born. Almost immediately after its birth, therefore, the occiput is turned toward the left and backward.

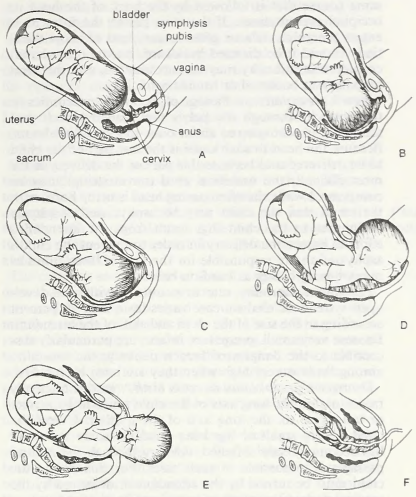


Figure 36: *Sequential changes in the position of the child during labour.*
(A) Onset of labour. (B) Flexion. (C) Internal rotation of the head. (D) Extension. (E) External rotation of the head. (F) Uterus immediately after birth; cord has been cut and placenta is separating from uterine wall.
Drawing by Ramon Goas after Alfred C. Beck

As the child's lower shoulder advances, it meets the sloping resistance of the pelvic floor on the right side and is shunted forward and to the left toward the middle of the pelvis in front. This position brings the long diameter of the shoulder circumference into relation with the antero-posterior, or long diameter, of the pelvic cavity. Because of this internal rotation of the shoulders, the occiput undergoes further external rotation backward and to the left so that the child's face comes to look directly at the inner aspect of the mother's right thigh (E in Figure 36).

Birth of the shoulders. Soon after the shoulders rotate, the one in front appears in the vulvovaginal orifice and remains in this position while the other shoulder is swept forward by a lateral bending of the trunk through the same upward and forward curve that was followed by the head as it was being born. After this shoulder is delivered, the shoulder in front and the rest of the child's body are expelled almost immediately and without any special mechanism.

An average of about one hour and 45 minutes is required for the completion of the second stage of labour in women who give birth for the first time. In subsequent labours the average duration of the stage of expulsion is somewhat shorter.

Posterior position of the occiput. The child may lie so that the back of its head is directed backward and toward either the right or left side. The leading pole is then in the right or left posterior quadrant of the mother's pelvis, and the condition is referred to as a posterior position of the occiput. In such cases the back of the child's head usually rotates to the front of the pelvis and labour proceeds as in transverse positions. Because of the longer rotation required, labour may be somewhat more prolonged than in transverse positions.

Face presentation. When the child's head becomes bent back (extended) so that it enters and passes through the pelvis face first, the condition is known as a face presentation. The chin is then the leading pole and follows the

same course that is followed by the back of the head in occipital presentations. If the chin lies to the front as it enters the pelvis, labour often is easy and of short duration. Should it be directed backward, on the other hand, considerable difficulty may be encountered, and the head may have to be flexed or rotated artificially.

Breech presentation. Passage of the lower extremities or the buttocks through the pelvis first, called breech presentation, is encountered about once in every 30 labours. Because the head in such cases is the last part of the child to be delivered and because this part of the delivery is the most difficult, the umbilical cord (navel string) may be compressed while the aftercoming head is being born, with the result that the child may be asphyxiated. Asphyxia or injuries to the child that result from the attendant's effort to hasten the delivery in order to prevent the child's asphyxiation are responsible for the loss of three times as many breech babies as head-on babies.

The infant mortality rate in countries with well-developed systems of medical care varies from 2 to 10 percent according to the size of the child and skill of the attendant. Because very small, premature infants are particularly susceptible to the dangers of breech delivery, the mortality among them is very high when they are born breech first.

Transverse presentation or cross birth. In this relatively rare situation the long axis of the child tends to lie across, or transverse to, the long axis of the mother. Unless the child is very small or has been dead for some time and has become greatly softened, delivery through the natural passages is impossible in such cases. For this reason the child must be turned by the attendant or delivered by the surgical procedure called cesarean section if it is alive.

Slight risk during pregnancy

That the above-mentioned complications are infrequent and can be cared for easily is shown by the excellent statistics of many maternity services, in which the maternal death rate is less than one per 1,000 and would be still lower if the deaths caused by complicating systemic diseases were excluded. The infant mortality rate is also usually low, ranging between 1.5 and 3 percent. It likewise would be much lower if the premature and poorly developed infants were excluded. In other words, the risk to a healthy mother who carries her child to maturity is less than one per 1,000, and the risk to her mature child is about 0.5 percent.

Third stage: placental stage. With the expulsion of the child, the cavity of the uterus is greatly diminished (F in Figure 36). As a consequence the site of placental attachment becomes markedly reduced in size, with the result that the placenta (afterbirth) is separated in many places from the membrane lining the uterus. Within a few minutes subsequent uterine contractions complete the separation and force the afterbirth into the vagina, from which it is expelled by a bearing-down effort. The third stage of labour, accordingly, is of short duration, seldom lasting longer than 15 minutes. Occasionally, however, the separation may be delayed and accompanied by bleeding, in which case artificial removal of the placenta is necessary. The attendant always remains with the mother until the possibility of hemorrhage has been eliminated by firm retraction of the uterus. (A.C.Bk.)

RELIEF OF PAIN IN LABOUR

Pain experienced in childbirth can be reduced or relieved by psychoprophylaxis, systemic drugs, regional nerve blocks, or a combination of these methods. One of the first drugs to be used for pain relief was chloroform, which was initially employed in the 1850s but eventually came into disuse because of its toxicity. In the early 20th century, a mixture of scopolamine, an amnesic drug, and morphine was given to produce "twilight sleep." On awakening from the induced dreamlike state, the woman would have no memory of her labour pains. The desire to be an active participant in the birth experience and to avoid the side effects of delirium and hallucinations have led to abandonment of this approach.

Since first described in the 1930s, psychoprophylaxis has gained in popularity as a method of psychologically and physically preparing a woman for childbirth, thereby helping her to anticipate and cope with the pain of labour. A more detailed discussion follows below. In addition, a comfortable and pleasant environment, supportive friends and family, and a competent and encouraging birth attendant can help to reduce or even eliminate the need for pharmaceutical pain relief. No one method, however, is suited to every woman. Drugs and techniques that are currently in use are described briefly in the following sections (see also DRUGS AND DRUG ACTION: *Anesthetics*).

Systemic drugs. Meperidine is the most common narcotic drug used for pain reief (analgesia) during labour. It has superseded morphine as the drug of choice because although some of the side effects associated with both drugs—namely, nausea and vomiting—are the same, those of meperidine are less severe. When promethazine is given in conjunction with meperidine, these side effects are further ameliorated. Other negative maternal effects caused by systemic analgesics are drowsiness, respiratory depression, and lowering of blood pressure (hypotension). Because systemic drugs cross the placental barrier, they can also affect the newborn, causing respiratory depression, decreased alertness, and abnormal reflexes. The longer the duration between the administration of the narcotic and the birth of the child, the higher the level of the drug in the infant and the greater its effects. Another drug commonly used for systemic analgesia, and which is reputed to produce less neonatal depression, is butorphanol.

Once important drugs used for lessening labour pains, barbiturates are no longer administered during active labour. Although they are sedatives, which normally induce a relaxed state, they are not analgesics and actually may increase sensitivity to pain. Barbiturates also cause respiratory depression in the newborn, which can be exacerbated by the concomitant use of narcotic analgesics. Sedatives such as secobarbital and pentobarbital are used only in the early stages of labour to help the woman relax and rest before the contractions of active labour begin.

Regional nerve blocks. Concerns about the negative effects that systemic drugs may have on the mother and newborn have led to the heavy reliance on regional, or local, anesthesia. The term regional nerve block refers to the prevention of the conduction of nerve impulses from the peripheral nervous system to the central nervous system by an anesthetic agent. The anesthesia is injected into an area in which the drug will be in direct contact with nervous tissue, thus numbing only this part of the body and allowing the woman to retain consciousness, lucidity, and control over the rest of her body.

The lumbar epidural block has become one of the most popular choices for management of labour pain in the United States. The most common anesthetics used are bupivacaine and chloroprocaine. When a catheter is used, the advantages of this technique include the ability to modify dose, volume, and type of anesthetic, as appropriate to the stage of labour. If a cesarean delivery becomes necessary, the epidural anesthesia can be extended to provide pain relief for the operation. Problems associated with a lumbar epidural block include lowering of maternal blood pressure and urinary retention. Because this procedure can slow labour, the hormone oxytocin is often administered concurrently to stimulate uterine contractions.

Epidural anesthesia is replacing spinal anesthesia, or saddle block, as it is sometimes referred to—the targeted area for anesthetization being the region that would come into contact with a saddle. (The numbing effect, however, occasionally extends beyond the intended "saddle" area, reaching as far as the toes.) Extreme maternal hypotension, a decrease in utero-placental perfusion, and loss of the urge of push are risks that can accompany a spinal block. These effects, as well as the popularity of more natural childbirth experiences, have contributed to the decline in the use of this method.

The pudendal block is a relatively simple and common procedure that numbs the birth canal and perineum for spontaneous delivery, forceps delivery, vacuum extraction, and episiotomy. The same anesthetic agents employed in epidural anesthesia are used and are injected through the vagina to the pudendal nerve. This technique will relieve the pain from perineal distension but not from uterine contractions.

NATURAL CHILDBIRTH

In the 1930s Grantly Dick-Read, a British obstetrician, developed a technique of delivery called natural childbirth that minimized the surgical and anesthetic aspects of delivery and concentrated upon the mother's conscious effort to give birth to her child. Although opposed by many physicians who felt that it denied the progress of modern medicine and needlessly primitivized the process of birth, the method was gradually accepted and by the late 1950s was practiced by a sizable percentage of women, especially in the United States and England.

Natural childbirth—sometimes called psychoprophylaxis, prepared childbirth, or the Lamaze method—as formulated by Dick-Read and later advanced by Fernand Lamaze, Elizabeth Bing, Robert Bradley, and Charles Leboyer, stems from the premise that childbirth need not be accompanied by excessive pain. It is believed that labour pains are the result of unnatural physical tension caused by fear, which can be counteracted by understanding and by developing the ability to relax. The various methods prescribe for the expectant mother and a partner a lengthy course of instruction in the mechanics of labour and birth as well as exercises to strengthen the musculature and to encourage proper breathing. Emphasis is placed on involving other family members, especially the father, in the birth process. During her labour the mother is aided by trained personnel and her partner, or "coach," and anesthetic is made available to her when needed. No claims are made that natural childbirth is totally painless; rather it enables the mother's physical response to transcend discomfort.

Arguments for natural childbirth

Natural childbirth presents the advantage of allowing the woman to participate actively, rather than passively, in labour and to experience the actual moment of birth. The prenatal instruction course also provides women with information about the birthing process, which affords them a greater sense of control over this event. (Ed.)

OPERATIVE OBSTETRICS

Most women deliver a baby spontaneously. Complications, however, that were present before labour or that develop during labour may threaten the life of the mother or of the baby and may require intervention by the attending physician.

Obstetrical forceps are used in vaginal delivery to grasp the fetal head in order to extract the fetus or rotate it so that it is in a satisfactory position for delivery. Some controversy surrounds the use of this procedure, but it is generally agreed that it should be used in situations dangerous to the mother or fetus that could be relieved by prompt delivery. If an expeditious delivery is desired to reduce maternal stress, especially if the woman suffers from conditions such as heart disease, acute pulmonary edema, or certain neurological conditions or if exhaustion or a prolonged second stage of labour jeopardize a successful vaginal delivery, forceps may be employed. Fetal indications for the use of forceps include prolapse of the umbilical cord, premature separation of the placenta, and particular abnormal fetal heart rates. It is important that a certain portion of the fetal skull be protruding from the cervix for this technique to be safe for the mother and fetus. Considerable care must be taken to avoid damaging maternal tissues and causing fetal deformation.

The vacuum extractor is a caplike device that is attached by suction to the fetal scalp and is used as an alternative to delivery by forceps. This technique is employed more frequently in Scandinavia and continental Europe than it is in the United States. Cervical and vaginal trauma have occurred in women undergoing this procedure, but it is less severe and less frequent than that experienced with forceps delivery and constitutes the main advantage of vacuum extraction over forceps delivery. Possible fetal complications include damage to the scalp and intracranial hemorrhage. The newer soft-cup vacuum extractor has generally replaced the original metal cup but has not removed the most serious threat to the fetus, which is the high pressure from the suction.

Manual rotation may be used instead of the forceps when the fetal head is in an abnormal position that makes delivery difficult or impossible. In carrying out the procedure, the obstetrician's hand is inserted into the birth canal and the fetal head is turned to a more favourable position.

Bleeding from the uterus during or after the third stage of labour may be controlled by manual removal of the placenta and packing of the uterus. In the former method the operator inserts a hand into the uterine cavity and separates the adherent placenta from the uterine wall. Modern obstetricians rarely pack the uterus with gauze to control bleeding, because it fails to stop bleeding and it carries with it an increased risk of infection.

When conditions are such that the delivery of a child through the vagina would be a hazard to the mother or to the fetus, it may be necessary or preferable to resort to cesarean section, a procedure in which the fetus is delivered through a surgical opening in the uterus made after the uterus has been exposed through an opening in the abdominal wall.

Cesarean section

The cesarean section evolved from being a surgical procedure that was used only in extreme cases and from which the mother rarely recovered to one of the most commonly performed operations in the United States. Prior to the 20th century, women undergoing a cesarean operation usually contracted peritonitis and died. Not until the advent of aseptic technique, dependable anesthesia, and proper suturing methods that controlled hemorrhage was the cesarean delivery considered a reasonable alternative to vaginal delivery.

Maternal complications, although less severe, are still associated with this procedure. Blood loss, injury to the bowel or bladder, and infection are common risks. Healing of the incision also lengthens convalescence. Although the procedure is often done for the benefit of the fetus at risk from asphyxia or trauma resulting from a vaginal birth, there are associated neonatal risks. Infants who have been delivered at various gestational ages sometimes suffer from respiratory illness. The cause is not completely understood, but the syndrome is most often seen in infants delivered abdominally in the absence of labour. Accidental lacerations of the fetus with the scalpel sometimes occur. Cesarean delivery also is linked with a higher incidence of placenta praevia in future pregnancies.

Cesarean delivery is considered appropriate in various situations in which the risks of vaginal delivery to the fetus or mother are deemed to be greater than the risks from abdominal delivery. It is preferred when there is bleeding from placenta praevia or a prematurely separated placenta to avoid further hemorrhage. If the mother is infected with recurrent genital herpes and lesions are apparent at the time of delivery, a cesarean delivery is usually recommended. It is resorted to if a woman's blood pressure rises precipitously during labour, as can occur with preeclampsia (although, in general, vaginal delivery is preferable to cesarean delivery for women with preeclampsia). Unusual cases, such as an anomaly of the genitalia or a paralytic muscular disorder that prevents the mother from pushing during labour, will generally require this operation.

Although the cesarean section is a lifesaving procedure in situations in which either the woman or child would not have survived delivery otherwise, concern that this technique is becoming overused has arisen in the past decade. Since the 1970s, obstetricians have increasingly relied on the cesarean birth as an alternative to vaginal birth. The four most frequent reasons cited for performing cesarean sections in the United States are prolonged labour, fetal distress, breech presentation, and previous abdominal delivery. The current percentage of pregnancies delivered by this technique is between 20 and 25 percent, which is considered too high because of the risks and complications that the cesarean section itself introduces to delivery. This recognition has led to efforts being undertaken by the health care community to limit the use of the procedure. The general trend to encourage vaginal delivery in women who have delivered abdominally is helping to lessen the number of cesarean sections performed.

In rare instances the obstetrician may make incisions in the incompletely dilated uterine cervix. The incisions permit vaginal delivery in those cases in which labour is unduly prolonged and the incomplete dilatation of the cervix is the only thing that is delaying delivery. (Ed.)

ACCIDENTS DURING LABOUR

Lacerations. *Lacerations of the perineum, vulva, and vagina.* Vaginal lacerations usually make themselves known by profuse bleeding after delivery of the baby. Not all extensive lacerations cause bleeding, however, and a large tear in the vaginal wall may not be discovered until the health-care provider inspects the vagina after the placenta is delivered. There is no difficulty in diagnosing lacerations near the external opening of the birth canal, because they are easily seen by the health-care provider. Even minor lacerations are repaired, because, if they are not, granulation tissue may form in the wounds and delay healing. Deep lacerations require anatomic reconstruction of the torn tissues. Extensive tears of the perineum (the tissues between the genital organs and the anus) can often be avoided by performing an episiotomy—an incision in the vulvar orifice, the external genital opening—before delivery of the infant's head. Also, attention on the health-care provider's part to the mechanism of labour, manual assistance in delivery of the head and shoulders, avoidance of too rapid delivery, delivery between pains, and the proper use of the forceps are some of the many measures that help to avoid injuries not only to the perineum but to all the genital tissues.

Cervical lacerations. The cervix, the lower end of the uterus that projects into the vagina, is usually inspected after the placenta has been delivered. Superficial tears look somewhat like a frayed edge on the cufflike cervix. Deeper lacerations usually cause serious bleeding immediately before or after delivery of the placenta, and these lacerations must be repaired promptly.

Small cervical lacerations are not repaired; they heal spontaneously. Deeper tears are sutured. The management of extensive tears into the body of the uterus or the broad ligaments that support the uterus depends on the extent of the injury and its location; an abdominal operation is sometimes required to control the bleeding and to repair the uterus. Occasionally hysterectomy—removal of the uterus—is necessary.

Other accidents. *Rupture of the uterus.* Rupture of the uterus may occur spontaneously; it may be caused by trauma, or it may occur when a cesarean-section scar gives way. The classical signs of impending spontaneous rupture are gradually increasing, constant, severe pain in the lower part of the abdomen, restlessness, a rising temperature, an increasing pulse rate, and a tense, tender uterus that does not relax between strong contractions. When rupture occurs, the patient complains, usually, of extreme pain and then a sensation of something tearing or giving way. Uterine contractions stop. There is extensive internal bleeding. The baby's body can be felt in the mother's abdomen beside the contracted uterus.

Signs of impending rupture

Prompt delivery, almost always by cesarean section, is the treatment of impending rupture. The patient is anesthetized to stop uterine contractions as soon as the diagnosis is made.

An immediate abdominal operation follows the diagnosis of uterine rupture. Bleeding from the torn uterine walls must be stopped as promptly as possible. The fetus is removed. A hysterectomy is usually performed, because the ragged uterine scar is likely to rupture again if the patient has another term pregnancy, and bleeding from the torn uterus is difficult to control. Such patients often require generous quantities of transfused blood. Antibiotics are given, because infection is, or may be, present.

Injuries to the pelvic supporting tissues. Injuries to the pelvic supporting tissues may not be evident during labour. Months or years later the diagnosis will be made, because the patient complains of something bulging from the vagina, involuntary loss of urine while coughing or laughing, a sensation of things falling down, dragging pelvic discomfort, and difficulty in emptying the lower bowel. The bulging mass formed by a cystourethrocele or rectocele, when seen at vaginal examination, confirms it. (A cystourethrocele is a sagging of the urethra and bladder out of position; a rectocele is a protrusion of the rectum into the vagina.) At times the result of the injuries to the pelvic supporting ligaments and muscles may be so severe that the uterus lies completely outside the vagina, and the vagina is turned inside out, forming a huge ball-shaped mass lying between the patient's thighs.

Treatment depends on the severity of the symptoms. Many women live out their lives without much distress from small cystourethroceles and rectoceles. Larger lesions are repaired surgically.

Inversion of the uterus. Another accident that may occur during labour is inversion of the uterus. The uterus turns inside out and upside down so that its inner surface lies outside and against the wall of the vagina. Inversion causes sudden shock. There may be severe bleeding. The diagnosis is made by noting the pear-shaped mass, covered by a shaggy, dark red, bleeding surface, filling the vagina or hanging outside. The placenta may be still attached to it.

Restoration of a uterus to its normal position is accomplished after the patient's shock and hemorrhage are treated, and she is anesthetized. The obstetrician inserts a hand into the patient's vagina and lifts up the uterus. The tension applied to the uterine ligaments by this procedure usually reinverts the uterus. Damage to the urethra or bladder may result.

Embolisms. Embolisms may occur. (An embolism is a blockage of a blood vessel, as by a blood clot or bubble of air.) Amniotic fluid embolism causes sudden, severe respiratory distress, signs of shock, cyanosis (blueing of the skin), heart collapse, and circulatory failure. If the diagnosis is made promptly, oxygen, blood transfusion, and the injection of fibrinogen, a clotting factor, into a vein may be lifesaving.

Effects of amniotic fluid embolism

Air embolism causes the patient to become suddenly short of breath and cyanotic. She may have heart pain and show signs of shock. The heart beats irregularly, and swishing sounds, caused by the presence of air mixed with blood in the heart, can often be heard. Death follows quickly unless the diagnosis is made at once. Treatment consists of drawing the air from the heart with a needle and syringe.

Fibrinogenopenia. Fibrinogenopenia, a shortage of fibrinogen, which causes defective blood clotting and serious bleeding, may result from a number of circumstances, including premature separation of the placenta and a hard, forceful labour. The diagnosis is confirmed by testing the mother's blood. It lacks fibrinogen if it does not clot firmly in 10 minutes or if the clot that does form dissolves when incubated at 37° C (98.6° F) for one hour. Treatment includes management of the condition responsible for the lack of fibrinogen, replacement of blood lost through bleeding, and the administration of fibrinogen intravenously.

Placenta accreta. Abnormal adherence of the placenta to the uterus, a condition called placenta accreta, is suspected when the placenta cannot be expelled. It is strongly associated with placenta praevia and previous cesarean-section deliveries. If the adherence is only partial, there is usually brisk bleeding from the remainder of the placental site. Placenta accreta is treated by immediate removal of the uterus.

Genital bleeding during labour. Genital bleeding during labour is a symptom of a number of obstetrical complications in addition to those discussed above. The diagnosis of premature separation of the placenta (abruptio placentae) is made when the patient complains of sudden abdominal pain and when there is uterine tenderness and vaginal bleeding. The mother may go into shock, and the fetus may die. There may be signs of hidden bleeding and concealed blood within the uterus. This condition is differentiated from placenta praevia by the fact that the placenta is not in the lower uterine segment.

Abruptio placentae

In cases of suspected placenta praevia, the placenta can be located with considerable accuracy by a careful abdominal examination and ultrasonography. Magnetic resonance imaging is beginning to be used to confirm diagnosis and may replace ultrasound.

In every case of serious blood loss, bleeding must be controlled and lost blood replaced in amounts sufficient to protect the mother. Most patients with placenta praevia and premature detachment of the placenta are treated by cesarean section. Cesarean section is not performed for premature detachment of the placenta (abruptio placen-

tae) until a shortage of fibrinogen is corrected. Removal of the uterus may be necessary in rare cases of abruptio placentae when the uterine wall has been so infiltrated by blood that it cannot contract.

Severe bleeding not related to the pregnancy, such as that resulting from rupture of one of the pelvic veins or rupture of the spleen, may occur during labour. When it occurs, it must be treated regardless of the pregnancy.

Accidents to the umbilical cord. An accident to the umbilical cord is suspected when there is marked irregularity in the fetal heart rate and particularly when the irregularity is accentuated by uterine contractions. A prolapsed cord—that is, a cord lying below the head—can be felt through the membranes on vaginal examination. After the membranes have ruptured, the cord can be felt and seen in the vagina. It may hang out of the vulva. The patient is delivered by cesarean section if the infant is alive and if the head can be prevented from pressing on the cord while preparations are made for the operation. The baby is delivered vaginally if the cervix is completely dilated and if conditions are favourable for prompt vaginal delivery. Attempts to replace the cord in the uterus are seldom successful. Vaginal delivery is allowed to continue if the infant is dead.

True knots in the cord and rupture of the cord with bleeding are seldom diagnosed until after delivery. They are usually associated with sudden and, at the time, inexplicable fetal death.

Puerperium, or period of involution

Uterine involution and other changes

Within six to eight weeks after childbirth, most of the structures of the maternal organism that underwent change during pregnancy return more or less to their prepregnancy state. The enlarged uterus, which at the end of gestation weighed about 1,000 grams (35 ounces), shrinks to a weight of about 60 grams (2 ounces). Along with this process of uterine involution, the lining membrane of the uterus is almost completely shed and replaced by a new lining, which is then (six to eight weeks after delivery) ready for the reception of another fertilized ovum (egg). The greatly dilated neck of the womb and lower birth passage likewise undergo marked and rapid involution, but they seldom return exactly to their prepregnancy condition. As a rule, examination of a woman who has given birth accordingly reveals evidence of this. The markedly stretched abdominal wall also undergoes considerable involution, particularly if abdominal exercises are carried out. Although the intradermal tears (striae gravidarum) become smaller and fade, they do not completely disappear but remain as evidence of the marked and rapid stretching of the skin that took place during pregnancy. (J.W.Hu.)

LACTATION

The breasts, unlike most of the other organs, continue to increase in size. Although mammary growth begins during pregnancy under the influence of ovarian and placental hormones and some milk is formed, copious milk secretion sets in only after delivery. Since lactation ensues after a premature birth, it would appear that milk production is held back during pregnancy. The mechanism by which this inhibitory effect is brought about, or by which lactation is initiated at delivery, has long been the subject of an argument that revolves around the opposing actions of estrogen, progesterone, and prolactin, as studied in laboratory animals, goats, and cattle. During pregnancy the combination of estrogen and progesterone circulating in the blood appears to inhibit milk secretion by blocking the release of prolactin from the pituitary gland and by making the mammary gland cells unresponsive to this pituitary hormone. The blockade is removed at the end of pregnancy by the expulsion of the placenta and the loss of its supply of hormones, as well as by the decline in hormone production by the ovaries, while sufficient estrogen remains in circulation to promote the secretion of prolactin by the pituitary gland and so favour lactation. It is also possible that a relative increase in the blood level of adrenal hormones favours the production of milk.

For lactation to continue, necessary patterns of hormone secretion must be maintained; and disturbances of the equilibrium by the experimental removal of the pituitary gland, in animals, or by comparable diseased conditions in human beings, quickly arrest milk production. Several pituitary hormones seem to be involved in the formation of milk, so that it is customary to speak of a lactogenic ("milk-producing") complex of hormones. To some degree, the role of the pituitary hormones adrenocorticotropin, thyrotropin, and growth hormone in supporting lactation in women is inferred from the results of studies done on animals and from clinical observations that are in agreement with the results of animal studies. Prolactin, growth hormone, and adrenal hormone seem of greatest value in restoring lactation after removal of the pituitary, although the precise response varies from species to species.

Suckling and the release of oxytocin

The stimulus of nursing or suckling supports continued lactation in two ways: it promotes the secretion of prolactin (and possibly other pituitary hormones of value in milk formation), and it triggers the release of yet another hormone from the pituitary gland—oxytocin, which causes contraction of special muscle cells around the alveoli in the breast and ensures the expulsion of milk. It is in this way that a baby's sucking at one breast may cause an increase in milk flow from both, so that milk may drip from the unsuckled nipple. About 30 seconds elapse between the beginning of active suckling and the initiation of milk flow.

The nerve supply to the mammary glands is not of great significance in lactation, for milk production is normal after the experimental severing of nerves to the normal mammary glands in animals or in an udder transplanted to the neck of a goat. Milk ejection, or "the draught," in women is readily conditioned and can be precipitated by the preparations for nursing. Conversely, embarrassment or fright can inhibit milk ejection by interfering with the release of oxytocin; alcohol, also, is known to block milk ejection in women, again by an action on the brain. Beyond its action on the mammary glands, oxytocin affects uterine muscle, so that suckling can cause contractions of the uterus and may sometimes result in cramp. Since oxytocin release occurs during sexual intercourse, milk ejection in lactating women has been observed on such occasions. Disturbance of oxytocin secretion, or of the milk-ejection reflex, stops lactation just as readily as a lack of the hormones necessary for milk production, for the milk in the breast is then not extractable by the infant. Many instances of nursing failure are due to a lack of milk ejection in stressful circumstances; fortunately, treatment with oxytocin, coupled with the reassurance gained from a successful nursing, is ordinarily successful in overcoming the difficulty.

Suckling can initiate lactation in nonpregnant women. This has been seen most often in women of childbearing age but also has been observed in older persons. A baby who had lost his mother was suckled by his 60-year-old grandmother, who had borne her last child 18 years before. The grandmother produced milk after a few days and continued to nurse the baby until he was a year old and could walk. Rarely, lactation has been reported to set in after operations on the chest; in such instances it is attributed to injury or irritation of the nerves in this region. Such observations argue against the possibility that lactation continues simply as a consequence of emptying the breasts.

WEANING AND THE CESSATION OF LACTATION

Composition and properties of milk. Milk can be regarded as an emulsion of fat globules in a colloidal solution of protein together with other substances in true solution. Two constituents of milk, the protein casein and milk sugar, or lactose, are not found elsewhere in the body. In the United States the frequency of breast-feeding has risen sharply in the past 25 years. Its advantages include nutritional, immunologic, and psychological benefits. Human breast milk is superior to modified cow's milk formulas, which may lack essential and beneficial components and are not absorbed as easily or as quickly by the infant. Maternal breast milk provides vitamins, minerals, protein, and anti-infectious factors; antibodies that protect

the infant's gastrointestinal tract are supplied, resulting in a lower rate of enteric infection in breast-fed than in bottle-fed babies. The bonding that is established through breast-feeding is advantageous to building the parent-child relationship.

The nutritional status of the mother is important throughout this period. The Food and Nutrition Board of the National Research Council recommends a daily caloric increase of approximately 400 kilocalories over nonpregnant diet. Most drugs that are taken during this time are secreted through the milk, and smoking reduces breast-milk volume and decreases infant growth rates.

Colostrum

The milk released from the breast when lactation starts differs in composition from the mature milk produced when lactation is well established. The early milk, or colostrum, is rich in essential amino acids, the protein building blocks essential for growth; it also contains the proteins that convey immunity to some infections from mother to young, although not in such quantity as among domestic animals. The human infant gains this type of immunity largely within the uterus by the transfer of these antibody proteins through the placenta; the young baby seldom falls victim to mumps, measles, diphtheria, or scarlet fever. For a short time after birth, proteins can be absorbed from the intestine without digestion, so that the acquisition of further immunity is facilitated. The growth of viruses and bacteria in the intestines is probably inhibited by immune factors in human milk. After childbirth the composition of milk gradually changes; within four or five days the colostrum has become transitional milk; mature milk is secreted some 14 days after delivery.

Some variations between human colostrum, transitional milk, and mature milk and cow's milk are shown in Table 2. The greater amount of protein in unmodified cow's milk is largely responsible for its dense, hard curd, which the infant cannot digest; the difficulty can be avoided by heat treatment or dilution of the milk. Ordinarily, when cow's milk is fed to young infants, it is modified so as to match its composition as far as possible to breast milk.

Varying customs with respect to weaning

There is no typical age at which human infants are weaned, for this varies from country to country and among the social classes of a nation. In India women in the higher socioeconomic groups tend to use artificial feeding, while the reverse relationship holds in Britain and the United States. Most commonly, weaning is a gradual process, with a gradual increase in the proportion of solid food supplied to the infant together with breast milk. Pediatricians in general have concluded that, on the basis of present knowledge, no nutritional superiority or psychological benefits result from the introduction of solid foods into the infant diet earlier than the age of $2\frac{1}{2}$ to $3\frac{1}{2}$ months and that normal full-term infants can be expected to thrive for the first six months of life on a diet consisting exclusively of milk, either normal human milk or properly modified milk from other sources.

With the reduced demand of the baby, lactation slowly declines and stops. Estrogen treatment is often used to suppress lactation, and the high doses used may accomplish this; but there is often a rebound effect at the end of treatment. Lactation may be slightly depressed when oral contraceptives are being taken in high dosage. Although ovulation is less frequent during lactation, it does occasionally occur. Breast-feeding should not, therefore, be used as a method of contraception. Menstruation usually resumes within six to eight weeks in women who are not breast-feeding; the length of its absence varies in women who breast-feed. (B.T.D./Ed.)

Table 2: Some Constituents of Human Colostrum and Transitional and Mature Milk and of Cow's Milk
(average values per 100 ml whole milk)

	colostrum (1–5 days)	transitional milk (6–14 days)	mature milk (after 14 days)	cow's milk
Energy, kcal*	58	74	71	69
Total solids, g	12.8	13.6	12.4	12.7
Fat, g	2.9	3.6	3.8	3.7
Lactose, g	5.3	6.6	7.0	4.8
Protein, g	2.7	1.6	1.2	3.3
Casein, g	1.2	0.7	0.4	2.8
Ash, g	0.33	0.24	0.21	0.72
Minerals				
Calcium, mg	31	34	33	125
Magnesium, mg	4	4	4	12
Potassium, mg	74	64	55	138
Sodium, mg	48	29	15	58
Iron, mg	0.09	0.04	0.15	0.10

*Kilocalorie: sufficient energy to raise the temperature of 1 kg of water 1° C.

BIBLIOGRAPHY

General features of reproduction. The nature of reproduction is addressed in JOHN TYLER BONNER, *Size and Cycle* (1965), a very general discussion of the evolutionary significance of life cycles in animals and plants; and JAMES WATSON, *Molecular Biology of the Gene*, 4th ed., 2 vol. (1987), an up-to-date summary of molecular replication by one of the pioneers in the field. The process of fertilization is discussed by CHARLES B. METZ and ALBERTO MONROY (eds.), *Fertilization: Comparative Morphology, Biochemistry, and Immunology*, 2 vol. (1967–69), a major source of information, and *Biology of Fertilization*, 3 vol. (1985); EDMUND B. WILSON, *The Cell in Development and Heredity*, 3rd ed. with corrections (1928, reprinted 1987), a classic work; JOHN F. HARTMANN (ed.), *Mechanism and Control of Animal Fertilization* (1983); and PAUL M. WASSERMAN, "The Biology and Chemistry of Fertilization," *Science*, 235(4788):553–560 (Jan. 30, 1987), and "Fertilization in Mammals," *Scientific American*, 259(6):78–84 (December 1988).

Plant reproduction. Investigations of plant reproductive systems are included in DAVID W. BIERHORST, *Morphology of Vascular Plants* (1971), at an advanced level, with many photographs and an extensive bibliography; MARY F. WILLSON, *Plant Reproductive Ecology* (1983); HAROLD C. BOLD and JOHN W. LA CLAIRE II, *The Plant Kingdom*, 5th ed. (1987), an account of the structure and reproduction of representative plant types; HAROLD C. BOLD, CONSTANTINE J. ALEXOPOULOS, and THEODORE DELEVORYAS, *Morphology of Plants and Fungi*, 5th ed. (1987), a profusely illustrated textbook designed for the advanced undergraduate and beginning graduate students; and JON LOVETT DOUST and LESLEY LOVETT DOUST (eds.), *Plant Reproductive Ecology: Patterns and Strategies* (1988).

Types and agents of pollination are analyzed in K. FAEGRI and L. VAN DER PIJL, *The Principles of Pollination Ecology*, 3rd rev. ed. (1979), a well-balanced work based on extensive experience in both tropical and temperate regions; MICHAEL PROCTOR and PETER YEO, *The Pollination of Flowers* (1973), an excellent work on the process of pollination; and BASTIAAN MEEUSE and SEAN MORRIS, *The Sex Life of Flowers* (1984).

Works treating the nature and function of seed and fruit include T.T. KOZLOWSKI (ed.), *Seed Biology*, 3 vol. (1972); P.F. WAREING and I.D.J. PHILLIPS, *The Control of Growth and Differentiation in Plants*, 2nd ed. (1978), written with a physiologist's appreciation for the experimental approach, admirably covering various aspects of dormancy; L. VAN DER PIJL, *Principles of Dispersal in Higher Plants*, 3rd rev. and expanded ed. (1982), a highly original and most stimulating book displaying a thorough knowledge of both tropical and temperate-region biology; J. DEREK BEWLEY and MICHAEL BLACK, *Seeds: Physiology of Development and Germination* (1985); and A.M. MAYER and A. POLJAKOFF-MAYBER, *The Germination of Seeds*, 4th ed. (1989), a very readable, responsible account presented in a well-organized, fairly concise manner.

Animal reproduction. Studies of reproductive systems in invertebrates are included in LIBBIE HENRIETTA HYMAN, *The Invertebrates*, 6 vol. (1940–67), a detailed work on Protozoa through Mollusca; JOSEPH G. ENGEMANN and ROBERT W. HEGNER, *Invertebrate Zoology*, 3rd ed. (1981), a college-level text covering major groups; P.A. MEGLITSCH and FREDERICK R. SCHRAM, *Invertebrate Zoology*, 3rd ed. (1991), a college-level text covering all major groups, highly readable and well illustrated; and ROBERT D. BARNES, *Invertebrate Zoology*, 6th ed. (1994), in which the reproduction of each major invertebrate group is discussed and illustrated. Reproductive systems in vertebrates, including some treatment of human reproduction, are discussed in EDWIN S. GOODRICH, *Studies on the Structure & Development of Vertebrates*, 2 vol. (1930, reprinted 1986), a classic, still useful for morphological details; ROBERT T. ORR, *Vertebrate Biology*, 5th ed. (1982), containing a good general discussion of vertebrate reproduction; C.R. AUSTIN and R.V. SHORT (eds.), *Reproduction in Mammals*, 2nd ed., 5 vol. (1982–86); *Marshall's Physiology of Reproduction*, 4th ed. by G.E. LAMMING, vol. 1, *Reproductive Cycles of Vertebrates* (1984); and ERNST KNOBIL and JIMMY D. NEILL (eds.), *The Physiology of Reproduction*, 2nd ed., 2 vol. (1994), a study of mammals.

Specific topics are treated in ARI VAN TIENHOVEN, *Reproductive Physiology of Vertebrates*, 2nd ed. (1983), primarily for the reproductive physiologist but containing valuable anatomic data relating to all vertebrate classes; JOHN G. VANDENBERGH

(ed.), *Pheromones and Reproduction in Mammals* (1983); PETER K.T. PANG and MARTIN P. SCHREIBMAN (eds.), *Vertebrate Endocrinology*, vol. 4, *Reproduction*, 2 parts (1991); A.D. JOHNSON, W.R. GOMES, and N.L. VANDEMARK (eds.), *The Testis*, 4 vol. (1970–77); B.P. SETCHELL, *The Mammalian Testis* (1978); HENRY BURGER and DAVID DE KRETSER (eds.), *The Testis*, 2nd ed. (1989); LORD ZUCKERMAN (SOLLY ZUCKERMAN) and BARBARA J. WEIR (eds.), *The Ovary*, 2nd ed., 3 vol. (1977), a detailed account of the development, structure, and function of vertebrate ovaries, commencing with protochordates; HANNAH PETERS and KENNETH P. MCNATTY, *The Ovary: A Correlation of Structure and Function in Mammals* (1980); and ELI Y. ADASHI and PETER C.K. LEUNG (eds.), *The Ovary* (1993).

The human reproductive system. In addition to the works on vertebrate reproduction cited in the previous section, the characteristics of human reproduction and reproductive organs are specifically reviewed in ROBERT H. WILLIAMS (ed.), *Textbook of Endocrinology*, 6th ed. (1981), covering in detail the endocrine aspects of the male and female reproductive systems; RICHARD E. JONES, *Human Reproductive Biology* (1991); RALPH M. WYNN and WILLIAM P. JOLLIE (eds.), *Biology of the Uterus*, 2nd ed. (1989), a useful summary of modern research on many aspects of uterine anatomy and physiology; ETHEL SLOANE, *Biology of Women*, 3rd ed. (1993); and CLAUDE GOMPEL and STEVEN G. SILVERBERG, *Pathology in Gynecology and Obstetrics*, 4th ed. (1994), which contains a chapter on each organ of the female genital tract and one on the breast. Reproductive technologies are explained in LORI B. ANDREWS, *New Conceptions* (1984); ARTHUR L. WISOT and DAVID R. MELDRUM, *New Options for Fertility* (1990); and SHERMAN J. SILBER, *How to Get Pregnant with the New Technology* (1991).

Additional information on menstruation and its cessation may be found in DOREEN ASSO, *The Real Menstrual Cycle* (1983); JOHN G. GRUHN and RALPH R. KAZER, *Hormonal Regulation of the Menstrual Cycle* (1989); LINDA R. GANNON, *Menstrual Disorders and Menopause: Biological, Psychological, and Cultural Research* (1985); and HERBERT J. BUCHSBAUM (ed.), *The Menopause* (1983).

Diseases of the human reproductive system are treated in JAMES B. WYNGAARDEN, LLOYD H. SMITH, JR., and J. CLAUDE BENNETT (eds.), *Cecil Textbook of Medicine*, 19th ed. (1992); *Campbell's Urology*, 6th ed. edited by PATRICK C. WALSH *et al.*, 3 vol. (1992), one of the most complete textbooks in the field of urology; DONALD R. SMITH, *General Urology*, 11th ed. (1984), well written; ROBERT W. KISTNER (ed.), *Gynecology*, 4th ed. (1986), a detailed work; *Novak's Textbook of Gynecology*, 11th ed. by HOWARD W. JONES III, ANNE COLSTON WENTZ, and LONNIE S. BURNETT (1988); STEPHEN C. RUBIN and GREGORY P. SUTTON (eds.), *Ovarian Cancer* (1993); CHARLES E. RINEAR, *The Sexually Transmitted Diseases* (1986); and L.C. PARISH and FRIEDRICH GSCHNAIT (eds.), *Sexually Transmitted Diseases: A Guide for Clinicians* (1989).

Human reproduction from conception to birth. General texts include *Human Embryology*, 4th ed. by W.J. HAMILTON and H.W. MOSSMAN (1972); and *Williams Obstetrics*, 19th ed. by F. GARY CUNNINGHAM *et al.* (1993), a standard textbook in the field.

Additional information on pregnancy, childbirth, and lactation may be found in the following specialized texts: IRWIN R. MERKATZ and JOYCE E. THOMPSON (eds.), *New Perspectives on Prenatal Care* (1990); ROBERT K. CREASY and ROBERT RESNIK (eds.), *Maternal-Fetal Medicine: Principles and Practice*, 3rd ed. (1994); *Burwell and Metcalfe's Heart Disease and Pregnancy*, 2nd ed. by JAMES METCALFE, JOHN H. MCANULTY, and KENT UELAND (1986); RICHARD S. ABRAMS, *Handbook of Medical Problems During Pregnancy* (1989); FRANCINE H. NICHOLS and SHARRON SMITH HUMENICK, *Childbirth Education: Practice, Research, and Theory* (1988); *Oxorn-Foote Human Labor & Birth*, 5th ed. by HARRY OXORN (1986); HELEN VARNEY, *Nurse-Midwifery*, 2nd ed. (1987); and RUTH A. LAWRENCE, *Breastfeeding: A Guide for the Medical Profession*, 4th ed. (1994).

(Ed.)

Reptiles

Reptilia is a class of vertebrate animals that includes the snakes, lizards, crocodiles, alligators, turtles, and the tuatara, among the living forms, and a great many extinct types such as dinosaurs, pterosaurs, and ichthyosaurs.

Intermediate between amphibians and the warm-blooded vertebrates, reptiles may be described as those air-breathing vertebrates with internal fertilization and a scaly body covering instead of hair or feathers.

For coverage of related topics in the *Macropædia* and the *Micropædia*, see the *Propædia*, section 313, and the *Index*.

This article is divided into the following sections:

Reptiles: class Reptilia 688
- General features 688
  - Importance
  - Size range
  - Distribution and ecology
- Natural history 690
  - Reproduction and life cycle
  - Growth and longevity
- Behaviour 692
  - Defense
  - Feeding habits
  - Locomotion
- Form and function 694
  - External covering
  - Internal features
  - Sense organs
  - Thermal relationships
- Evolution and paleontology 698
  - Historical development
  - Fossil distribution
- Classification 700
  - Distinguishing taxonomic features
  - Annotated classification
  - Critical appraisal

Major reptilian groups 701
- Chelonia (turtles) 701
- Rhynchocephalia (tuatara) 706
- Sauria (lizards) 707
- Serpentes (snakes) 713
- Crocodilia (crocodiles) 720

Bibliography 723

REPTILES: CLASS REPTILIA

General features

Importance. The economic and ecological importance of reptiles to humans is not as great as the other major vertebrate groups—birds, fishes, and mammals. Locally, some species are eaten on occasion, if not regularly. The green turtle (*Chelonia mydas*) is the most widely eaten species of reptile. The giant Galápagos tortoise was especially popular as food among 19th-century seafarers and, for this reason, nearly became extinct. Among the lizards, iguanas are perhaps the most popular as a local food.

Leather goods, including luggage, gloves, belts, handbags, and shoes, are made from the skins of lizards, crocodilians, and snakes. This has led to the virtual extinction of several species of crocodilians and to severe reduction of populations of large lizards, snakes, and turtles. As living subjects for biological research, lizards in particular have been useful to the scientist. Venomous species constitute little hazard to humans except in limited rural areas.

Size range. Although persistent but unsubstantiated reports have been made of 12-metre (40-foot) anacondas, this gigantic South American snake, while it probably is the largest living species, does not usually exceed nine metres (30 feet) in length. The reticulated python of Southeast Asia and the East Indies has been recorded at 8.4 metres (27.6 feet). The rock python (*Python sebae*) of Africa reaches 7.5 metres (24.6 feet). No other group of living snakes approaches the pythons and boas in weight, although the king cobra of Asia and the East Indies comes close in length (5.4 metres, or 17.7 feet) and is the longest venomous snake. The heaviest venomous snake is probably the eastern diamondback rattlesnake (*Crotalus adamanteus*), which, though not exceeding 2.4 metres (7.9 feet), may weigh as much as 15.5 kilograms (34 pounds). The largest of the common nonvenomous snakes of the family Colubridae is probably the Oriental rat snake, *Ptyas carinatus* (3.73 metres; 12.2 feet).

Largest snakes

Four living species of crocodilians grow larger than six metres (20 feet): the American crocodile (*Crocodilus acutus*), the Orinoco crocodile (*C. intermedius*), the saltwater crocodile (*C. porosus*), and the gavial (*Gavialis gangeticus*). The last two may approach nine metres (30 feet).

The giant among living turtles is the marine leatherback (*Dermochelys*), which reaches a total length of about 2.7 metres (8.9 feet) and a weight of about 680 kilograms (1,500 pounds). The largest of the land turtles is a Galápagos tortoise weighing 255 kilograms (560 pounds).

The largest modern lizard is a monitor, the Komodo dragon of the East Indies; it attains a length of three metres (10 feet). Two or three other species of monitors reach 1.8 metres (5.9 feet). The common iguana comes close to that size, but no other lizard does.

None of the living reptiles, with the possible exception of snakes, is as large as the largest extinct representative of its particular group. The 2.7-metre leatherback turtle is smaller than the extinct 3.6-metre (11.8-foot) marine turtle *Archelon*, and no modern crocodile approaches the estimated 15-metre (49-foot) length of *Phobosuchus*. The Komodo dragon does not compare with the six-metre (20-foot) or more mosasaur *Tylosaurus*. Lengths exceeding 30 metres (100 feet) and weights of 91,000 kilograms (200,000 pounds) or more may have been achieved by some browsing, quadruped dinosaurs.

The smallest reptiles are the geckos, some of which grow no longer than three centimetres (slightly more than one inch). Certain blind snakes (Typhlopidae) are less than 10 centimetres (four inches) in length when fully grown. The smallest turtles weigh less than 450 grams (one pound) and reach a maximum length of 12.5 centimetres (about five inches). The smallest crocodilians are the dwarf crocodile (*Osteolaemus tetraspis*) and the smooth-fronted caiman (*Paleosuchus*), about 1.7 metres (5.6 feet) in length.

The smallest reptiles

Distribution and ecology. *North Temperate Zone.* Although living reptiles, which number some 6,000 species, are primarily tropical animals, many inhabit the temperate zones. The northernmost ranges are those of the lacertid lizard (*Lacerta vivipara*) and the common viper (*Vipera berus*), both of Europe and Asia. These ovoviviparous (live-bearing, but the unborn snakes develop within eggs) species live north of the Arctic Circle, at least in Scandinavia. Two other lizards, the slowworm (*Anguis fragilis*) and the sand lizard (*Lacerta agilis*), and two snakes, the grass snake (*Natrix natrix*) and the smooth snake (*Coronella austriaca*), reach 60° N in Europe. Of the six

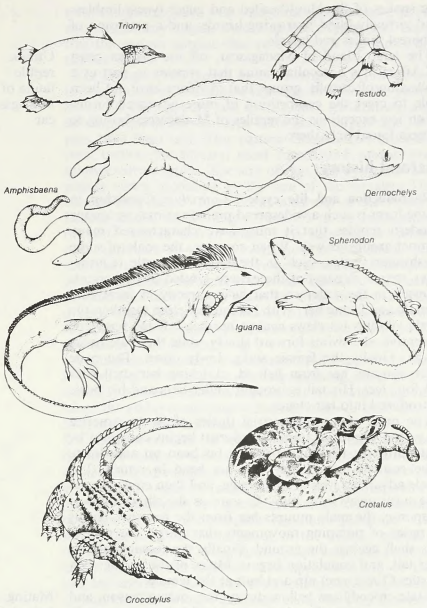


Figure 1: Body plans of living reptiles.

Drawing by J. Helmer based on (*Crotalus*) photograph courtesy of the National Marine Fisheries Service; (*Crocodylus*) *Reader's Digest Living World of Animals*; (others) J.Z. Young, *The Life of Vertebrates*, The Clarendon Press, Oxford

northern species, all but the grass snake are ovoviviparous. Across Siberia only *Lacerta vivipara* and *Vipera berus* live north of 60°.

In North America no reptile reaches the 60th parallel. Two species of garter snakes live as far north as 55° in western Canada. In North America and Eurasia the northern limit of turtles is about 55° N.

It is only south of 40° N that reptiles become abundant. In the eastern United States and eastern Asia, water snakes (*Natrix*), rat snakes (*Elaphe*), racers (*Coluber*), green snakes (*Opheodrys*), northern skinks (*Eumeces*), glass "snakes" (lizards of the genus *Ophisaurus*), and soft-shelled turtles (*Trionyx*) are common. One of the two living species of *Alligator* lives in southeastern United States; the other lives in China. Even though both regions are characterized by many species of emydid turtles (family Emydidae), the genera to which the species belong are found in only one region or the other. Many lizards of temperate Eurasia belong to the families Agamidae and Lacertidae, which do not occur at all in the Americas. On the other hand, many lizards of North America are in the families Iguanidae and Teiidae, which do not live in Eurasia.

The fauna of the eastern United States is almost as distinct from that of the western United States and northern Mexico (which is faunistically part of the same region) as it is from that of eastern Asia. The eastern United States has many genera and species of emydid turtles; the western United States (defined by a diagonal line running southeast to northwest through Texas, then northward along the Continental Divide) has only four or five species. Few genera and species of iguanid lizards inhabit the eastern United States, whereas the western United States has many. Although the eastern United States has more species of water snakes, the western United States contains more garter snakes. More species of snakes appear in the eastern United States than in the western areas, while the converse is true of lizard species.

Ecological types of the North Temperate Zone

Reptiles of the North Temperate Zone include many ecological types. Aquatic groups are represented in both hemispheres by the water snakes, many emydid turtles, and the two alligators. Terrestrial groups include tortoises, ground-dwelling snakes, and many genera of lizards. Arboreal snakes are few, and arboreal lizards are almost nonexistent. Burrowing snakes are common. Specialized burrowing lizards are few.

Central and South America. In Central America the reptile fauna becomes richer. Besides several turtle families found in the eastern United States, Central America has three genera of turtles (*Dermatemys, Claudius,* and *Staurotypus*) not living elsewhere. Crocodilians become more numerous both in species and individuals. Lizards and snakes are particularly more abundant.

Many of the genera of iguanid lizards occurring in the western United States have species in Mexico; one genus of spiny lizards (*Sceloporus*) reaches its peak of numbers of species in Mexico. South of Mexico the North American iguanid genera disappear and are replaced by tropical groups such as the black iguanas (*Ctenosaura*), the helmeted iguanids (*Corythophanes*), the casque-headed iguanids (*Laemanctus*), and the basilisks (*Basiliscus*). The lizard family Teiidae, though represented in the United States by the race-runner genus (*Cnemidophorus*), is tropical, and its real development begins in Central America with the large, conspicuous, and active ameivas (*Ameiva*) and several small genera that live in concealment.

Among snakes, the fer-de-lance genus *Bothrops,* the coral snakes (*Micrurus*), the rear-fanged snakes such as the cat-eyed snakes (*Leptodeira*), and nonvenomous genera such as the tropical green snakes (*Leptophis*) either appear for the first time or begin their proliferation of species in Central America.

Reptiles become increasingly numerous in northern South America. Vine snakes (*Oxybelis* and *Imantodes*), false coral snakes (*Erythrolamprus*), slender ground snakes (*Drymobius*), and the burrowing spindle snakes (*Atractus*) are most abundant there. Most of the genera of the lizard family Teiidae occur in this area. Iguanid lizards of the anole genus (*Anolis*) are represented in northern South America by approximately 165 species. Other iguanid genera—*e.g.,* the long-legged *Polychrus*—make their appearance.

Crocodilians, in terms of species, are more numerous in South than in Central America, and turtles are also abundant. Some of the North American groups—for example, the mud turtles (*Kinosternon*) and sliders (*Chrysemys*)—are represented, but the majority of species are members of genera and even families (*e.g.,* the side-necked turtles, families Pelomedusidae and Chelidae) unknown in temperate North America.

Several groups that form important, if not dominant, elements of the fauna of the Eastern Hemisphere are largely or completely absent from the American tropics: the lizard families Scincidae, Lacertidae, Chamaeleontidae, and Agamidae and the snakes of the cobra (*Naja*) and water snake (*Natrix*) genera.

South of the tropics, in the temperate zone of South America, the reptilian fauna diminishes rapidly. Crocodilians and turtles do not occur south of northern Argentina. An ovoviviparous pit viper reaches almost 50° S there; two iguanid lizards range almost to 55° S.

Asia. Apart from the genera of reptiles listed above as common to the eastern United States and eastern Asia, the temperate zone of Eurasia is noted for its many lizards of the families Agamidae, Lacertidae, and to lesser degrees Gekkonidae and Scincidae. Most of the lizards are terrestrial; extremely specialized burrowers include desert-dwelling skinks (*Ophiomorphus* and *Scincus*). Most of the snakes characteristic of this vast area are also terrestrial. Arboreal snakes are represented almost exclusively by the rat snakes (*Elaphe*). The leaf-nosed snakes (*Lyterhynchus*) and the sand boas (*Eryx*) are the distinctive burrowing snakes of the region. Except for the Chinese alligator and the Indian gavial, temperate Eurasia lacks crocodilians. A few species of turtles are found.

A few types characteristic of the Oriental tropics extend into the temperate zone—*e.g.,* several rear-fanged snakes

(*Boiga trigonata* and *Psammodynastes*), a cobra or two (*Naja*), several species of soft-shelled turtles (*Trionyx*), and some species of true chameleons (*Chamaeleo*).

In the Oriental tropics the reptilian fauna is extremely rich in species and diverse types. Aquatic groups are represented by snakes of various genera (*e.g., Natrix, Enhydris, Acrochordus*), several groups of lizards (*Tropidophorus* among the skinks and *Hydrosaurus* among the agamids), many emydid and soft-shelled turtles, and five species of crocodiles. The numerous terrestrial reptiles include the small kukri snakes (*Oligodon*), the big Oriental rat snakes (*Ptyas*), cobras, monitor lizards (*Varanus*), many species and genera of skinks, some geckos, and several land turtles (*Cuora, Geochelone*). Specialized burrowing snakes (*e.g.,* the family Uropeltidae and the colubrid genus *Calamaria*) and lizards (*e.g.,* the family Dibamidae and the skink genus *Brachymeles*) are also abundant.

Arboreal reptiles of tropical Asia

The distinctive life-forms of reptiles in tropical Asia are arboreal. They include pythons and Oriental pit vipers (*Trimeresurus*), vine snakes (*Ahaetulla*), slug-eating snakes (*Pareas*), "flying" snakes (*Chrysopelea*), and tree racers (*Gonyosoma*). Some lizards climb only with the aid of claws (*e.g.,* the monitors), a few with the help of prehensile, or grasping, tails (*e.g.,* the deaf agamids, *Cophotis*), and many with the help of clinging pads under the digits (*e.g.,* many geckos). The most striking arboreal reptiles of this area are the flying lizards (*Draco*) and the parachuting gecko (*Ptychozoon*), which has fully webbed digits, a fringed tail, and wide flaps of skin along its sides.

Australia. Because New Guinea, although geographically part of the East Indies, has a reptilian fauna more akin to that of Australia, the two areas are considered here as one. The Australian region is the only area in the world in which venomous species of snakes outnumber harmless ones. The family Colubridae, comprising the majority of the nonvenomous or slightly venomous snakes of the world, is poorly represented in Australia, which has only 12 species. The Australian region has many snakes of the cobra family (family Elapidae), but no vipers. The fauna also include several pythons and minute blind snakes (family Typhlopidae); a variety of geckos, skinks, and agamid lizards; side-necked turtles; and three species of crocodiles.

Africa. The reptilian fauna of Africa forms two main divisions. The first, the fauna of the North African coast, is akin to that of central and southwestern Asia and southern Europe and is therefore mainly a temperate-zone fauna. The racers, the burrowing sand skink (*Scincus*), and the emydid turtle (*Mauremys caspica*) are elements of temperate fauna in North Africa. Some species of the great tropical fauna lying south of the Sahara Desert occur in North Africa and in Southwest Asia. Examples are the sand snakes (*Psammophis*), cobras, and chameleons (family Chamaeleontidae). As is true of the temperate fauna of Eurasia, the North African reptiles, though representing many families, are principally terrestrial and burrowing. Many lacertid and agamid lizards scamper over rocks and sand by day; they are replaced at night by small geckos and are preyed upon by the racers (*Coluber*) and sand snakes (*Psammophis*). In addition to cobras, the venomous snakes of North Africa include the common vipers, the saw-scaled viper (*Echis carinatus*), and the horned vipers (*Cerastes*). The last two are true desert animals. Land tortoises (*Testudo*) are common in the semi-arid land.

Reptiles of tropical Africa

The second and much larger division of the African fauna is the great tropical assemblage that ranges from the Sahara southward to the Cape of Good Hope. In common with tropical Asia, this vast area has cobras, many skinks, and many geckos. Its fauna differs from that of Asia in the absence of pit vipers (subfamily Crotalinae), the near absence of emydid turtles, and the poor representation of agamid lizards. These groups are replaced in tropical Africa by the many true vipers (subfamily Viperinae), the side-necked turtles (family Pelomedusidae), and the lacertid and cordylid lizards. Chameleons and land tortoises are abundant. Three species of crocodiles occur in Africa.

In Africa are found all of the diverse reptilian types characteristic of a tropical area: aquatic turtles, crocodiles, and snakes; terrestrial turtles, snakes, and lizards; burrowing snakes of the blunt-headed and auger types; limbless and virtually blind burrowing lizards; and a profusion of arboreal snakes and lizards.

Unique reptile fauna of Madagascar

The large island of Madagascar, off the eastern coast of Africa, has a peculiar fauna that appears in part as a collection of castoffs, groups that in Africa have not been able to meet the competition of more advanced forms. With few exceptions the reptiles of Madagascar belong to genera found only there.

Natural history

Reproduction and life cycle. *Courtship.* Courtship in some form is such a widespread prelude to mating among modern reptiles that it must have characterized many extinct groups as well. When courting, the male of some freshwater turtles, such as the red-eared turtle (*Chrysemys scripta elegans*) of the eastern United States, orients himself in the water so that he is directly in front of a female and facing her. With his forefeet close together, the male vibrates his claws against her head. If the female is receptive, she swims forward slowly while the male backs away. Finally the female sinks slowly down. The male then mounts her from behind, clutching her shell with all four feet. His tail is brought under hers and his penis introduced into her cloaca.

The male of some terrestrial turtles of North America (*e.g.,* the gopher tortoises, *Gopherus*) begins courtship by extending his neck and bobbing his head up and down. The courted female may bob her head in return. The male advances, nips at the female, and then circles her as she turns away from him. As soon as she shows signs of response, the male mounts her from the rear and begins a series of pumping movements that bump the rear of his shell against the ground. Finally the female extends her tail, and copulation begins. Males of the smaller box turtles (*Terrapene*) nip and butt at the female.

Mating behaviour of crocodilians

Male crocodilians bellow during the mating season, and in many cases the females respond with an answering call. The male American alligator, which copulates in water, grasps the female's neck with his jaws and slips the rear part of his body under hers to enable him to insert his penis into her cloaca.

Lizards have rather elaborate courtship patterns usually involving display and posturing by the males, who often have distinctive patches of colour on their throats or low on their sides. The male bobs up and down, thus exposing the patches of colour, which may be blue, orange, red, or black depending on the species. Males of some species, such as the green anole (*Anolis carolinensis*) of the southeastern United States, have brightly coloured folds of skin at the throat (dewlaps) that are expanded during courtship. If the female seems receptive, the male straddles her back, often gripping her back or legs with his jaws. Just before copulation, his tail is bent under hers.

The courtship patterns of snakes are simpler; they usually consist of the male's crawling over the back of the female and adopting every curve her body takes, then vibrating against her body or nudging it with waves of his own. Male boas and pythons stroke or scratch the female's body with their vestigial hind limbs. The male water snake rubs his chin against the female's back. The male rattlesnake (*Crotalus*) frequently nudges the female with his head. The male bull snake (*Pituophis*) grasps the female's neck with his jaws during copulation.

When the male king cobra (*Ophiophagus hannah*) crawls onto the back of a receptive female, he flicks his tongue against her repeatedly. The female raises her head and spreads her hood. The male nudges her neck and head with his snout and lifts the rear of her body with his tail. Copulation between cobras may last more than two hours.

The amnion

The embryo. The embryo of a reptile develops a thin sac, the amnion, that envelops the embryo and becomes filled with a watery fluid. The entire structure serves to protect the embryo from desiccation and mechanical injury. A parchment-like shell, which also contributes mechanical protection, is produced by the female parent and surrounds the amnion; between the shell and the amnion a second sac, the allantois, becomes inserted. The allan-

tois, which is supplied with many fine blood vessels, serves as a respiratory organ, absorbing the oxygen and emitting the carbon dioxide that pass through the somewhat porous shell.

Egg laying. The typical mode of reptilian reproduction is oviparous (*i.e.,* the female lays eggs in which the young develop). The eggs are laid shortly after fertilization, and development of the embryos takes place largely after the eggs have been laid. This pattern characterizes crocodilians, turtles, the tuatara, most lizards and snakes, and many extinct reptiles. The size of eggs laid by lizards and snakes varies according to the size of the females. The banded rock lizard (*Petrosaurus mearnsi*) of the western United States, which ranges from 7.5 to 10 centimetres (three to four inches) in length, lays eggs that are about one centimetre (0.4 inch) long; those of the 30- to 60-centimetre (12- to 24-inch) ringneck snake (*Diadophis punctatus*) from the eastern United States are 1.25 centimetres (0.5 inch) long. Eggs of the three-metre Komodo dragon lizard (*Varanus komodoensis*) and of the six-metre (20-foot) Indian python (*Python molurus*) are about 11.25 centimetres (4.5 inches) long.

A minority of modern and extinct reptiles are (and were) live-bearing, or viviparous. Strictly speaking, most of this minority are not truly viviparous but ovoviviparous, because the embryos develop with their shells or shell membranes intact and are nourished wholly by the yolk. In a few modern reptiles the embryonic membranes and the tissues lining the oviducts of the females come into close contact and are modified in one of several ways to provide a temporary organ, through which food and respiratory gases are exchanged; *i.e.,* a structure similar to the placenta of mammals. In the simplest reptilian "placenta," the most superficial layer (the lining of the oviduct partially degenerates, thereby bringing the blood vessels of embryo and mother closer together. The approximation of the two bloodstreams facilitates the exchange of oxygen and carbon dioxide, this gas exchange being the only function of the organ at this stage of evolution. Several Australian snakes (*Denisonia superba* and *D. suta*) and a number of lizards—*e.g.,* the common East Indian brown-sided skink (*Mabuya multifasciata*) and the cylindrical skink (*Chalcides ocellatus*) of southern Europe and North Africa—are known to have this type of organ, as presumably do many ovoviviparous reptiles.

The best developed reptilian "placentas" consist of apposed, thickened, folded elliptical areas of the outer embryonic membrane and lining of the oviduct. The ridges of the oviductal areas are filled with blood vessels, and the epithelium between ridges is thickened and glandular. Usually, eggs developing with this type of "placenta" have less yolk; food and oxygen are transmitted from mother to embryo. Several species of Australian lizards, American water snakes, and the common European viper (*Vipera berus*) are known to provide this type of internal environment for their developing young.

Oviparity and ovoviviparity in lizards and snakes

The line between oviparity (egg laying) and ovoviviparity (hatching of eggs in the mother's body) is arbitrary. Females of some lizards and snakes retain the fertilized eggs in their bodies for a few days before laying them. Other species retain the eggs for most of the developmental period, hatching occurring shortly after laying. For the grass snake of England (*Natrix natrix*), the lapse of time between copulation and egg laying is usually two months; the young hatch six to ten weeks later. The interval between mating and egg laying is one month in the Texas horned lizard (*Phrynosoma cornutum*). A given species may be ovoviviparous in parts of its range and oviparous elsewhere.

The nest. The eggs of modern reptiles may be deposited in a nest prepared by the female or simply laid under some convenient cover, such as a rock or log. Crocodilians invariably prepare a nest, and the female invariably does the work. Most turtles dig their nests, scooping out a flask-shaped cavity in the ground with their hindfeet. When the hole has reached the proper size, the oval or spherical eggs, 1.5 to 3.75 centimetres (0.6 to 1.5 inches) in diameter depending on the species, are dropped into the nest from the female's cloaca. The female scratches soil over the eggs, usually obliterating the nest site. Crocodilians either dig a nest along the bank of a river or lake or heap together a mass of dead vegetation in which the eggs are laid; their oval eggs are usually about five centimetres (two inches) long.

Most oviparous lizards merely hide their eggs under some convenient cover such as under a rock or in a hole in a tree. Nest construction among lizards, though appearing in such diverse families as the iguanids, skinks, and true chameleons, is neither so elaborate nor so rigid in pattern as among turtles. The nest consists of a small hole made by either the snout or the limbs. Soil or leaves usually are pushed on top of the eggs to hide the nest, although the entrance to the nest cavity is kept open in a few species. Snakes, like lizards, usually lay their eggs under natural, pre-existing cover. The king cobra, one of the very few nest-building snakes, drags dead vegetation into a low heap by bending its body. The eggs are laid in a cavity at the centre. Other snakes deposit their eggs in holes they have scooped out of sand or soft earth with their snouts.

Number of offspring. The number of eggs in a clutch or offspring in a brood varies from one to 200 among living reptiles; presumably similar variation occurred among the extinct types. Crocodilians lay from 20 to 70 eggs, turtles from one to 200. In turtles, more so than in crocodiles, the number varies with the species and roughly with the size attained by the females. The big marine turtles have the largest clutches (usually more than 100); the smaller land and freshwater turtles have much smaller ones. The number of eggs or young is not so closely related to the size of mature females in species of lizards and snakes. With few exceptions, lizards of the family Gekkonidae lay two eggs at a time, regardless of the size of the female. Lizards of the family Scincidae have broods varying from two to about 30; one of the largest members of this family, the 30-centimetre-long (12-inch) stump-tailed skink (*Tiliqua rugosa*) of Australia, has only two young at a time, whereas the Great Plains skink (*Eumeces obsoletus*) of the United States usually lays between ten and 20 eggs in a clutch.

Relationship of brood size to adult size

Clutch size in the 1.5-metre (five-foot) bull snake (*Pituophis catenifer*) of western United States is usually ten or 12; in the grass snake of England and Europe, which measures 60 to 90 centimetres (two to three feet), it is 30 to 40; in the giant reticulated python of Southeast Asia and the East Indies, it may reach 100.

Parental care. Parental care of eggs and newborn young is neither well developed nor elaborated among reptiles. Female crocodilians generally remain in the vicinity of their nests and chase would-be predators from the site. In a few lizards the female returns to the nest between feeding excursions to coil around the eggs and turn them at intervals. The male and female king cobra remain in the vicinity of the nest, and one of the parents usually is coiled above the egg cavity.

Female pythons coil around their eggs and pull them into a heap. Females of some species remain with the eggs for the entire two-month incubation period; others leave the eggs only to drink. In at least one species (*Python molurus*) the female provides heat by muscular contraction to keep the eggs at incubation temperature on cool nights. Turtles and the majority of egg-laying lizards and snakes abandon their eggs after they are laid.

Incubation period. The incubation (or gestation) period of reptilian eggs is affected by many factors and to such an extent that it is difficult to assign a figure characteristic of a given species. One source of complication is the combination of oviparous and ovoviviparous habits by certain species. The developmental period of the embryo, whether it occurs within the female's body or outside it, is referred to as the gestation period.

In general, the gestation period lasts from 60 to 105 days in most American and European reptiles. The eggs of the American alligator hatch about 63 days after they are laid, those of the small Eastern fence lizards (*Sceloporus undulatus*) in about the same time. The gestation period in the common European viper lasts from 60 to 90 days. Eggs of marine turtles hatch between 30 and 75 days after they are laid, depending on the site. The temperatures to which a brood is subjected shorten or lengthen gestation according to whether the temperatures are high or low.

Growth and longevity. Giant Galápagos tortoises kept under nearly ideal conditions have been known to increase their weight from 3.2 to six kilograms (seven to 13 pounds) to about 82 kilograms (180 pounds) in nine years. Smaller species also grow rapidly. The box turtle of the United States has a shell about 3.75 centimetres (1.5 inches) long at the end of its first year; at the end of five years, the length has doubled.

Under favourable conditions a one-year-old American alligator is about 60 centimetres (24 inches) long and weighs about 1.8 kilograms (four pounds). At the end of six years, males average about 190 centimetres (about six feet) and about 36 kilograms (80 pounds). The red diamond rattlesnake is about 30 centimetres long at birth, grows to about 65 centimetres in its first year, reaches about 85 centimetres (33 inches) by the end of its second year, and grows more slowly after that. The pattern for lizards is much the same: rapid growth early in life and slow growth afterward. The significant difference between growth in reptiles and that in mammals is that a reptile has the potential of growing throughout its life, whereas a mammal reaches a terminal size and grows no more, even though it may subsequently live many years in ideal conditions.

Achievement of sexual maturity

The length of time needed to attain sexual maturity varies greatly among reptiles and, although roughly related to the size usually attained by the species, is even more closely related to the climate in which the animal lives. The red diamond rattlesnakes in southern California, for example, bear their first young when three years old; on the other hand, the northern Pacific rattlesnake (*Crotalus viridis oreganus*) bears its first litter when four years old. The much smaller common garter snake is sexually mature shortly before the age of two years.

The 12.5- to 15-centimetre (five- to six-inch) northwestern sagebrush lizard (*Sceloporus graciosus gracilis*), living in the Sierra Nevada range of California at an elevation (about 1,800 metres [6,000 feet]) where it has, at most, six months of activity each year, requires two years or more to reach sexual maturity. Another lizard, the green anole (*Anolis carolinensis*), similar in size to the sagebrush lizard but living in the lowlands of the southern United States, may reach maturity in four or five months in Florida.

Turtles mature at a slower rate. Females of the red-eared turtle of the central United States lay their first eggs when they are from three to eight years old, depending upon how long it takes them to reach a shell length of 15 centimetres (six inches). Females of the musk turtle, or stinkpot (*Sternotherus odoratus*), in Michigan require nine to 11 years to mature, at which time their shells are 7.5 to 10 centimetres (three to four inches) long. Presumably, turtles living in the tropics mature more rapidly.

The maximum age, meaning the potential longevity, of modern reptiles varies greatly and can be determined only from records of captive animals. Turtles as a group seem capable of living longer than the others, and about 30 species have been kept in captivity more than 20 years. Several species, said to have lived 150 years or more, may be cases of two individuals whose periods of captivity overlapped. There is no reliable evidence for believing that the giant land tortoises live much longer than some smaller species. Two crocodilians (*Alligator mississippiensis* and *A. sinensis*) have survived in zoos for more than 50 years. Several species of pythons and boas have lived longer than 20 years. Lizards seem to have an upper limit near that of snakes. A slowworm, *Anguis fragilis*, has been kept in captivity for more than 30 years.

Behaviour

Defense. *Avoidance and noise.* Avoidance, the commonest form of defense in the animal kingdom, is also the commonest one in reptiles. At the first recognition of danger, most snakes and lizards crawl or scamper away into the undergrowth; turtles and crocodilians plunge into water and sink out of sight. But should the danger arise so suddenly and so close at hand that flight may be hazardous, other expedients are adopted.

Crocodiles, some lizards, turtles, and some snakes hiss loudly when confronted by an enemy. Rattlesnakes rapidly vibrate the tip of the tail, which consists of loose, dry, horny rings. A few snakes without rattles (*e.g.*, the fox snake, *Elaphe vulpina*, of the United States) vibrate the ends of their tails rapidly, and if, as often happens, the tail hits dry leaves, it makes a sound deceptively like the rattle of a rattlesnake.

Hissing and tail rattling

Body form and posturing. Change in body form, which is relatively common in snakes, usually involves spreading the neck, as in cobras (family Elapidae), or the whole body, as in the harmless hognose snakes (*Heterodon*) and DeKay's snake (*Storeria dekayi*) of the United States. Some snakes inflate the forward parts of their bodies; inflation is one of the defensive actions of the large South American tree snake *Spilotes* and of the African boomslang (*Dispholidus*).

Threatening postures may be assumed by snakes as they change their body form. A cobra raises the forepart of its body and spreads its hood when endangered. The typical defensive posture of a viper is with the body coiled and the neck held in an S-curve, the head poised to strike.

Some lizards flatten their bodies, puff out their throats, and turn broadside to the enemy. The helmeted iguanids (*Corythophanes*) of Central America and the chameleons of Africa increase their apparent size in this way when approached by snakes. The Australian bearded lizard (*Amphibolurus barbatus*) spreads its throat downward and outward. The Australian frilled lizard (*Chlamydosaurus kingi*) suddenly raises a wide membrane, or frill, which extends backward from the throat. Many lizards and snakes open their mouths when threatened, but do not strike. A common African lizard, *Agama atricollis*, faces an enemy with head held high and mouth open to show the brilliant orange interior.

Display of colour. Display of colour in *Agama atricollis* may not be part of a threatening mechanism, but it is so in the instances of certain red- or yellow-bellied snakes that turn over or curl up their tails, exposing the brightly coloured undersurface. This behaviour, known in harmless (*e.g.*, the American ring-necked snake, *Diadophis*) as well as venomous snakes (*e.g.*, the coral snake, *Micrurus frontalis*), is displayed only by snakes having red, orange, or yellow undersides. These colours must have some significance, as yet not fully understood, to predacious animals, for they are also the common colours in insects having warning coloration.

The defense mechanism of camouflage involving form and colour is common. Many arboreal snakes and lizards (*e.g.*, chameleons) are green; some of the green snakes (*e.g.*, the vine snakes of South America, *Oxybelis*, and of southern Asia, *Ahaetulla*) are very slender, resembling plants common in the habitat. Lizards of semi-arid and rocky country frequently are pale in colour and blotched in pebble fashion—*e.g.*, the leopard lizard (*Crotaphytus wislizeni*) of the southwestern United States.

Mimicry of dangerous species by harmless snakes is a passive defense. Its validity as an actual mechanism of defense is, however, sometimes challenged. The venomous coral snakes (*Micrurus*) of the Western Hemisphere are ringed with bright red, yellow, and black. A series of relatively harmless snakes, such as *Erythrolamprus* and *Anilius* of South America and the scarlet king snake, *Lampropeltis triangulum doliata*, of southeastern United States, have similar colours and patterns that may confer some protection against predators.

Mimicry

Striking and biting. If a threatening posture does not succeed in driving off an enemy, many reptiles become more aggressive. Some snakes (*e.g.*, DeKay's snake) strike, but with their mouths closed. Others (*e.g.*, the hognose snakes) strike with their mouth open but do not bite. Still others strike and bite viciously. Among the nonvenomous snakes of North America, few are as quick to bite as the water snakes (*Natrix*). The sole danger from the bites of these snakes is infection of the wound.

Most of the dangerously venomous snakes (vipers, pit vipers, and cobras) bite in self-defense. Vipers and pit vipers usually strike from a horizontally coiled posture. From this position the head can be shot forward, stab the enemy, and be as rapidly pulled back in readiness for

the next strike. From the typical raised posture a cobra sweeps its head forward and downward to bite. To strike again it raises its head and neck once more; such aggressive, defensive movements of cobras are slower than those of pit vipers.

Many lizards, regardless of family and size, bite in defense. *Gekko gecko* of Southeast Asia bites if sufficiently threatened. Although small lizards have a bite effective against only the smallest predators, a large monitor lizard (*Varanus*) can inflict a painful wound with its large teeth and strong jaws. Some turtles, particularly the soft-shelled turtles (*Trionyx*), bite frequently, vigorously, and effectively.

Spitting. The spitting of venom by certain African cobras, the ringhals (*Hemachatus haemachatus*), and the black-necked cobra (*Naja nigricollis*) is a purely defensive act directed against large enemies. A fine stream of venom is forced out of each fang, which, instead of having a straight canal ending in a long opening near the tip as in most cobras, has a canal that turns sharply forward to a small round opening on the front surface well away from the tip. At the moment of ejection the mouth is opened slightly, and venom is forced out of the fangs by contraction of the muscle enveloping the poison gland. Usually a spitting cobra raises its head and the forepart of its body in the characteristic cobra defensive posture prior to spitting, but venom can be ejected from any position. The effect on skin is negligible; the eyes, however, may be severely damaged, and blindness can result unless the venom is washed out quickly.

Use of the tail. A few lizards, representing different families, have in common thick tails covered by large, hard, spiny scales. Such a tail swung vigorously from side to side is an effective defense against snakes, especially when the head and body of the lizard are in a burrow or wedged between rocks.

Voluntary tail shedding

Lizards' tails are useful in defense in another way. When captured, many lizards voluntarily shed their tails, which wriggle violently, temporarily confusing the predator and allowing the lizard to escape. Each vertebra of the tails of lizards with this capacity has a fracture line and can be split on that line when tail muscles contract violently. Simultaneous stimulation of the nerves in the severed portion keeps it twitching for a few seconds after separation. Usually the tail is broken in only one place, but a few lizards, particularly the so-called glass snakes (*Ophisaurus*), break their tails into several pieces. The stump heals quickly, and a new tail grows; often, however, the regenerated tail is not so long as the original and has simpler scales.

Snakes, turtles, and crocodiles may have their tails bitten off by predators, but they cannot break them voluntarily or regenerate them. Some snakes use their tails in diversionary tactics by raising them and moving them slowly. Species with this habit commonly have thick, blunt, brightly coloured tails. The small African python *Calabaria* and the Oriental venomous snake *Maticora* wave their tails in the air as they move slowly away from a threat.

Balling. Many snakes, both harmless and venomous, attempt to hide their heads under coils of their bodies. The body may be coiled loosely, as it is in most species with this habit, or tightly so that it forms a compact ball with the head in the centre. Balling, as the latter habit is called, is a characteristic response of *Calabaria* and another African python, *Python regius*.

The African armadillo lizard (*Cordylus cataphractus*), a species with heavy scales on its head and hard spiny scales covering its body and tail, rolls on its back and grasps its tail in its mouth. It thus presents a ring of hard spines to a predator.

Odours. Some reptiles use musk-secreting glands when other defensive measures fail. The water snakes (*Natrix*), the garter snakes (*Thamnophis*), the alligator lizards (*Gerrhonotus*), and the musk turtles (*Sternotherus*) emit a foul-smelling substance from anal glands.

Feeding habits. With few exceptions, modern reptiles feed on some form of animal life: insects, mollusks, birds, frogs, mammals, fishes, or other reptiles. Land tortoises are vegetarians, eating leaves, grass, and in some cases even cactus. The big green iguana (*Iguana iguana*) of Central and South America, its relative the chuckwalla (*Sauromalus obesus*) of southwestern United States and northern Mexico, and the spiny-tailed agamids (*Uromastix*) of North Africa and southwestern Asia also are herbivorous. The marine iguana (*Amblyrhynchus cristatus*) of the Galápagos Islands dives into the sea for seaweed.

The majority of carnivorous reptiles have nonspecialized diets, feeding on a variety of animals. In general, the smaller the reptile, the smaller is its prey. Only the very largest of living snakes—the reticulated python (*Python reticulatus*), the Indian python (*P. molurus*), and the anaconda (*Eunectes murinus*)—are capable of eating large mammals such as small pigs and deer. Among crocodiles the largest species—the Nile crocodile (*Crocodylus niloticus*), the East Indian saltwater crocodile (*C. porosus*), and the Orinoco crocodile (*C. intermedius*)—have been known to attack and to eat men. Presumably, even larger prey was devoured by the great carnivorous dinosaurs such as *Allosaurus* and *Tyrannosaurus*, which were almost certainly capable of killing the largest of their herbivorous contemporaries.

Man-eating reptiles

Locomotion. *Walking and crawling.* The majority of reptilian orders are quadrupedal—*i.e.*, four-legged. Among the land vertebrates, the limbs gradually shifted from a lateral to a ventral position. In most amphibious reptiles the limbs projected out to the side and then bent downward to the ground at the knee and elbow. With few exceptions, the quadrupedal reptiles have the same awkward position. With such an orientation, the centre of gravity of the body is not in the same axis as the hands and feet, resulting in a sideways as well as a forward component of thrust when the animal walks. The typical reptile throws its body into a slight horizontal curve to progress straight forward. In mammals the limbs are directly underneath the body, the centre of gravity is in the axis of the limbs, and all of the thrust of the limbs is directed forward. The latter position and type of motion are more efficient. The lateral orientation of the limbs in amphibians and reptiles also makes it more difficult to raise the body off the ground.

Despite the awkwardness of the orientation of their limbs, some reptiles are (and many extinct forms probably were) capable of moderate speeds. Crocodilians raise their bodies off the ground and make short, fast rushes. Short-bodied lizards also can move fast for short distances; longer bodied lizards have greater difficulty in raising their bodies. They usually have short legs and proceed in a serpentine fashion, with the body, thrown in horizontal curves, doing much of the work.

A snake moves by pushing backward against rocks, sticks, or any relatively fixed point—a lump of earth or a small depression in uneven ground—with the rear surface of the horizontal curves of its body. Each joint of the body passes through the same curves, pressing against the same object and thrusting the forepart of the body forward. Heavy-bodied snakes such as pythons and certain rattlesnakes can move forward without throwing their bodies into curves. This rectilinear movement depends on the ability of snakes to stretch or contract their bodies in the longitudinal axis. By raising a part of its belly, stretching that part forward, lowering it to the ground, and repeating the process alternately with other parts of the body, a heavy snake moves forward smoothly in a straight line.

Bipedal locomotion

Some modern lizards have adopted semi-bipedal locomotion. The collared lizard (*Crotaphytus collaris*) of the United States and the frilled lizard (*Chlamydosaurus kingi*) of Australia show the early stages of bipedalism, a phenomenon widespread among the dinosaurs and therefore important in reptilian history. These lizards run on their long hindlegs with the forward parts of their bodies at an angle of about 60° off the horizontal.

Presumably, bipedalism among the dinosaurs began as it did among modern lizards, as an occasional means of obtaining bursts of speed. Because the centre of gravity is in front of the hips, modern bipedal lizards must move forward continuously in order to maintain a semi-erect posture; they can stand still in that position only for very short periods.

The awkward sideways orientation of the limbs forces bipedal lizards to swing each leg outward as it is brought forward and to push the body sideways and forward when each leg thrusts backward against the ground. Bipedal dinosaurs eliminated this inefficient rocking motion, for during the course of evolution their hind limbs were rotated forward so that they were directly under their bodies. Thus, they delivered their full force in the forward direction. So successful was this mode of locomotion that dinosaurs utilizing it dominated terrestrial life for millions of years.

Clinging and climbing. Associated with arboreal life are groups of anatomical features mainly concerned with clinging. The commonest clinging structures in vertebrates are claws; they seem to be the only arboreal adaptations of some lizards, such as the common iguana (*Iguana iguana*). Similar structures appear in many lizards of the family Gekkonidae, in the anoles (*Anolis*) of the family Iguanidae, and in some skinks of the family Scincidae.

Pads on the feet consist of wide plates or scale under the fingers and toes. The outer layer of each plate or scale is composed of innumerable tiny hooks formed by the free, bent tips of cells. These minute hooks catch in the slightest irregularities and enable geckos to run up apparently smooth walls and even upside down on plaster ceilings. Because the hooklike cells are bent downward and to the rear, a gecko curls its toes upward to disengage them. Thus, when walking or running up a tree or wall, a gecko must curl and uncurl its toes at every step.

Prehensile tails

The giant Solomon Islands skink (*Corucia*), true chameleons (Chamaeleontidae), arboreal vipers, boas, and pythons use prehensile tails—that is, tails capable of supporting most of the weight of the animal or used habitually for grasping—for clinging to their aerial supports. For this purpose, however, true chameleons rely mainly on a tonglike arrangement of their digits, which are united into two opposed bundles on each foot—three on the inside and two on the outside of the front foot, and two on the inside and three on the outside of the hindfoot.

Slender vine snakes of several genera of the family Colubridae are capable of extending half the body length in a horizontal plane without support; they do so habitually in bridging the gap between branches. Most snakes can reach across an open space, but all except the vine snakes can extend only a short length of the body, and that portion invariably sags like a cable. The vine snakes bridge an open space like an I-beam. This ability is based partly on reduced body weight and partly on deepened and strengthened vertebrae.

Swimming. In water, of course, neither bipedal nor quadrupedal locomotion is very effective. Aquatic reptiles, with few exceptions, use the same means of propulsion as do fish and whales—that is, powerful beats of the tail. Crocodilians and aquatic lizards such as some monitors (Varanidae) lash their tails from side to side while holding the limbs against the body. The same method was used by the ancient mesosaurs (Mesosauria) and ichthyosaurs (Ichthyosauria). The marine ichthyosaurs, which were the reptilian counterpart of the porpoises, may have used their very short limbs for steering.

A fishlike method of swimming requires a flexible body and at least a moderately long tail. Turtles propel themselves by using their feet as paddles—the hindfeet, which have webbed toes, in the case of freshwater turtles, and the forefeet, which are modified into large paddles, in the case of marine turtles.

The extinct marine plesiosaurs (suborder Plesiosauria), with their short bodies and tails and their large paddlelike limbs, swam the way marine turtles do, although they may have used their hindlimbs for more than just steering. Both pelvic and pectoral (shoulder) girdles were modified in the plesiosaurs into structures having small upper portions and very large lower portions. As the upper element, especially that of the pelvic girdle, has the important function of transferring the weight of the body to the limbs, it is likely that the limbs of plesiosaurs could not support the body weight on land and that the plesiosaurs never came out of water.

Most plesiosaurs had long necks. By moving toward their prey with the neck curved, they probably could strike suddenly. The heavy trunk would provide the inertia against which the neck could move, thus preventing a significant backward shift of the animal as the head shot forward. The modern sea snakes (Hydrophidae) show the same adaptation. Though they swim with an eel-like undulation of the body, the sea snakes have relatively small heads, slender necks, and very heavy middle and rear sections. With most of the body mass concentrated in the second half of the animal, almost all of the force of the strike is used to drive the head forward.

Locomotion of sea snakes

Flying. Three groups of reptiles have experimented with flight. Thecodontia, a group of Archosauria (the so-called ruling reptiles, which included dinosaurs and crocodilians), became highly successful at this means of locomotion and evolved into birds.

A second group of archosaurs, the Pterosauria, developed wings that were supported along the front margin by the arm and an extremely elongated finger. The pterosaur wing was made of skin; since it lacked both internal supports and feathers, it probably lacked the flexibility or durability of a bird wing. Flight of the pterosaurs presumably amounted to soaring and gliding. It is not understood how they moved when not flying and how they managed to take off if they happened to land on level ground. Since most remains have been found in marine deposits, it is assumed that they lived along ocean shores, probably roosting on cliffs from which takeoff would have been easy.

"Flying" lizards

The third experiment with flight was made by a group of modern lizards (*Draco*). The "wing" of these small lizards consists of skin supported by five or six elongated ribs between the arm and leg. At rest the ribs and the wings are folded against the sides of the body. In flight the wings form broad semicircles from arm to leg on each side. These flying lizards, which live in the forested country of Southeast Asia and the East Indies, are capable only of gliding. A flying lizard launches itself from a tree into the air and glides toward another tree, turning upward sharply just before lighting on the new perch. Since the arms and legs are not modified, this lizard is capable of scampering about like any strictly arboreal lizard.

Form and function

External covering. The external covering of reptiles is characteristically dry. It bears few glands or none at all and differs in this respect from the skin of amphibians and mammals. The so-called malpighian layer of the epidermis secretes the outer layer, which is tough and horny. Bony plates develop in the dermis, which lies just below the epidermis. The arrangement of scales is usually characteristic for each species.

Internal features. *Skeletal system and dentition.* The skeleton of reptiles fits the general pattern of vertebrates. They have a bony skull, a long vertebral column that encloses the spinal nerve cord, ribs that form a bony basket around the viscera, and a framework of limbs.

Each group of reptiles developed its own particular variations on this major pattern in accord with the general adaptive trends of the group. Snakes, for example, have lost the limb bones, although a few retain vestiges of the hindlimb. The limbs of several types of marine reptiles became modified into fins or flippers with obvious functional significance. In groups such as the extinct ichthyosaurs and plesiosaurs, the bones of the limbs, no longer supporting the weight of the body against the pull of gravity, became much shortened. At the same time the bones that in other reptiles composed the digits multiplied in number, forming a long flipper.

Groups of reptiles whose modes of life came to depend heavily on passive defense also developed specializations of the skeleton. The bony and horny shell of turtles and rows of bony plates on the back of ankylosaurs (Cretaceous dinosaurs) are cases in point.

The skulls of the several subclasses and orders vary in the ways mentioned below. In addition to differences in openings on the side of the skull and in general shape and size, the most significant variations in reptilian skulls are those affecting movements within the skull.

Reptilian skulls as a group differ from those of early amphibians, the vertebrates from which reptiles arose, in lacking on otic notch (an indentation at the rear of the skull) and several small bones at the rear of the skull roof. The skulls of modern reptiles are sharply set off from those of mammals in many ways, but the clearest differences are in the lower jaw and adjacent regions. Reptiles have a number of bones in the lower jaw, only one of which, the

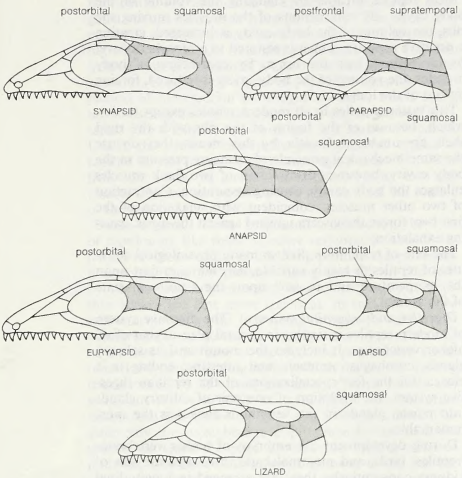


Figure 2: Diagrammatic reptilian skulls.

Differences between reptile and mammal skulls

dentary, bears teeth. Behind the dentary a small bone, the articular, forms a joint with the quadrate bone near the rear of the skull. In mammals the lower jaw consists of a single bone, the dentary, and the articular and quadrate have become part of the chain of little bones in the middle ear. An almost complete transition between these two very different arrangements is known from fossils of mammal-like reptiles (order Therapsida).

The dentition of most reptiles shows little specialization along the tooth row. A division into distinctive bladelike incisors, tusklike canines, and flat-crowned molars, such as characterize mammals, does not occur in reptiles. Instead, the entire tooth row usually consists of long conical teeth. Venomous snakes have one or several hollow or grooved fangs, but they have the same shape as most snake teeth. The principal differences between species lie in the number, length, and position of the teeth. Crocodilians among the living forms and dinosaurs among the extinct forms have but a single upper and a single lower tooth row. Snakes and many extinct reptilian groups have teeth on the palatal bones (vomer, palatine, pterygoid) and on the bones of the upper jaw (premaxilla, maxilla); only one row of teeth is present on the lower jaw.

Lizards have conical or bladelike bicuspid or tricuspid teeth. Some species have conical teeth at the front of the jaws and cuspid teeth toward the rear, but the latter are not comparable to the molars of mammals in either form (they are not flat-crowned) or function (they do not grind food). Turtles, except for the earliest extinct species, lack teeth, having instead upper and lower horny plates that serve to bite off chunks of food.

The teeth of reptiles are also less specialized in function than are mammalian teeth. The larger carnivorous reptiles are equipped only to tear off or bite off large pieces of their prey and to bolt them without chewing. Insectivorous lizards (the majority of lizards) usually crack the exoskeleton of their insect prey, and then swallow the prey without grinding it up. Snakes simply swallow their prey whole without any mechanical reduction.

Skull and joint structures. Many reptiles developed joints (in addition to the hinge for the lower jaw) within the skull, permitting at least slight movement of one part relative to others. The capacity for such movement within the skull, called kinesis, enables an animal to increase the gape of the mouth and thus is an adaptation for swallowing large objects. Apparently some of the large carnivorous theropod dinosaurs *(e.g., Allosaurus)* had a joint between the frontal and parietal bones in the roof of the skull. All reptiles of the subclass Lepidosauria (lizards, snakes, rhynchocephalians, and the extinct eosuchians) have had kinetic skulls, but they differ from the dinosaurs in having the joint on the floor of the skull at the juncture of basisphenoid and pterygoid bones.

Kinesis

The skulls of the lepidosaurians became increasingly kinetic as new groups evolved. The Rhynchocephalia (which include the living tuatara) and their antecedents, the Eosuchia, had only the basisphenoidal–pterygoidal joint. The lizards lost the lower temporal bar, thereby freeing the quadrate bone and allowing greater movement to the lower jaw, which is hinged to the quadrate. Finally, in the snakes, this trend culminates in the most kinetic skull among the vertebrates—a skull having the ancestral basisphenoidal–pterygoidal joint, a highly mobile quadrate (which gives even greater mobility to the lower jaw), upper jaws capable of rotating on their longitudinal axes and of moving forward and backward, and often a hinge on the roof of the skull between the nasal and frontal bones that allows the snout to be raised slightly. In short, the only part of a snake's skull incapable of movement is the braincase.

Nervous system. As in all vertebrates, the nervous system of reptiles consists of a brain, a spinal nerve cord, nerves running from the brain or spinal cord, and sense organs. Reptiles have small brains compared with mammals. The most important difference between the brains of these two vertebrate groups lies in the size of the cerebral hemispheres, the principal associative centres of the brain. In mammals these hemispheres make up the bulk of the brain and, when viewed from above, almost hide the rest of the brain. In reptiles the relative and absolute size of the cerebral hemispheres is much smaller. The brain of snakes and alligators forms less than $^{1}/_{1,500}$ of the total body weight, whereas, in mammals such as squirrels and cats, the brain accounts for about $^{1}/_{100}$ of the body weight. A stegosaur (Stegosauria), roughly the size of an elephant, had a braincase no larger than that of a 2.4-metre (8-foot) crocodile, about large enough to contain a brain the size of a large walnut.

Circulatory system. Modern reptiles do not have the capacity for rapid sustained activity found in birds and mammals. It is generally accepted that this lower capacity is related to differences in the circulatory and respiratory systems. Before the origin of lungs, the vertebrate circulatory system had a single circuit: in the fishes, blood flows from heart to gills to body and back to the heart. The heart consists of four chambers arranged in a linear sequence.

Evolution of the reptile heart

With the evolution of lungs in amphibians, a new and apparently more efficient circulatory system evolved. Two chambers of the heart, the atrium (or auricle) and ventricle, became increasingly important, and the beginnings of a double circulation appeared. An early stage in this evolution can be seen in amphibians today, where one of the main arteries from the heart (the pulmonary artery) goes directly to the lungs, whereas the others (the systemic arteries) carry blood to the general body. The blood is aerated in the lungs and carried back to the atrium of the heart. From the left side of the atrium, which is at least partially divided for the first time, the aerated blood is pumped into the ventricle and there mixes with the nonaerated blood from the body that was returned to the heart via the right half of the atrium. Then the cycle begins again. One of the features of the amphibian system is that the blood leaving the heart for the body is only partially aerated; part of it is the deoxygenated blood returned from the body.

All groups of modern reptiles have a completely divided atrium; it is safe to assume, therefore, that this was true of most, if not all, extinct reptiles. In reptiles, the ventricle

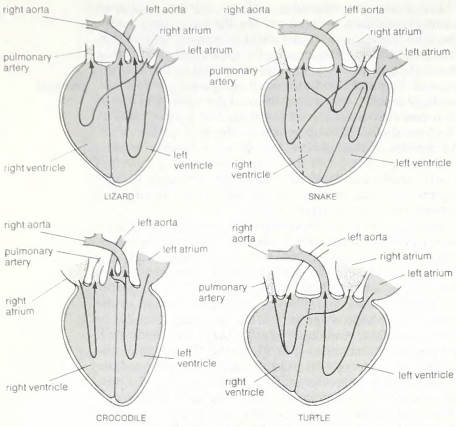


Figure 3: Types of reptilian hearts.

for the first time becomes at least partially divided in the four major living groups.

When the two atria of a lizard's heart contract, the two streams of blood (aerated blood from the lungs in the left atrium and nonaerated blood from the body in the right atrium) flow into the left chamber of the ventricle. As pressure builds up in that chamber, the nonaerated blood is forced through the gap in the partition into the right chamber of the ventricle. Then, when the ventricle contracts, nonaerated blood is pumped into the pulmonary artery and thence to the lungs, while aerated blood is pumped into the systemic arteries (the aortas) and so to the body.

In snakes all three arterial trunks come out of the chamber of the ventricle that receives the nonaerated blood of the right atrium. During ventricular contraction, a muscular ridge forms a partition that guides the nonaerated blood into the pulmonary artery, while the aerated blood received by the other chamber of the ventricle is forced through the opening in the ventricular septum and out through the aortas.

In crocodiles the ventricular septum is complete, but the two aortas come out of different ventricular chambers. A semilunar valve at the entrance to the left aorta prevents nonaerated blood in the right ventricle from flowing into the aorta. Instead, part of the aerated blood from the left ventricular chamber pumped into the right aorta flows into the left by way of an opening.

The ventricle of the turtle is not perfectly divided, and some slight mixing of aerated and nonaerated blood takes place.

Oxygen capacity of reptile blood

Despite the peculiar and complex circulation, a double system has been achieved by lizards, snakes, and crocodilians. Tests of the blood in the various chambers and arteries have shown that the oxygen content in both systemic aortas is as high as that of the blood just received by the left atrium from the lungs and is much higher than that of the blood in the pulmonary artery. Except for the turtles, limitation of activity in reptiles cannot be explained on the basis of imperfect heart circulation.

An explanation may lie in the chemistry of the blood. Apparently, the blood of reptiles has less hemoglobin and thus can carry less oxygen that that of mammals.

Respiratory system. The form of the lungs and the methods of irrigating them may also influence activity by affecting the efficiency of respiration. In snakes the lungs are simple saclike structures having small pockets, or alveoli, in the walls. In the lungs of many lizards and turtles and of all crocodilians the surface area is increased by the development of partitions that, in turn, have alveoli. Because exchange of respiratory gases takes place across surfaces, an increase of the ratio of surface area to volume leads to an increase in respiratory efficiency. In this regard, the lungs of snakes are not so effective as those of crocodilians. The elaboration of the internal surface of lungs in reptiles is simple, however, compared with that reached by mammalian lungs with their enormous number of very fine alveoli.

Most reptiles breathe by changing the volume of the body cavity. By contractions of the muscles moving the ribs, the volume of the body cavity is increased, creating a negative pressure, which is restored to atmospheric level by air rushing into the lungs. By contraction of body muscles, the volume of the body cavity is reduced, forcing air out of the lungs.

This system applies to all modern reptiles except turtles, which, because of the fusion of the ribs with the rigid shell, are unable to breathe by this means; they do use the same mechanical principle of changing pressure in the body cavity, however. Contraction of two flank muscles enlarges the body cavity, causing inspiration. Contraction of two other muscles, coincident with relaxation of the first two, forces the viscera upward against the lungs, causing exhalation.

The rate of respiration, like so many physiological activities of reptiles, is highly variable, depending in part upon the temperature and in part upon the emotional state of the animal.

Digestive and urogenital systems. The digestive system of modern reptiles is similar in general plan to that of all higher vertebrates. It includes the mouth and its salivary glands, esophagus, stomach, and intestine, ending in a cloaca. Of the few specializations of the reptilian digestive system, the evolution of one pair of salivary glands into poison glands in the venomous snakes is the most remarkable.

During development the embryos of higher vertebrates (reptiles, birds, and mammals) use three separate sets of kidneys consecutively; these are arranged in longitudinal sequence in the body cavity. The first set, the pronephroi, are vestigial organs left over from the evolutionary past that soon degenerate and disappear without having had any function. The second set, the mesonephroi, are the functional kidneys of adult amphibians, but their only contribution to the lives of reptiles is in providing the duct (the wolffian duct) that forms a connection between the testis and the cloaca. The operational kidneys of reptiles, birds, and mammals are the last set, the metanephroi, which have separate ducts to the cloaca. The principal function of the kidney is the removal of nitrogenous wastes resulting from the oxidation of proteins. Vertebrates eliminate three kinds of nitrogenous wastes: ammonia, urea, and uric acid. Ammonia and urea are highly soluble in water; uric acid is not. Ammonia is highly poisonous, urea slightly so, and uric acid not at all. Among reptiles the form taken by the nitrogenous wastes is closely related to the habits and habitat of the animal. Aquatic reptiles tend to excrete a large proportion of these wastes as ammonia in solution. This method, involving a great loss of body water, is no problem for an alligator, which eliminates between 40 and 75 percent of its nitrogenous wastes as ammonia. Terrestrial reptiles, such as most snakes and lizards, which must conserve body water, convert their nitrogenous wastes to insoluble, harmless uric acid, which forms a more or less solid mass in the cloaca. In snakes and lizards these wastes are eliminated from the cloaca together with wastes from the digestive system.

Discharge routes for eggs and sperm

Prior to the evolution of the metanephric kidney, the products of the male gonad, the testis, travelled through the same duct with the nitrogenous wastes from the kidney. But with the appearance of the metanephros, the two systems became separated. The female reproductive system never shared a common tube with the kidney. Oviducts in all female vertebrates arise as separate tubes with openings usually near, but not connected to, the ovaries. The oviducts, like the wolffian ducts of the testes, open to the cloaca. Both ovaries and testes lie in the body cavity near the kidneys.

With the evolution of the reptilian egg, internal fertiliza-

tion became necessary. The males of all modern reptiles, with the exception of the tuatara, have copulatory organs. The structures vary from group to group, but all include erectile tissue as an important element of the operating mechanism, and all are protruded through the male's cloaca into that of the female during copulation. Unlike the penis of turtles and crocodilians, the copulatory organ of lizards and snakes is paired, each unit being called a hemipenis. The hemipenes of lizards and snakes are elongated tubular structures lying in the tail. The penis of a crocodile or turtle is protruded through the cloacal opening wholly by means of a filling of blood space (sinuses) in the penis; protrusion of a lizard's or snake's hemipenis, however, is begun by a pair of propulsor muscles. Completion of the erection is brought about by blood filling the sinuses in the erectile tissue. Only one hemipenis is inserted into a female, but which one is a matter of chance. Unlike the penis of mammals, the copulatory organs of reptiles do not transport sperm through a tube. The ducts from the testes, as already mentioned, empty into the cloaca, and the sperm flow along a groove on the surface of the penis or hemipenis.

Sense organs. *Sight.* In general construction the eyes of reptiles are like those of other vertebrates. Accommodation for near vision in all living reptiles except snakes is accomplished by pressure being exerted on the lens by the surrounding muscular ring (ciliary body), which thus makes the lens more spherical. In snakes the same end is achieved by the lens being brought forward under pressure built up on the vitreous humour by contractions of muscles at the base of the iris. The pupil shape varies remarkably among living reptiles, from the round opening characteristic of all turtles and many diurnal lizards and snakes to the vertical slit of crocodilians and nocturnal snakes and the horizontal slits of a few tree snakes. Undoubtedly the most bizarre pupil shape is that of some geckos, in which the pupil contracts to form a series of pinholes, one above the other. The lower eyelid has the greater range of movement in most reptiles. In crocodilians the upper lid is more mobile. Snakes have no movable eyelids, their eyes being covered by a fixed transparent scale. The tuatara and all crocodilians have a third eyelid, the nictitating membrane, a transparent sheet that moves sideways across the eye from the inner corner, cleansing and moistening the cornea without shutting out the light.

Visual acuity

Visual acuity varies greatly among living reptiles, being poorest in the burrowing lizards and snakes (which often have very small eyes) and greatest in active diurnal species (which usually have large eyes). Judging by the size of the skull opening in which the eye is situated, similar variation existed among the extinct reptiles. Those that hunted active prey (*e.g.*, the ichthyosaurs) had large eyes and presumably excellent vision; many herbivorous types (*e.g.*, the horned dinosaur *Triceratops*) had relatively small eyes and weak vision. Colour vision has been demonstrated in few living reptiles.

Hearing. The power of hearing is variously developed among living reptiles. Crocodilians and most lizards hear reasonably well. Snakes and turtles are sensitive to low-frequency vibrations, thus they "hear" mostly earth-borne, rather than aerial, sound waves. The auditory apparatus in reptiles typically consists of a tympanum, a thin membrane located at the rear of the head; a small bone, the stapes, running between the tympanum and the skull in the tympanic cavity (the middle ear); the inner ear; and a eustachian tube connecting the middle ear with the mouth cavity. In reptiles that can hear, the tympanum vibrates in response to sound waves and transmits the vibrations to the stapes. The inner end of the stapes abuts against a small opening (the foramen ovale) to the cavity in the skull containing the inner ear. The inner ear consists of a series of hollow interconnected parts: the semicircular canals; the ovoidal or spheroidal chambers called the utriculus and sacculus; and the lagena, a small outgrowth of the sacculus. The tubes of the inner ear, suspended in a fluid called perilymph, contain another fluid, the endolymph. When the stapes is set in motion by the tympanum, it develops vibrations in the fluid of the inner ear; these vibrations activate cells in the lagena, the seat of the sense of hearing. The semicircular canals are concerned with equilibrium.

Most lizards can hear; details of the acuity of hearing, however, are largely unknown. The majority have a tympanum, tympanic cavity, and eustachian tube. The tympanum, usually exposed at the surface of the head or at the end of a short open tube, may be covered by scales or may be absent. In general the last two conditions are characteristic of lizards that lead a more or less completely subterranean life and presumably do not hear airborne sounds. The middle ear of these burrowers is usually degenerate as well, often lacking the tympanic cavity and eustachian tube.

Snakes have neither tympanum nor eustachian tube, and the stapes is attached to the quadrate bone on which the lower jaw swings. Snakes are obviously more sensitive to vibrations in the ground than to airborne sounds. A loud sound above a snake does not elicit any response provided the object making the sound does not move or, if it does, the movements are not seen by the snake. On the other hand, the same snake will raise its head slightly and flick its tongue in and out rapidly if the ground behind it is tapped or scratched. Snakes undoubtedly "hear" these vibrations by means of bone conduction. Sound waves travel more rapidly and strongly in solids than in the air and are probably transmitted first to the inner ear of snakes through the lower jaw, which is normally touching the ground, thence to the quadrate bone, and finally to the stapes. Burrowing lizards presumably hear ground vibrations in the same way.

Crocodilians, all of which have an external ear consisting of a short tube closed by a strong valvular flap and ending at the tympanum, have rather keen hearing. The American alligator (*Alligator mississippiensis*) can hear sounds within a range of 50 to 4,000 cycles per second. The hearing of crocodilians is involved not only in detection of prey and enemies but also in their social behaviour, for males roar or bellow either to threaten other males or to attract females.

Although turtles have well-developed middle ears and usually large tympana, their ability to hear airborne sounds is still an open question. Measurements of the impulses of the auditory nerve between the inner ear and the auditory centre of the brain show that the inner ear in several species of turtles is sensitive to airborne sounds in the range of 50 to 2,000 cycles per second, but this does not prove that the animals are aware of the sounds.

Chemoreception. Chemical-sensitive organs, used by many reptiles to find their prey, are located in the nose and in the roof of the mouth. Part of the lining of the nose consists of cells subserving the function of smell and corresponding to similar cells in other vertebrates. The second chemoreceptor is Jacobson's organ, originally an outpocketing of the nasal sac in amphibians and remaining so in the tuatara and crocodilians. It has been lost by turtles. Jacobson's organ is best developed in lizards and snakes, in which its connection with the nasal cavity has been closed and is replaced by an opening into the mouth. The nerve connecting Jacobson's organ to the brain is a branch of the olfactory nerve.

Jacobson's organ

The use of Jacobson's organ is most obvious in snakes. If a strong odour or vibration stimulates a snake, its tongue is flicked in and out rapidly. With each retraction the forked tip touches the opening of Jacobson's organ in the roof of the mouth, transmitting any chemical fragments adhering to the tongue. In effect, Jacobson's organ is a supplement to taste and is a short-range chemical receptor, as contrasted with the long-range testing of the true sense of smell located in the nasal tube.

Some snakes (notably the large vipers) and lizards (especially skinks and burrowing species of other families) rely upon the olfactory tissue and Jacobson's organ to locate food, almost to the exclusion of other senses. Other reptiles, such as certain diurnal lizards and crocodilians, appear not to use scent in searching for prey, though they may use their sense of smell for locating a mate.

Some snakes, notably pit vipers, boas, and pythons, have special heat-sensitive organs on their heads as part of their food-detecting apparatus. Just below and behind the nos-

tril of a pit viper is the pit that gives the group its common name. The lip scales of many pythons and boas have depressions (labial pits) that are analogous to the viper's pit. The labial pits of pythons and boas are lined with skin thinner than that covering the rest of the head and are supplied with dense networks of blood capillaries and nerve fibres. The facial pit of the viper is relatively deeper than the boa's labial pits and consists of two chambers separated by a thin membrane bearing a rich supply of fine blood vessels and nerves. In experiments using warm and cold covered electric light bulbs, pit vipers and pitted boas have been shown to detect temperature differences of less than 0.6° C (1.1° F).

Since many pit vipers, pythons, and boas are nocturnal and feed largely on mammals and birds, the facial sense organs enable them to direct their strikes accurately in the dark, once their warm-blooded prey arrives within range. The approach of the prey to that point is probably detected by the chemical receptors—either the nose, Jacobson's organ, or both.

Thermal relationships. Reptiles are often described as being cold-blooded, which is not always true. Their body temperatures are not always low, but they have no internal mechanism for regulating body temperature and thus approximate closely the temperature of their surroundings. This condition is termed poikilothermy. Mammals and birds maintain their relatively high body temperatures at a fairly constant level by physiological means that are independent of the external environment, a condition called homoiothermy. When the body temperature of a dog or a man falls below the normal range, he begins to shiver, blood vessels in the skin contract, muscular activity generates heat, and the contraction of the superficial blood vessels, by reducing the volume of blood flow at the surface, reduces heat loss by radiation. By contrast, a reptile, when its body temperature falls below the optimum, must move to some portion of the environment having a higher temperature; in less than optimal temperatures, its activity drops, its movements become sluggish, its heartbeat slows, and its rate of breathing drops. In short, it becomes incapable of the normal activities required for growth, reproduction, and survival.

Mammals and birds have some physiological means of cooling their bodies (*e.g.*, panting and sweating, expansion of superficial blood vessels), but a reptile must ordinarily move away from a spot in which the temperature is too high or it will perish very quickly. Some reptiles also pant, but most of their temperature accommodations are behavioral (*e.g.*, orienting to sun or wind, raising body from the ground).

Range of body temperatures

Each group of reptiles has its own characteristic thermal range. One genus of lizards, for example, may require temperatures of 29°–32° C (84°–90° F) for maximum efficiency, and another may require 24°–27° C (75°–81° F). As a result of such physiological differences, lizards of the two groups will be active at different times of the day or occupy slightly different habitats.

In general the normal activity temperatures of reptiles are lower than those of most mammals; however, a few sun-loving (heliothermic) lizards (*e.g.*, the greater earless lizard, *Holbrookia texana*, of southwestern United States) have average activity temperatures above 38° C (100° F), several degrees higher than the average human body temperature. Such high temperatures are exceptional, and the majority of lizards have normal activity temperatures in the 27°–35° C (81°–95° F) range.

Evolution and paleontology

Historical development. Reptiles occupy an evolutionary position between amphibians, on the one hand, and the birds and mammals on the other, the last two classes having evolved from reptilian ancestors. Reptiles first appear in the fossil record of the Carboniferous Period, more than 280,000,000 years ago. By the Triassic, about 50,000,000 years later, they began to dominate the terrestrial life of the world and continued that dominance through the Mesozoic Era (65,000,000–225,000,000 years ago). Reptiles succeeded in adapting to deserts, swamps,

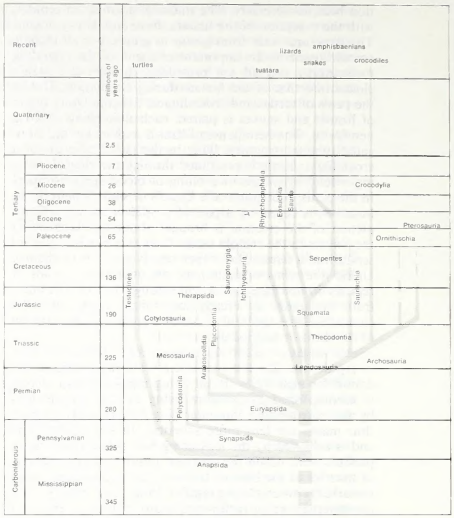


Figure 4: Family tree of the reptiles.

forests, grasslands, rivers, lakes, and even the air and the seas. Coincident with the rise of mammals at the end of the Mesozoic, most reptilian groups became extinct.

The big evolutionary step made by reptiles was the final emancipation from life in water, for, until that step was taken, vertebrates could not exploit all of the Earth's surface. That breakthrough required two basic changes, the first of which took place in the skin. Modern amphibians have naked skins that lack horny scales, hair, and other protective devices. One small amphibian group, the caecilians, has small fishlike scales embedded in the skin; similar scales occurred in certain extinct amphibians. Because of their thinness and position, such amphibian scales are no protection against desiccation, one of the principal hazards of life for all animals, vertebrate and invertebrate. This susceptibility to drying out forces amphibians to remain in water or in very humid places, thus limiting their exploitation of the terrestrial environment. Reptiles evolved a different type of scale consisting of keratin (or horn) deposited in the outermost layer of the skin. This type of scale was in a position and of a thickness to prevent desiccation.

Consequences of amniote egg

The second basic change made by the reptiles was the development of the amniote egg. This development expanded the possibilities of exploiting terrestrial environments. The egg could be laid under rocks or logs, in holes in the ground, in deserts or forests—in fact, almost anywhere except in water.

The development of the reptilian egg had several other consequences. An egg that is enclosed by a shell must be fertilized before that shell is deposited, thus necessitating internal fertilization. The evolution of a land egg also increased the efficiency of the life cycle of terrestrial vertebrates. An amphibian, hatching from an aquatic egg, must develop and grow in water in a larval form, the tadpole. The history of a reptile, on the other hand, is one of development and growth of adult structures adapted for a terrestrial life. It need not develop gills or a lateral line system (series of sense organs), which are needed by the aquatic tadpole and which must be resorbed and reworked into other structures. This important change in the type of development was made possible by a great increase in the amount of yolk in the reptilian egg.

The large amount of yolk also permitted the lengthening

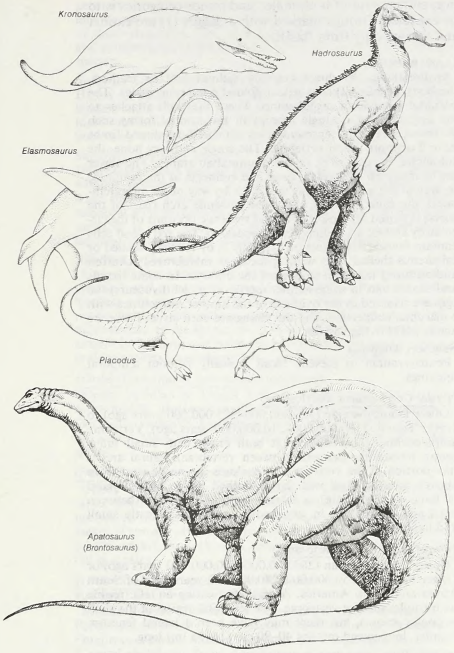


Figure 5: Body plans of extinct reptiles.
Drawing by J. Helmer based on (*Hadrosaurus, Brontosaurus*) photographs courtesy of Field Museum of Natural History, Chicago; (others) Colbert, *Age of Reptiles*, Weidenfeld & Nicolson

of the embryonic period, which in turn allowed the development of all structures that are necessary for successful existence on land. When a reptile hatches, it is ready to carry out all the activities of the adult (with the exception of reproduction), in the same environment and manner.

After reptiles acquired scaly protection for their skin and an egg that did not have to be laid in water, they were free to move over most of the Earth's surface. That freedom set the stage for the evolution of the many varied types of reptiles and, ultimately, for the evolution of birds and mammals.

Fossil distribution. What is known of the fossil record of reptiles shows that most of the major groups, or orders, were worldwide or nearly so at some time in their individual histories. Few orders are known from South America and Australia; the absence of most major groups from these areas is explained more likely on the basis of lack of preservation or lack of discovery of fossil beds than on the basis of a genuine absence of the animals throughout such a long interval as the Mesozoic. In the following discussion the names of present-day continents are used, though it should be understood that continental outlines in the past did not always coincide with those of today.

Few orders of reptiles are known from Paleozoic times—*i.e.*, Carboniferous (280,000,000–345,000,000 years ago) and Permian (225,000,000–280,000,000 years ago). The stem reptiles, the Cotylosauria, have been found in Carboniferous deposits of eastern and western North America and western Europe and in Permian beds of what was formerly the Soviet Union and of Africa. Presumably they also lived in Asia during this interval of more than 100,000,000 years, but their remains have yet to be found there. In the same period the Pelycosauria lived in North America and Europe, where their fossils are well known, and possibly in Africa and Asia. The mammal-like Therapsida and perhaps other forms were fossilized in Africa and Europe. The probable ancestors of turtles appeared in Africa, and the first diapsids (reptiles having two-arched temporal structures) emerged in Africa and Europe in the Permian.

By the Triassic (190,000,000–225,000,000 years ago), the earliest portion of the Mesozoic, the mammal-like reptiles had spread to all of the continents except Australia. Turtles were still in Africa and had spread at least as far as Europe. The Ichthyosauria were living in seas covering what is now western North America and western Europe and may have been much more widely distributed, considering their oceanic habitat. North America at that time was also the home of primitive diapsids (Thecodontia), phytosaurs (suborder Phytosauria), and the earliest dinosaurs (Saurischia) and crocodiles. Besides the mammal-like therapsids, Eurasia and Africa in the Triassic had phytosaurs, the first Rhynchocephalia, and the early dinosaurs. The major groups of reptiles, therefore, were essentially worldwide in distribution by the Triassic.

Appearance of giant dinosaurs

The giant dinosaurs began their efflorescence in the Jurassic (136,000,000–190,000,000 years ago). The big carnivorous types, such as *Allosaurus*, roamed the landscapes of the major continents, presumably preying on even larger herbivorous dinosaurs (*Apatosaurus, Diplodocus*, etc.), whose remains have been found in North America, Europe, Africa, and Australia. Marine ichthyosaurs and



Figure 6: Body plans of extinct reptiles.

plesiosaurs (order Sauropterygia) still swam in the shallow seas of both hemispheres. The ornithischian dinosaurs (Ornithischia), the ancestors of the duckbilled and horned dinosaurs, became widely distributed in the Jurassic. One group, the armoured stegosaurians (suborder Stegosauria), left fossils in North America, Europe, and Africa. The flying reptiles (Pterosauria) also made their appearance at least in Africa and Europe during the Jurassic.

The culmination of dinosaur evolution occurred in the Cretaceous 65,000,000–136,000,000 years ago), when every part of the world had herbivorous ornithischian dinosaurs: the ankylosaurs (suborder Ankylosauria), medium-sized dinosaurs armoured with heavy plates and large spines, ranging from South America to Africa; the duck-billed dinosaurs (suborder Ornithopoda), ranging from North America to Africa; and the horned dinosaurs (suborder Ceratopsia) in North America and Asia. Everywhere they were preyed upon by the big carnivorous types, which culminated in North America in the gigantic carnivore *Tyrannosaurus*. The skies over North America, Africa, and Europe (and probably Asia and South America as well) were the province of the flying reptiles (Pterosauria). There were also inconspicuous groups that later inherited the reptilian world. Lizards (suborder Sauria) appeared in most continents, and snakes (suborder Serpentes) appeared in some places. Birds, having splintered off from reptilian ancestors in the Jurassic, became more numerous in the Cretaceous; mammals, which ultimately replaced most of the reptiles, were represented on most continents by small creatures.

Through all this evolutionary activity, the conservative turtles continued their plodding evolutionary pace, changing but little, yet lasting through all. (Ed.)

Classification

Distinguishing taxonomic features. The major reptile groups are distinguished on the basis of vertebral and skull features, particularly the number and positions of the temporal fenestrae (*i.e.*, large openings in the temporal bone). Beyond these, the pelvic structure and that of the teeth and

From *The Procession of Life* by Alfred S. Romer (1968)
Universe Books, New York, and Weidenfeld & Nicolson Ltd.

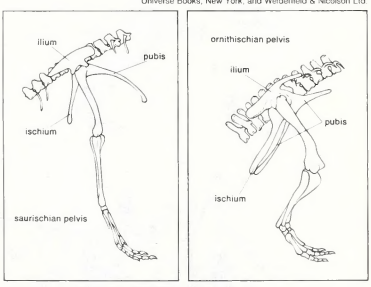


Figure 7: The types of dinosaur pelvis.

limbs are important. Any extraordinary structural development (*e.g.*, the wings of pterosaurs, the shell of turtles) is also a major factor in the system of classification. In recent reptiles (four out of some 17 orders), the structure of the heart, the male secondary sex organs, the extent and kind of dermal armour, and the structure of other soft organs are used. Discoveries in serology (the study of blood serum) and karyology (the study of chromosomes) have little effect, as yet, on the systems of classification.

Annotated classification. The following classification of the reptiles is based on that of A.S. Romer (1956, 1966), the American vertebrate paleontologist, as later modified by himself (1968) and in minor respects by E.H. Colbert (1965), B.W. Halstead (1969), and H.G. Dowling (1971). No single system of reptile classification is acceptable to all herpetologists, and widely differing views are held by various authorities. Considerable change in the recognition and content of both major and minor categories is to be expected. Groups marked with a dagger (†) are extinct and known only from fossils.

CLASS REPTILIA

Air-breathing, vertebrate animals without hair or feathers, the body usually covered with infolded epidermal scales. The occipital condyle (a protuberance where the skull attaches to the 1st vertebra) is single (except in transitional forms such as Therapsida), and representatives with well-developed limbs have 2 or more sacral vertebrae. The single auditory bone, the columella, is equivalent to the mammalian stapes. The lower jaw is made up of several bones and connects to the braincase by way of the quadrate bone and often by way of the supratemporal, the squamosal, or both. The systemic arch (part of the aorta) is paired; the respiratory and systemic portions of the circulatory system are incompletely separated. The red blood cells contain nuclei. Reproduction typically is by leathery-shelled or calcareous-shelled eggs with specialized membranes (chorion and amnion) that help to protect the embryo. In some lizards and snakes and in some extinct reptiles (*e.g.*, ichthyosaurs) the eggs are retained in the oviducts of the mother, sometimes with a placental connection, and the young are born alive. There are about 6,000 living species.

Subclass Anapsida

Pennsylvanian to present. Skull typically without temporal openings.

†*Order Cotylosauria* (cotylosaurs)

Lower Pennsylvanian (300,000,000–325,000,000 years ago) to Upper Triassic (190,000,000–210,000,000 years ago). Vertebrae amphicoelous (*i.e.*, concave at both ends) with small intercentra (crescentic elements between vertebrae); neural arches (the portion of the vertebra that encloses the nerve cord from above) convex. Skull with large toothed flange on pterygoid (a bone of the palate); a lateral temporal fenestra (between jugal and squamosal) in some advanced forms. Mostly small and lizard-like.

†*Order Mesosauria* (mesosaurs)

Upper Pennsylvanian (280,000,000–300,000,000 years ago) or Lower Permian (250,000,000–280,000,000 years ago) of South Africa and South America. A small fish-eating aquatic reptile with cotylosaur-like vertebrae. The temporal region of the skull is poorly known, but there may have been a lateral fenestra. Slender, long-jawed reptiles 40–90 cm (16–35 in.) long.

Order Chelonia (or Testudines; turtles)

Upper (possibly Middle) Triassic to present. Skull without pineal (on the midline of the "forehead") or temporal fenestrae, though the temporal region may be emarginated, or indented. Jaws toothless (though there were palatine teeth in some extinct forms), with a horny beak. A shell (dorsal carapace and ventral plastron) covers the body and encloses the pectoral and pelvic girdles. Most modern turtles have shells less than 60 cm (24 in.) long, but oceanic forms have larger ones (2.7 m; 8.9 ft), and those of some extinct turtles exceeded 3.6 m (11.8 ft) in length.

Subclass Lepidosauria (lepidosaurians)

Upper Permian (225,000,000–250,000,000 years ago) to present. Primitive forms had 2 openings in the temporal region of the skull; most of the descendants have lost the lower (jugal–quadratojugal) temporal arch. The earliest known forms already had a jaw that was shortened. Usually a pineal eye (degenerate, median eyelike structure), no trend toward bipedalism. Ribs single-headed in advanced forms.

†*Order Eosuchia* (eosuchians)

Upper Permian to Eocene (38,000,000–54,000,000 years ago). Primitive lizard-like reptiles with 2 temporal arches and without a beaked snout.

Order Rhynchocephalia (beaked reptiles)

Lower Triassic to present. Scaled reptiles with 2 temporal arches. The premaxillary overhangs the lower jaw as a beak; the teeth are acrodont (*i.e.*, attached to the edge of the jaw rather than inserted in sockets). The vertebrae are amphicoelous, and the ribs are single-headed. Mostly lizard-like forms 30–90 cm (12–35 in.) long; one group (rhynchosaurs) attained lengths of up to 180 cm (71 in.). One living species.

Order Squamata (scaly reptiles)

Lizards, snakes, and amphisbaenians. Upper Triassic to present. The quadrate is freed by loss of the lower (jugalquadratojugal) arch and reduction of the squamosal to allow some movement at both ends. Vertebrae procoelous (*i.e.*, with the centre part concave on the anterior side, convex on the posterior side) except in a few geckos (Sauria), in which they are amphicoelous. Living species possess unique paired copulatory structure (hemipenes).

Suborder Sauria (lizards). Upper Triassic to present. Most

generalized suborder; most species with well-developed limbs, an external ear opening, movable eyelids, or some combination of these structures. The skull typically has a pineal opening and epipterygoid, lacrimal, and jugal bones. About 3,000 living species.

Suborder Amphisbaenia (amphisbaenians). Eocene to present. Highly specialized, limbless, burrowing reptiles with the eyes hidden under the skin, no pineal opening, and the body scales fused into annuli, or rings. The skull is solidly constructed as a burrowing wedge. About 140 living species.

Suborder Serpentes (snakes). Upper Cretaceous (65,000,000–100,000,000 years ago) to present. A highly specialized group, without pectoral limbs or girdle, pelvic limbs rudimentary when present. Without external ear or movable eyelids. Upper temporal arch (postorbital–squamosal) absent, leaving quadrate movable at both ends. No pineal opening; no epipterygoid, lacrimal, or jugal bones. About 2,500 living species.

Subclass Archosauria (ruling reptiles)

Upper Permian to present. Reptiles with 2 temporal openings (diapsid) and a tendency toward bipedalism. Most have long hindlegs and short forelimbs. Typically without a pineal opening in the skull, but with an antorbital fenestra and one on the outer surface of the lower jaw. Ribs typically 2-headed, at least anteriorly. Ischium and pubis (bones of pelvis) elongated. Teeth in deep sockets (thecodont). Most with some armour.

†*Order Thecodontia* (primitive archosaurs)

Upper Permian to Upper Triassic. Lightly built bipedal or more heavily armoured reptiles, some (phytosaurs) crocodile-like and presumably amphibious, but with nostrils far back on the snout. Most had at least 2 rows of bony plates along the spine.

Order Crocodilia (crocodilians)

Upper Triassic to present. Aquatic or amphibious reptiles, rather generalized in body form, but with a flattened skull, the nostrils on the tip of the snout, and a well-developed secondary palate. Typically, the pubis is excluded from the acetabulum, or hip socket, and the 5th toe is reduced to a stump; 21 living species.

†*Order Saurischia* (carnivorous dinosaurs and giant herbivorous dinosaurs)

Middle Triassic (200,000,000 years ago) to Upper Cretaceous. Pelvis triradiate (*i.e.,* 3-branched). Some reduction in digits. Forelimbs usually distinctly shorter than hindlegs. Three to 7 sacral vertebrae. Some herbivorous forms were more than 24 m (78 ft) long.

†*Order Ornithischia* (herbivorous dinosaurs)

Upper Triassic to Upper Cretaceous. Pelvis tetraradiate (*i.e.,* 4-branched). Typically with a beaklike structure in the front part of the mouth and grinding teeth in the rear. Both obligate bipedal and quadrupedal forms. Toes often with hooflike structures. Many with heavy armour and horns. Largest about 9 m (30 ft) long.

†*Order Pterosauria* (pterodactyls)

Lower Jurassic (150,000,000–190,000,000 years ago) to Upper Cretaceous. Highly specialized flying reptiles with hollow bones; 4th digit of the forelimb greatly elongated to support the flying membrane of the wing. Early forms toothed and with long tails; later forms tended to be larger and to have lost both teeth and tail.

†**Subclass Euryapsida** (plesiosaurs and relatives)

Lower Permian to Upper Cretaceous. Mainly aquatic reptiles with an upper temporal opening (between postorbital, squamosal, and parietal bones) and a broad plate of bone below.

†*Order Araeoscelidia* (primitive euryapsids)

Lower Permian to Upper Triassic. Primitive terrestrial reptiles with lizard-like proportions, some with a specialized cervical region.

†*Order Sauropterygia* (nothosaurs and plesiosaurs)

Middle Triassic to Upper Cretaceous. Aquatic reptiles with strongly developed ventral ribs, dorsally placed nostrils, and a highly modified palate (pterygoids and often palatines are joined in the midline). Limbs paddle-like in advanced forms.

†*Order Placodontia* (placodonts)

Lower to Upper Triassic. A side branch of euryapsids, apparently mollusk eaters. In some the body was armoured and turtle-like in form.

†*Order Ichthyosauria* (ichthyosaurs)

Middle Triassic to Upper Cretaceous. Highly aquatic reptiles with porpoise-like bodies, a dorsal fin, and a reversed-heterocercal tail (*i.e.,* with the lower lobe longer than the upper). Limbs paddle-like; snout often elongated and beaklike.

†**Subclass Synapsida** (mammal-like reptiles)

Lower Pennsylvanian to Middle Jurassic (160,000,000 years ago). A single lateral temporal opening with the postorbital and squamosal bones joining above it in primitive forms; the opening extends upward to the parietal in later forms. Pineal present. Teeth differentiated. Two coracoids in pectoral girdle.

†*Order Pelycosauria* (primitive synapsids, pelycosaurs)

Lower Pennsylvanian to Middle Permian (250,000,000 years ago), especially in Europe and North America. Neural arches higher and less swollen than in the contemporary cotylosaurs. Abdominal ribs present in most.

†*Order Therapsida* (advanced synapsids, therapsids)

Middle Permian to Middle Jurassic, mainly in South Africa. Temporal opening expanded in advanced forms, a secondary palate formed. Occipital condyle double and dentary bone much enlarged.

Critical appraisal. A natural classification of reptiles is more difficult than that of many animals because the main evolution of the group was during Mesozoic time; 13 of 17 recognized orders are extinct. The consequent reliance on osteological (bone) characters may have obscured some important evolutionary trends, and there is little agreement among herpetologists and paleontologists on reptile taxonomy. Even the major categories of reptile classification are still in dispute. Although the ideas of Watson, Colbert, and Romer have dominated the field, some authorities question most of the basic elements—from subclass to suborder—on which the framework of their classification depends. Halstead, for example, would discard the entire system of reptile subclasses and recognize a "superclass Reptilia" with seven reshuffled "classes."

On the other hand, there is general agreement that the base reptilian stock is the Cotylosauria, which evolved from an amphibian labyrinthodont stock (the captorhinomorphs) at about the Mississippian–Pennsylvanian transition. It is also quite clear that the cotylosaurs early divided into two lines, one of which (the pelycosaurs) represented the stock that gave rise to the mammals. Another branch led to all of the other reptiles, and, later, to the birds as well. Thus, most of the questions of reptilian evolution and classification deal with inter-reptilian relations, rather than with their relationships with other animals. (H.G.Do.)

MAJOR REPTILIAN GROUPS

The remainder of this article consists of a review of the major living orders and suborders of the class Reptilia, arranged in accordance with the *Annotated classification* above. Extinct groups (designated in the *Annotated classification* by a dagger [†]) are treated in the following sections under the headings *Evolution* or *Paleontology*; additional information about the origin and evolution of the reptiles is contained in the *Macropædia* article GEOCHRONOLOGY: *Fossil record.*

Chelonia (turtles)

Turtles are members of the Chelonia, an ancient order of reptiles chiefly characterized by a shell that encloses the vital organs of the body and more or less protects the head and limbs. Although there has been much confusion over the scientific as well as the common name of the group, most scientists now accept the term Chelonia rather than Testudines or Testudinata. Two common names are in wide use: "tortoise" and "turtle." "Tortoise" is applied in the British Isles to all members of the group except the few marine species, all of which have paddle-shaped limbs. "Turtle" has long been much more broadly applied in the United States, with the addition of "terrapin" for some edible species. Usage both in the British Isles and in the United States has left the group without a general name comparable to "bird" or "mammal." The American Society of Ichthyologists and Herpetologists standardized

the common names of the reptiles found in the United States, assigning "turtle" to all of those with a shell. The name "tortoise" is employed secondarily for the slow-moving terrestrial species, primarily those of the genera *Testudo* and *Gopherus*.

NATURAL HISTORY

Food habits

Most turtles prefer a varied diet. Fibrous parts of plants are avoided because the jaws are not sharp enough to cut well and are entirely incapable of grinding. Small invertebrates, such as worms, snails, slugs, insects, thin-shelled bivalves and crayfishes and other crustaceans make up the bulk of the animal food in the turtle diet. Large aquatic turtles are able to catch fish and occasionally a few birds and small mammals.

Turtles have been toothless for more than 150,000,000 years, but in some modern types the moderately sharp and jagged edges of the horny jaws function as teeth. Food is chewed, the claws of the forelimbs often assisting in manipulation, until it is reduced to fragments that can be swallowed.

A few turtles have special ways of securing prey. The gigantic alligator snapping turtle (*Macrochelys temminckii*) of the southern United States has a wormlike lure on the floor of the mouth with which it entices prey into its open jaws. The grotesque matamata, or fringed turtle (*Chelys fimbriata*), of South America has, on the neck and chin, soft projections with which, apparently, it detects the presence of prey by water movements. The floor of the large throat is quickly lowered as the head thrusts forward, the mouth agape. Water rushing into the mouth takes the unwary prey with it.

Turtles, like other reptiles, can survive long fasts, being able to live on weekly or even monthly feedings; however,

Drawing by M. Moran based on (all except *Sternotherus* and *Malaclemys*) J. Z. Young, *The Life of Vertebrates*, 2nd ed. (1962), The Clarendon Press, Oxford

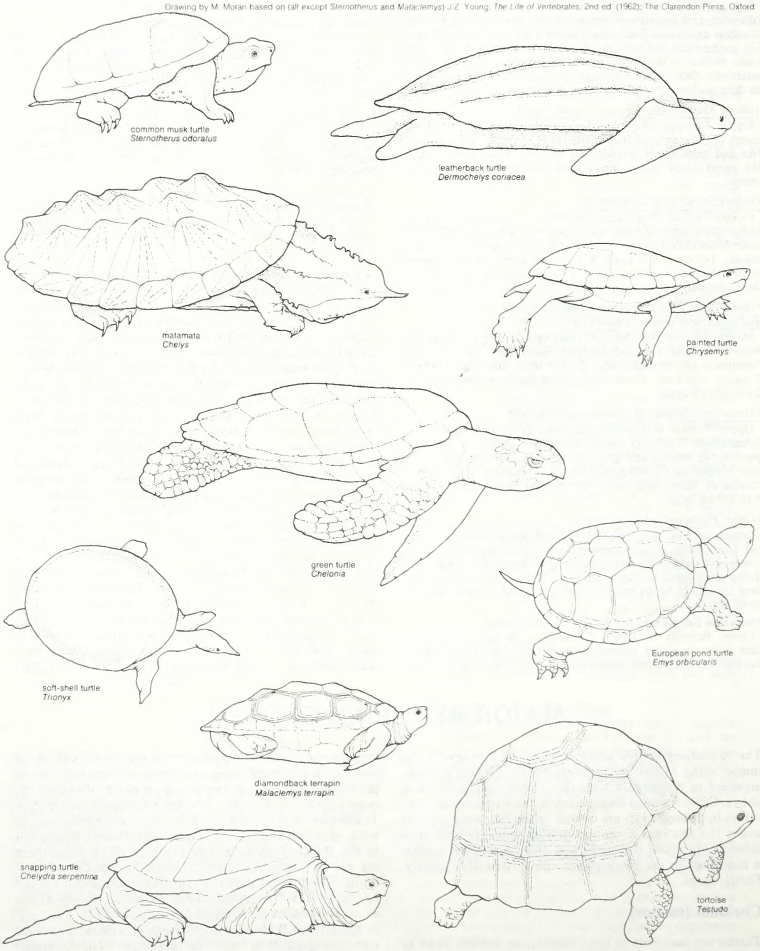


Figure 8: Representative chelonians.

when food is readily available, they may eat frequently and grow very fat. The rate of digestion varies with the temperature. In the wild, they probably maintain a relatively constant temperature by seeking suitable surroundings, which may require constant activity. Turtles drink readily, and some store water in cloacal bladders, an ability that allows them to survive long droughts.

Copulation is usually preceded by a courtship that is highly characteristic for a given species or related group of species. The male's part in the mating display may include various types of head-waving, antics such as lunging at the female while roaring, or, in some aquatic species, gracefully swimming backward in front of her while stroking her lores (cheeks) with the excessively long nails of his forelimbs. The penis, paired in snakes and lizards, is single in turtles.

The most unusual aspect of turtle reproduction is the ability of females to lay fertile eggs for years after a single mating; among higher animals the rule that copulation must precede each pregnancy has few exceptions. In the diamondback terrapin (*Malaclemys*), fertile eggs may be laid for as many as four years after copulation. In a controlled experiment, ten old females were penned without the males they had been living with. During the first season after penning, they laid 124 eggs, all but one of which hatched; during the fourth season they deposited 108 eggs, only four of which hatched, however.

Egg-laying behaviour

Both terrestrial and aquatic turtles lay their eggs on land. The female usually selects a sunny place for the nest. She then digs a hole about as deep as her hindlimbs are long, into which she deposits round or elliptical, whitish eggs, the shells of which may be either flexible or brittle, but never pigmented like bird eggs. Several clutches may be laid in a season, though this is by no means the rule. A clutch may have more than 200 eggs (as in some sea turtles) or as few as one. The sea turtles and the largest tortoises lay eggs about five centimetres (two inches) in diameter, whereas nearly all of the other turtles lay much smaller ones.

Many nests are carefully constructed and hidden; a few are crudely made. The most spectacular laying is that of the sea turtles. The female green turtle (*Chelonia mydas*), for instance, crawls up the beach to a point above high tide and excavates a shallow depression, using all four flippers, then digs an egg pit in the bottom of the depression. The sand is removed by hind flippers, used alternately. During this delicate procedure the flipper must be curled and gently lowered to get its load, a feat often accomplished without loss of sand. A final flip sends the sand directly backward and clear of the hole. She then deposits the eggs, usually two at a time, and carefully covers them. Before leaving she obliterates the exact site by flinging sand about with the front flippers. The entire process consumes a few hours.

The period of incubation depends to an appreciable extent on the temperature. After the nest is completed, the mother appears to take no further interest in it, or, for that matter, in the hatchlings. As a result, the nest is often robbed of eggs, and freshly hatched turtles, whose shells are still soft, are frequently preyed on by large birds and small mammals. Again, in the case of the sea turtles, favoured beaches may become the centre of aggregation of countless animals, devouring the hatchlings as they scurry down the beach. Even in the water they are not safe, since predacious fishes lurk there.

Longevity

These shelled reptiles outlive all other vertebrates, including man, the longest-lived mammal. A marked eastern box turtle (*Terrapene carolina*) of the U.S. survived 138 years in the wild. There is good evidence of a turtle of another species having lived more than a century and a half. Turtles do not grow very slowly; maturity is reached in less than ten years, and growth in a large species may be more rapid than in man himself.

Age estimates are sometimes made by counting the growth rings on the horny shields of the shells, but this method is of little practical value, because the rings become indistinct with maturity. Size is sometimes untrustworthy as an indication of age.

Although scientific evidence is lacking, turtles appear to depend heavily on their sense of smell. Their sight is good. They easily distinguish differences in intensity of light and it is known that they recognize at least four colours: blue, green, yellow, and red, especially the last, which often figures in the adornment of the turtles themselves. In experimental situations they learn to choose among these colours and among various complex black-and-white patterns to receive reward or avoid punishment. The receptors of the eye are all cones, which are responsible for the ability of colour recognition. Many semi-aquatic species can make alterations in curvature of the lens for vision under water, but only the eyes of sea turtles are fully adapted for such vision.

Senses

Turtles respond readily to vibrations of the substratum, but those of the air are another matter. Casual observations do not suffice to evaluate turtle sensitivity and response to aerial sounds, and even scientific workers have disagreed. Experiments reported in 1915 by an American indicated that turtles do hear, but this was contradicted in 1925 by a Japanese. A few years later a Soviet investigator concluded that turtles do hear, but his paper was published only in the U.S.S.R. and largely ignored elsewhere. It was not until 1966 that two Americans demonstrated that at least some turtles hear low-frequency airborne waves, ranging from about 20 to about 1,000 cycles per second. The corresponding figures for man are 20 to 20,000, a fact that may help to explain the difficulty of arriving at a suitable conclusion for the turtle.

Turtles seldom emit sounds, except when courting or mating; even then they do little more than grunt or roar. Exceptions are a few sea turtles that can give a loud cry under extreme provocation.

Experiments on the intelligence of turtles indicate that they learn readily; in some ways they are comparable to the laboratory rat. Persons with extensive experience with pets sometimes elicit from turtles a degree of intelligence usually credited only to mammals. To the casual observer, however, a turtle often appears to be very stupid, especially when it repeatedly climbs over a large object it could easily go around.

Locomotion

The turtle is proverbially one of the slowest animals, and there is some justification for this reputation, at least for certain land forms. It is odd that, in general, aquatic species move faster, even on land. The tortoises of the genus *Gopherus* have been clocked at rates of 0.21 to 0.48 kilometres (0.13 to 0.30 miles) per hour, whereas the rate on land of a normally aquatic cooter (*Pseudemys floridana*) has been recorded at 1.7 kph (1.07 mph) in spite of, or possibly because of, the fact that it was out of its element.

The marine green turtle (*Chelonia mydas*) has been known to swim 480 kilometres (300 miles) in 10 days; it must have travelled at an appreciable rate, since it scarcely could have swum steadily ahead without taking time to eat, sleep, or rest. Soft-shell turtles (Trionychidae) are able to move their limbs at a rate comparable to that in birds and mammals.

The sense of location is well developed; turtles released in an enclosure will usually pick out a resting place and spend much time in it. Many species have a "home area" to which they will return if they are taken a short distance away. The sea turtles (Cheloniidae) are exceptions to the general rule of living in a restricted area; some make long migratory journeys and mass migrations from breeding beach to feeding ground and back.

When a sea turtle hatches out high on a beach, it is faced with something of a problem; from its low eye-level the ocean probably is not visible. Considerable attention has been paid to its problem and the conclusion drawn that scanning of the horizon and landscape enables the turtle to pick out the proper direction, by moving away from the darker areas or by going toward the lighter ones. This method works both day and night.

ECONOMIC VALUE

Man has always relished turtles, and it is likely that almost every species has at one time or another satisfied the broad human appetite. The green turtle (*Chelonia mydas*), with its distribution extending around the world, no doubt has

supplemented the diet of peoples of more different cultures than has any other wild vertebrate. Tortoise populations of many oceanic islands have been decimated; it has been estimated that 10,000,000 land giants were taken from the Galápagos Islands as food supply for the early whaling ships. Turtle eggs are also prized as food. These are deposited in such abundance on certain beaches that harvesting them has become a national industry in Malaysia.

Just as turtle meat has long satisfied man's appetite, so has "tortoise shell" gratified his sense of beauty. Plastics have come to the rescue of the hawksbill turtle (*Eretmochelys*), chief source of the horny shields from which tortoise-shell ornaments are made.

FORM AND FUNCTION

Structure of the shell

The protective shell, to which the evolutionary success of turtles is largely attributed, is a casing of bone covered by horny shields. Plates of bone are fused with ribs, vertebrae, and elements of shoulder and hip girdles. There are many shell variations and modifications from family to family, some of them extreme. At its highest development, the shell is not only surprisingly strong but also completely protective. A box turtle (*Terrapene*) of North America, for example, can readily support a weight 200 times greater than its own; a man with a proportionate supporting power could bear up two large African elephants. The lower shell (plastron) can be closed so snugly against the upper (carapace) that a thin knife blade cannot be inserted between them. This, of course, means that the plastron of the box turtles has a hinge allowing it to move upward to fit into the carapace. Such a movable joint occurs here and there among turtles and is not as simple as it might at first seem. Since the horny shields of the surface do not ordinarily coincide with the bony plates below, the presence of a hinge calls for adjustment to bring about coincidence of the borders of certain shields and plates. Although the hinge is usually in the plastron, the carapace of the African genus *Kinixys* has the hinge, permitting it to move up and down to a limited extent.

The protection of the turtle shell was acquired by the development of structural peculiarities, two of which are especially worthy of mention. As a rule, the limb girdles of vertebrates lie outside the rib box; in turtles, however, they are partly within it, due to a fusion of the developing ribs with the growing shell, which carries the ribs to a position partially surrounding the limbs.

In man the ribs play an important part in the chest expansion, enabling him to breathe. The turtle's ribs are immovable, so the task of chest expansion has been transferred to abdominal muscles; two muscles enlarge the chest cavity for inspiration, and the others press the organs against the lungs to force the air out. Some aquatic species have additional methods of breathing: the vascularized mucous membrane of the cloacal region or of the throat can function like the gill of a fish. Such accessories to ordinary respiration enable turtles to lie quietly submerged for hours or even days.

The vertebral column, with very little to support, underwent drastic modification. The trend in turtles, in contrast to that in most other reptiles, has been to reduce the number of vertebrae. The ability to retract the head into the shell is related to the retention of eight specialized vertebrae. The result is that the neck has almost as many vertebrae as does the central part of the column (between neck and tail). The manner in which the neck bends is of importance in classification (see below).

Turtles run the gamut in size. The Atlantic leatherback (*Dermochelys coriacea*), largest of living kinds, may weigh more than 680 kilograms (1,500 pounds) and measure 3.7 metres (12 feet) from the tip of one front flipper to that of the other. An extinct marine giant, *Archelon*, of the family Protostegidae, was probably much larger, and an extinct land turtle of Asia (*Colossochelys atlas*) had a shell 2.1 metres (seven feet) long. The largest living tortoises may weigh more than 225 kilograms (500 pounds).

At the opposite extreme, the adults of some species weigh less than a pound and have a shell less than five inches long. The shell length of most adult turtles falls between five and 15 inches.

ORIGIN AND EVOLUTION

Origin of turtles

The evolution of the turtle is one of the most remarkable in the history of vertebrates. Unfortunately, the origin of this highly successful order is obscured by the lack of early fossils, although turtles leave more and better fossil remains than do other vertebrates. By the middle of the Triassic Period (about 200,000,000 years ago) turtles were numerous and in possession of basic turtle characteristics. Teeth are lacking in all living turtles, but were present in a few fossils. The teeth of even the Triassic turtles were oddly placed, apparently having been confined to the palate. Intermediates between turtles and cotylosaurs, the primitive reptiles from which turtles probably sprang, are entirely lacking. The most likely link is a small, toothed reptile (*Eunotosaurus*) of the Permian Period of southern Africa. The skull roof of *Eunotosaurus* is not known, but the shoulder and pelvic girdles were overlapped by the ribs, possibly a beginning of the condition found in turtles.

Many turtles today live in marshes and swamps and it is likely that this way of life has been a common one throughout their history. A fairly large group has become terrestrial, but a stronger tendency has been in the opposite direction, toward aquatic life. The extreme development of aquatic types has been large marine species with limbs

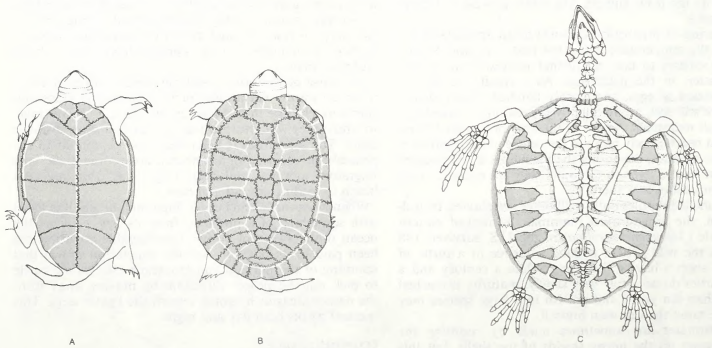
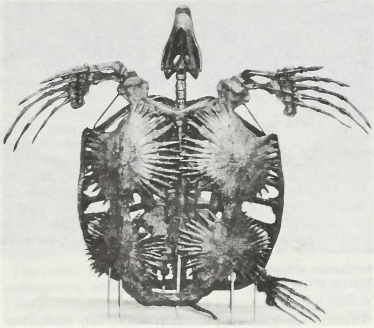


Figure 9: *Turtle skeleton.*
(A) Plastron and (B) carapace, showing the relationship between bony and horny parts of shell in a freshwater turtle. Shaded parts show parts of horny shell; black lines indicate joints in underlying bone. (C) Relationship between the dermal bones (plastron and carapace) and the axial skeleton in a marine turtle.



Skeleton of the Cretaceous marine turtle *Archelon*, height 325 cm (10 ft 8 in.).
By courtesy of the Peabody Museum of Natural History, Yale University

turned into paddles. The shell may become modified and reduced as seen in the gigantic leatherback (*Dermochelys*), in which the horny shields are gone and the shell has become leathery.

While some reptiles flourished and vanished (as the dinosaurs), others persisted, some as once-successful groups and a few as initiators of expanding groups (snakes and lizards). The turtles, however, have plodded a stolid and steady course through evolutionary time, changing very little in basic structure.

CLASSIFICATION

Distinguishing taxonomic features. The suborders and superfamilies of the Chelonia are defined primarily on the basis of the completeness of the skull and degree of posterior emargination. The most obvious diagnostic feature separating the two modern suborders is the manner of folding the neck (vertical or horizontal) when the head is withdrawn. Family distinctions are based primarily on characters of the skull and shell.

Annotated classification. The classification presented here is adapted from A.S. Romer, 1956 and 1966. Three unimportant fossil families have been omitted from the first two suborders.

ORDER CHELONIA

Reptiles with temporal region of skull complete or emarginate, but without true fenestrae. 18 presacral vertebrae, typically 8 cervicals and 10 dorsals; 2nd to 8th dorsals fused to neural elements of carapace (when present). Shoulder girdle internal to ribs and shell. Clavicles and interclavicle incorporated in shell.

†Suborder proganochelydia

Fossil only.

†Superfamily Proganochelyoidea

†Family Proganochelyidae. Skull solidly roofed, sculptured. 7 cervicals. Teeth present. Oldest and most primitive of true turtles. Found only in the Late Triassic of Europe. Presumably amphibious.

†Suborder Amphichelydia

Fossil only. Skull roof complete, rarely emarginate from behind. Little or no retraction of neck. Teeth absent. Suborder includes the more primitive turtles, which were almost entirely Mesozoic and lacked the distinctive features of the cryptodires and pleurodires.

†Superfamily Pleurosternoidea

Skull generally elongate, sculptured, or tuberculate; roof at times moderately emarginate. 8 cervicals. Chiefly amphibious but some marine.

†Family Pleurosternidae. Temporal region well roofed. Common aquatic chelonians of the Jurassic and Lower Cretaceous. Europe and North America. About 10 genera.

†Family Plesiochelyidae. Temporal roof completed by enlarged postorbital. Jurassic and Cretaceous. Europe and Asia. Ancestral to the following family. 2 genera.

†Family Thalassemyidae. Temporal region usually less roofed than in foregoing families. Initial development, in the Upper Jurassic, of an important marine group; also Cretaceous. Europe and Asia.

†Superfamily Baenoidea

Skull short. Neck possibly a little retractile. Includes forms leading to cryptodires.

†Family Baenidae. Skull roof complete. Upper Cretaceous to Eocene. North America.

†Family Meiolaniidae. Skull roof complete and with protuberances. Upper Cretaceous to Pleistocene. South America and Australian region.

†Family Eubaenidae. Skull roof emarginate. Transitional to cryptodires. Upper Cretaceous. North America.

Suborder Cryptodira

Temporal region of skull frequently emarginate from behind; if completely roofed, probably secondarily so. Neck retracted in a vertical plane. Shell primitively complete, but elements lost or reduced in aquatic, especially marine, species. These are the dominant turtles of today.

Superfamily Testudinoidea

Temporal region usually reduced. Bony shell usually complete and always covered by horny shields.

Family Dermatemydidae. Central American river turtle. Temporal region emarginate from behind. 1 Recent species in Central America and Mexico; others from Late Cretaceous and Early Tertiary of North America. Size medium; adult length to 25 cm (10 in.).

Family Chelydridae. Common and alligator snapping turtles. Head proportionately large, shell, especially plastron, reduced. Temporal region emarginate. Inhabit swamps, rivers, and shallow lakes of North and a little of South America. Aggressive but not dangerous to man. Miocene to Recent. Adult size: Common snapper 20–46 cm (8–18 in.), 5–30 kg (11–66 lb); alligator snapper 42–70 cm (16.5–27.5 in.), 18–100 kg (40–220 lb).

Family Kinosternidae. Mud and musk turtles. Plastron often reduced and in some, singly or doubly hinged. Some exude a disagreeable odour when handled. Pliocene to Recent. Aquatic. Widespread in New World temperate and tropical regions. 21 living species, adult length 7–20 cm (3–8 in.).

Family Testudinidae. Terrestrial turtles, or tortoises. Shell usually with high dome; hinged plastron in 1 species. Temperate and tropical regions of all continents except Australia; also on islands of eastern Africa and Galápagos Islands. Eocene to Recent. About 40 living species. Moderate to large size; adult length 20–100 cm (8–40 in.), weight 1–200 kg (2–440 lb).

Family Platysternidae. Big-headed turtle. Broad, flat plastron; large head. Inhabits moving streams in southeastern Asia. 1 species; adult length about 15 cm (6 in.).

Family Emydidae. This family contains most of the familiar turtles. Temporal region primitively and usually emarginate from behind; bony carapace and horny covering complete. Abundant in Northern Hemisphere; a few species in South America and Africa. Eocene to Recent. About 76 living species; adult length 7–22 cm (3–8.7 in.).

Superfamily Chelonioidea

Marine turtles. Temporal region well roofed. Limbs generally paddle-like.

†Family Toxochelyidae. Nasal bones small or absent (except in one genus). Upper Cretaceous to Eocene. North America and Europe.

†Family Protostegidae. Powerful beak, jaws with large crushing surfaces. Upper Cretaceous to Oligocene of North America and Europe. Some species very large, length up to 4 m (13 ft; *Archelon*).

Family Cheloniidae. Modern sea turtles. Secondary palate, formed by vomer and palatine bones. Distribution worldwide in warmer oceans. Fossils from Cretaceous of Europe, North America, Africa and Asia. 5 to 6 living species; adult length 60–210 cm (24–83 in.); weight 20–500 kg (44–1,100 lb). Most are economically important, for eggs and meat.

Superfamily Dermochelyoidea

Family Dermochelyidae. Leatherback turtle and fossil species. Shell of leatherback without horny plates but with prominent ridges of cornified skin down the back. Worldwide in warm seas. Eocene to Recent. Adult leatherback very large, length to 225 cm (88 in.), weight to 600 kg (1,320 lb).

Superfamily Carettochelyoidea

Family Carettochelyidae. New Guinea plateless turtle and fossil species. Upper Cretaceous to Recent; fossils from North America, Europe, and Asia.

Superfamily Trionychoidea

Family Trionychidae. Soft-shell turtles. Horny shields absent; considerable bony carapace supporting flexible leathery covering of the back. Temporal region widely open. Flat, round body; webs between toes. Nostrils in a projecting snout. Fossils from Cretaceous. Freshwater habitats on temperate and tropical parts of continents except South America and Australia. Adult length 5–60 cm (2–24 in.).

Suborder Pleurodira

Temporal roof may be emarginate from behind, and is invariably emarginate from below. Neck retracted in a horizontal plane.

Family Pelomedusidae. Side-necked turtles. Skull moderately emarginate behind, variably so below. Aquatic habitats in Africa, Madagascar, and South America. Fossils from Cretaceous onward, all continents except Australia. Adult shell length 20–75 cm (8–30 in.).

Family Chelyidae. Snake-necked turtles. Skull slightly emarginate behind, greatly from below. Head and neck may be half of total length. Fossils few, Pliocene to Recent. Present distribution: Australia and South America. Family includes all living land and freshwater turtles of Australia. Adult shell length 12–40 cm (5–16 in.).

†Suborder Eunotosauria

Known from a single partial specimen from Middle Permian of South Africa. 10 dorsal vertebrae known, 8 with expanded ribs. Incomplete carapace with dermal ossifications (in rows). Shoulder and pelvic girdles overlapped by ribs. Upper skull unknown; palate with marginal and palatal teeth.

†*Family Eunotosauridae.* Characters are as for the suborder.

Critical appraisal. This classification now has wide acceptance. The chief difference of opinion has been in regard to the modern soft-shell types such as the aquatic trionychid group and especially the marine leatherback (*Dermochelys*) with its highly atypical leathery shell. The leatherback has, in fact, been considered very primitive and consequently assigned to subordinal rank, rather than being placed among the cryptodires. There is no certainty that *Eunotosaurus* is ancestral to turtles, but the adaptations of its vertebrae and ribs suggest that it is on an evolutionary line leading toward the turtles. (C.H.P.)

Rhynchocephalia (tuatara)

The Rhynchocephalia constitute one of the four orders of living reptiles; the only surviving representative of the group is the tuatara, or sphenodon (*Sphenodon punctatus*). Structurally, the tuatara is not much different from related forms, also assigned to the order Rhynchocephalia, that may have appeared as early as the Lower Triassic Period (more than 200,000,000 years ago).

Distribution

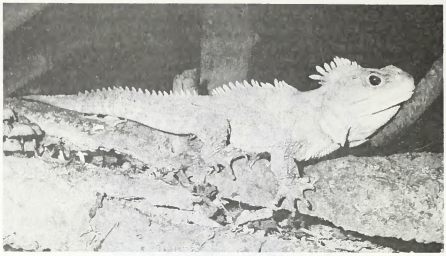
Until recently the tuatara lived on the two main islands of New Zealand. Today, it is found only on certain islets in Cook Strait between the main islands and on islets between East Cape and North Cape of the North Island of New Zealand.

NATURAL HISTORY

Male tuataras may attain a length of 60 centimetres (about 24 inches) and a weight of 1,000 grams (about 2.2 pounds). Females grow to about 50 centimetres (20 inches) and sometimes weigh as much as 500 grams (1.1 pounds). Eight to 15 eggs are laid in a nesting burrow covered by several centimetres of soil. The young emerge about 13 months later, in early summer to midsummer. Adult size is not reached for 50 to 60 years, and the animal may live to the age of 100. Sexual maturity is believed to occur after about 20 years.

During mating, the more prominent crest of the male becomes turgid and erect as he stalks the female, approaching her in a slow, jerky fashion. He then grips her over the shoulders with his forelimbs. Unlike the lizards, the tuatara does not use the jaw for grasping the female. The length of time between mating and egg laying is not known.

The mode of locomotion is primitive. The animal moves in a sprawling fashion, the belly leaving the ground only momentarily. The pattern of limb movement is unusual in that the forelimb contacts and leaves the ground before the hindlimb on the opposite side. As it moves, the body is thrown into marked lateral, or sideward, bends.



Male tuatara with white spines erected, a behaviour characteristic of the tuatara when excited or hunting.
W.H. Dawbin

Tuataras are essentially solitary, nocturnal, burrowing animals, seldom travelling more than a few metres from their burrows during the day. Feeding takes place mainly at night. The principal diet of the young is insects; adults eat, in addition to insects, snails, lizards, young seabird chicks, and eggs. They will drink water if it is available but can survive for months on water obtained from dew and solid food.

The islands inhabited by the tuatara have large numbers of ground insects, which are an important source of food. Bird burrows often serve as ready-made homes for the tuatara, but the animal is capable, from the time it hatches, of making its own burrow if none is already available. The burrow may be shared by a bird and a tuatara, as well as by the bird's chick or egg. (Occasionally the tuatara preys on the chick or adult sharing the burrow.) The largely nocturnal habits of the animal conceal it from many potential predators, but occasionally it is preyed upon by hawks, gulls, or kingfishers. Rats, introduced by man, are serious enemies of eggs and the young. The animal may be regarded as seriously threatened with extinction on islands where rats occur.

Natural enemies

FORM AND FUNCTION

The tuatara has two pairs of well-developed limbs, a strong tail, and a scaly crest down the neck and back. The scales, which cover the entire animal, vary in size. The tuatara also has a bony arch, low on the skull behind the eye, that is not found in lizards. This arch is formed by the presence of two large openings (temporal fossae) in the region of the temple. It has been used as evidence that the tuatara is a survivor of the otherwise extinct order Rhynchocephalia and is not a lizard. The teeth of the tuatara are acrodont—*i.e.*, attached to the rim of the jaw rather than inserted in sockets. The tuatara is also unique among reptiles in not possessing a male copulatory organ. As in birds, sperm transfer is effected during contact between the male and female cloacae.

The heartbeat and respiratory rate of the tuatara are relatively slow; oxygen consumption, accordingly, is low. Tuataras actively forage in temperatures as low as 6° C (43° F) and can briefly tolerate temperatures above 37° C (99° F). Tests indicate that the preferred temperature is about 22° C (72° F)—rather low for a reptile. The excretion of nitrogenous wastes, a by-product of the animal's metabolism, are in the form of uric-acid masses, or concretions. About 10 to 30 percent of this waste may occur as the substance urea in the urine; the urea concentration rises as the amount of protein increases in the food consumed.

EVOLUTION AND CLASSIFICATION

The tuatara has changed remarkably little in skeletal features since the Jurassic Period (136,000,000–190,000,000 years ago), when the closely related *Homoeosaurus* occurred in Europe. There is some evidence that the line of rhynchocephalians began as early as the Lower Triassic Period. Most forms of that time were of moderate size and, except for one specialized family (Rhynchosauridae), they

were never abundant. Fossils are unknown in any region during Tertiary times (2,500,000–65,000,000 years ago). Remains from the Pleistocene Epoch (10,000–2,500,000 years ago) in New Zealand are structurally identical with the living tuatara.

The order Rhynchocephalia belongs to the subclass Lepidosauria of the class Reptilia. Three families (Sphenodontidae, Rhynchosauridae, and Sapheosauridae) are assigned to the order; two families (Claraziidae and Pleurosauridae) are tentatively assigned.

In addition to the living species, *Sphenodon punctatus*, the family Sphenodontidae is represented by European fossils of three genera from the Upper Triassic (200,000,000 years ago); by Asian, European, and North American fossils of genera from the Upper Jurassic Period; and by a southern African fossil of a genus from the Lower Triassic.

The family Rhynchosauridae is represented by fossils of eight genera, some tentatively assigned. All are from the Triassic Period and occurred in Europe, South America, south Asia, southern Africa, and East Africa.

The family Sapheosauridae is represented by one genus from the Upper Jurassic Period in Europe. (W.H.D.)

Sauria (lizards)

Lizards (suborder Sauria) are familiar scaly-skinned reptiles, closely allied to snakes but usually distinguished by the possession of legs, movable eyelids, and external ear openings. Most of the 3,000 living species of lizards inhabit warm regions of the Earth, but some species of lizards are found nearly to the Arctic Circle in Eurasia and others to the southern tip of South America. The lizards are usually considered part of the reptilian order Squamata. The "worm lizards," or amphisbaenians, which are included in this section for convenience of comparison with lizards, are placed in their own suborder, Amphisbaenia, by many modern taxonomists. (For a further discussion of this problem, see the *Annotated classification* in the first section of this article.)

Distribution

GENERAL FEATURES

Lizards are by far the most diverse group of modern reptiles in body shape and size. They range in total length from geckos of three centimetres (1.2 inches) to monitor lizards of three metres (10 feet), and in adult weight from less than a gram (0.04 ounce) to more than 150 kilograms



Figure 10: *Variation in body form among lizards.*
(A) Old World chameleon (*Chamaeleo*). (B) Flying lizard (*Draco*). (C) Wall lizard (*Lacerta*). (D) Basilisk (*Basiliscus*). (E) Horned lizard (*Phrynosoma*). (F) Gecko (*Gekko*). (G) Glass snake (*Ophisaurus*). (H) Worm lizard (*Amphisbaena*). (I) Gila monster (*Heloderma*). (J) Cylindrical skink (*Chalcides*).

(330 pounds). The popular conception of a lizard as a scampering reptile with head and body about 15 centimetres (six inches) long and a slender tail of about equal length may be applied accurately to only a small number of species. Representatives of several families are limbless, resembling snakes; others have elongate hind legs and can run bipedally. A variety of ornaments is found, including extensible throat fans, tail crests, horns or casques on the head, and spines and frills around the throat.

Lizards occupy diverse habitats: underground, on the surface, and in elevated vegetation. Some are slow-moving, relying on cryptic coloration for protection; others can run swiftly across desert sands. The mosasaurs, an extinct family, were strictly marine. Some were giants, attaining lengths of 10 metres (33 feet), with an elongate head, short neck, and long thin body and tail. One living lizard, the marine iguana of the Galápagos Islands, feeds on algae in the sea. No other extant species is marine, but several are partially aquatic, feeding on freshwater organisms.

Importance to man

Modern lizards do not play a major role in human ecology. Some large species (iguanas) are eaten, and others are used for leather goods. Predators, such as the tegu, can be pests around chicken farms, but the vast majority are primarily insectivorous or consume undesirable rodents. Small species such as geckos, which tend to live around houses, are often welcomed because of their efficiency in destroying insects; they may transmit *Salmonella* bacteria, however. In general though, lizards are not disease vectors.

Only two species, the Gila monster of the southwestern United States (*Heloderma*) and a close relative in Mexico, are venomous. They bite man only when provoked and fatalities are very rare. Unfortunately, diabolical powers attributed to lizards have made them objects of fear in many countries. Lizards are valued as subjects for biological research: the varied modes of reproduction and ability to regulate body temperatures are only two of many areas studied by comparative physiologists; the great abundance and easy observability of numerous species make them ideal subjects for ecologists and ethologists; the ability of some species to regenerate broken tails has led to their use by developmental biologists. In addition, because they are clean and easy to keep, lizards are quite popular as house pets.

NATURAL HISTORY

Life cycle. Most lizards reproduce by laying eggs. In some small species the number of eggs is rather uniform for each laying or clutch. For example, all anoles lay but a single egg at a time; many geckos lay one or two eggs (depending upon the species), and some skinks have fixed clutches of two eggs. A more general rule is that clutch size varies with size, age, and condition of the mother. A clutch of four to eight eggs may be considered typical, but large species such as iguanas may lay 50 or more eggs at one time. Lizard eggs are usually leathery shelled and porous and can expand by absorption of moisture as the embryos grow. An exception occurs in the majority of egg-laying geckos, whose eggs have shells that harden soon after deposition and then show no further change in size or shape.

Some lizards bear live young. In the family Scincidae this is true for about one-third of the species, many of which are tropical. In most other families that have live-bearing representatives, the species that are exposed to cold—either at high altitude or at extreme latitude—generally are the live bearers. For example, all New Zealand geckos have live young, yet all other geckos lay eggs. Live bearers are divided into two categories. Ovoviviparous forms essentially retain eggs in the oviduct with no definite shell laid down, whereas truly viviparous forms have a chorioallantoic placenta. Whether there is an actual exchange of nutrients across this placenta, as in mammals, is currently under study.

Sex in lizards is genetically and rigidly determined; a hatchling normally has either only male or female reproductive structures. In representatives of at least three diverse families (Iguanidae, Teiidae, Pygopodidae), the males have dissimilar sex chromosomes. This is similar to the sex-chromosome system of most mammals but differs from snakes, in which sex-chromosome differences, when present, are found in the female.

Most lizard populations are evenly divided between females and males. Deviations from this pattern are found in parthenogenetic (all female) species, in which the young are produced from unfertilized eggs. In the Caucasus all-female races of *Lacerta* and in the southwestern United States and parts of Mexico all-female species of whiptailed lizards (*Cnemidophorus*) have interesting parallels. They appear to live in areas ecologically marginal for representatives of their genera. In the case of *Cnemidophorus*, there is convincing evidence that parthenogenetic forms arose through hybridization between two bisexual species.

Parthenogenesis

Parental care among lizards tends to be minimal following egg deposition. Many species dig holes in which the eggs are placed; others bury them under leaf litter or utilize a cranny in a tree or cave. Some species, however, notably the five-lined skink (*Eumeces fasciatus*) of the United States, remain with their eggs throughout incubation time (about six weeks), leaving only infrequently to feed. These skinks turn the eggs regularly and, if the eggs are experimentally scattered, will return them to the nest cavity. As soon as the young disperse, family ties are severed.

Certain lizards, particularly some geckos, are known to be communal egg-layers, many females depositing their eggs at the same site. Whether this is due to social interaction or simply to ideal site conditions has not been determined.

Juvenile lizards are essentially miniature adults; they do not go through any larval phase nor any stage of dependence upon adults. They often differ from the adult in colour or pattern and in certain body proportions. For example, heads of hatchling lizards of some species tend to be proportionally larger than those of adults. Certain ornamental structures, such as the throat fan of the male American chameleon (*Anolis*) or the horns of some true chameleons, develop as the lizards become sexually mature.

Some of the smaller lizards mature very quickly, and population turnover is essentially an annual event. The small, side-blotched lizards of the iguanid genus *Uta* of western North America have this type of population dynamics. The young hatch in July and reach sexual maturity that autumn. At this time males undergo spermatogenesis and mating takes place. Females accumulate large quantities of fat, which appears to be utilized in the production of eggs the following spring. Losses of 90 percent or more adults per year may come from predation, inclement weather, or other environmental variables. A single species living under a variety of environmental conditions may have very different population dynamics in different regions. For example, in areas with long winters there are periods of hibernation, coupled with greater longevity and slower population turnover.

Population turnover

Large lizards may take several years to reach sexual maturity. Unfortunately, there is little information on the dynamics of natural populations of most lizard species. In captivity many species are long-lived. There is a report of a male slowworm (*Anguis fragilis*) 46 years old mating with a female 20 years old. Gila monsters (*Heloderma*) have been kept in captivity for more than 25 years, and even some small geckos have been kept for as long as 20 years.

Ecology. The most important environmental variable to a lizard is almost certainly temperature. Like fish and amphibians, lizards are ectothermic ("cold-blooded"); they tend to assume the temperature of their surroundings. Yet all temperatures are not equally acceptable to lizards. Most species seek out relatively specific body temperatures, called "preferred temperatures," mostly ranging from 28° to 38° C (82°–100° F). Although metabolic energy is not utilized to control body temperature, considerable thermoregulation can be accomplished largely through behavioral means, if the lizard has a choice. Typically, a diurnal lizard will emerge early in the morning and will sun itself, orienting the body to maximize exposure to the sun, until the preferred temperature is approached. The ability to absorb heat from solar radiation may permit the lizard to warm itself well above air temperatures. For example, *Liolaemus multiformis*, a small iguanid that lives high in

Temperature regulation

the Andes, has the ability to raise its body temperature to 35° C (95° F), while air temperatures are at 10° C (50° F) or lower.

The preferred body temperature plays a critical physiological role in the life of a lizard. It has been demonstrated in the laboratory that animals kept without the opportunity to achieve the preferred temperature may be unable to attain the normal reproductive condition or may become sterilized. There is also evidence that enzymes catalyze reactions most efficiently at or near the preferred body temperature.

A lizard living in the tropics often will find the immediate environmental temperatures within its preferred range. To be successful in other habitats, a species must be able either to function at less than optimal temperature or, at least, to have less time at the optimal temperature and spend more time attaining this temperature. Thus it is not surprising that diversity of lizard species decreases with increasing latitude or with high altitude within the tropics.

Water is less of a problem to lizards than is temperature. All reptiles excrete uric acid and hence do not need great amounts of liquid to get rid of nitrogenous wastes. Many lizards have salt glands for active excretion of mineral salts. Deserts, therefore, do not pose severe problems to lizards and actually serve as a major habitat for the group. Lizards form a conspicuous portion of the fauna of oceanic islands, where amphibian and mammalian species' diversity is generally low. The ability of lizards to withstand desiccation may well account for their success at colonizing oceanic islands. They are believed to arrive on mats of floating vegetation washed up on the island shore. Hard-shelled gecko eggs seem to be particularly equipped for such journeys.

Other variables that affect lizards are day length (photoperiod) and rainfall. Lizards living far from the equator experience marked variation in photoperiod, with short winter days and long summer days. Certain species are adapted to respond to such cues. *Anolis carolinensis* of the southeastern United States ceases reproduction in the late summer and fattens up for winter hibernation. This change occurs while the days are still warm and appears to be triggered by decreasing day length. It is adaptive for the species because eggs laid in September would essentially be wasted, the young hatching in November when food is not readily available and temperatures are too low to permit efficient, rapid growth. Some tropical species respond to alternations between rainy and dry seasons, and egg-laying may cease during the driest months of the year. This may be adaptive for a variety of interrelated reasons. When food abundance is low, it is of direct advantage to the parent not to channel energy into production of eggs. Additionally, eggs might be less viable because of desiccation.

Hibernation

Lizards provide valuable models for the study of competition between species. On some Caribbean islands as many as ten species of *Anolis* may live in a single restricted area. For so many species to be accommodated, each must be specialized for a rather precise niche. The species come in a variety of sizes; feed on different sizes of prey, and have different preferences for structural and climatic niches. Some live in tree crowns, others on trunks, others in grass; some live in open sun, others in "filtered" sun, and still others in deep shade. Thus, with ten anoles in a single area, each species has its own characteristic microhabitat.

Behaviour. Most lizards are active during daylight hours, when acute binocular vision is necessary for most nonburrowing species. The family Gekkonidae, however, is composed predominantly of species that are most active from dusk to dawn. In conjunction with night activity, geckos are highly vocal and communicate by sound, whereas most other lizards are essentially mute.

Lizards spend considerable time obtaining food, usually insects. Iguanids and agamids tend to perch motionless at familiar sites and wait for prey. They are attracted by motion and dart suddenly for the capture. The true chameleons move slowly, carefully observant, each eye moving independently, and capture their prey by shooting out the sticky tongue. Lizards of other groups actively search for prey by probing and digging, using scent as well as visual cues. Plants are eaten by many of the largest lizards, such as the iguanas and the spiny-tailed agamid (*Uromastix*).

Lizards themselves are food for many birds, mammals, and reptiles, and they have many defensive mechanisms. The chuckwallas (*Sauromalus*) remain close to rock piles. When danger threatens, they move into crevices and puff themselves up so that extrication is most difficult. A number of spiny-tailed forms also move into crevices and expose only the formidable tail. Perhaps the highest development of this sort of defense is found in the African armadillo lizard (*Cordylus*), which holds its tail in its mouth with the forefeet and presents a totally spiny appearance to an attacker. To intimidate intruders on its territory, the frilled lizard of Australia (*Chlamydosaurus*) extends a throat frill almost as wide as the lizard is long. The tails of many lizards break off (autotomize) easily. The broken-off section wriggles rapidly, distracting the predator as the tailless lizard scurries for cover.

Defense

Social interactions among lizards are best understood for the species that respond to visual stimuli. Many lizards defend certain areas against intruders of the same or

Based on photographs in (A,B) K.P. Schmidt and R.F. Inger, *Living Reptiles of the World* (1957), Doubleday & Company Inc. (C,D) R. Mertens, *The World of Amphibians and Reptiles*, and (E) R. Stebbins, *A Field Guide to Western Reptiles and Amphibians*, Houghton Mifflin Company

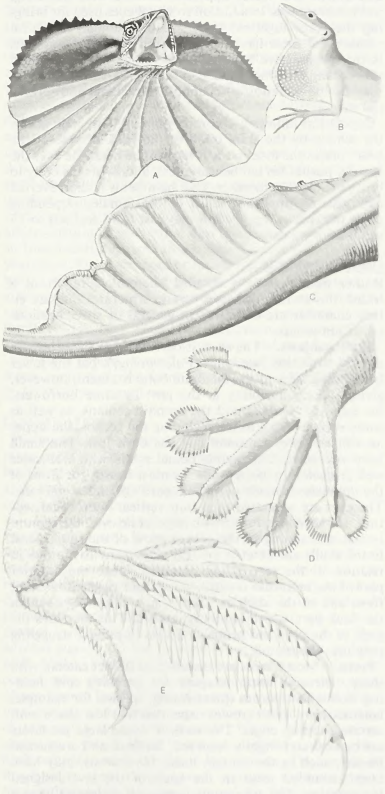


Figure 11: *Specialized structures of some lizards.* (A) Frill of the frilled lizard (*Chlamydosaurus*). (B) Dewlap of the anole (*Anolis*). (C) Fin of the water lizard (*Hydrosaurus*). (D) Toe fans of the fan-footed gecko (*Ptydoctylus*). (E) Toes of the fringe-toed lizard (*Uma*).

closely related species. Territorial defense does not always involve actual combat. Elaborate, ritualized displays have evolved, presumably to avoid physical harm. The displays often involve erection of crests along the back and neck, increasing apparent size; showing bright colours, either by extending a throat fan or exposing a coloured patch of skin; and stereotyped movements such as pushups, head bobbing, and tail waving. Impressive ornamentation is often restricted to males, but females of many species defend territories by stereotyped movements similar to those of males. Energy is expended in territorial defense, and a displaying male is very conspicuous and vulnerable to predation; but territoriality is evidently advantageous and has evolved through natural selection. Successful defense of a territory increases the chances of obtaining the necessary amount of food and at the same time increases the probability of a successful mating with any females living within the defended area. Thus a successful territorial male has a higher probability of reproductive success than one living in a marginal area. The displays used by males in establishing territories also may function to "advertise" their presence to females; in species that breed seasonally, territoriality diminishes during the nonbreeding season. In iguanids, actual courtship displays differ from territorial displays in that males approach females with pulsating, jerky movements. In addition to visual cues used for bringing the sexes together, there are undoubtedly chemical stimuli, but these have not been well studied. Numerous lizard species have femoral pores, which are small blind tubes along the inner surface of the thighs, whose function may be the secretion of chemical attractants and territorial markers.

Courtship

Copulation follows a common pattern. The male grasps the female by the skin, often of the neck or side of the head, places the fore and rear legs on the near side over the female, pushes his tail beneath hers, and twists his body to bring the cloacae together. One hemipenis is then everted and inserted into the cloaca of the female. Depending upon the species, copulation may last from seconds to 15 minutes or more.

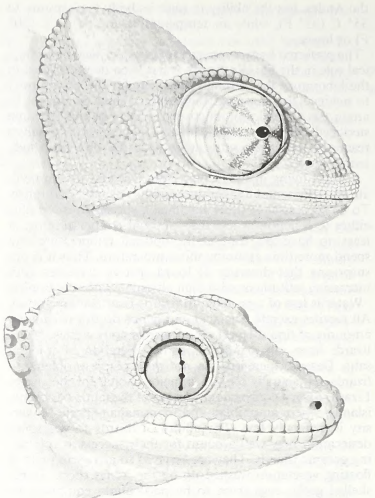


Figure 12: *Specialized eyes of two lizards.* (Top) Turreted eye of chameleon (*Chamaeleo*). (Bottom) Slit pupil of gecko (*Gekko*).

FORM AND FUNCTION

Rather than present a detailed anatomical report of a lizard, this section discusses certain structures that are either characteristic of lizards in general or specializations of certain groups.

Skull and jaws. The skull is derived from the primitive diapsid condition (see below *Paleontology*); but the lower bar leading back to the quadrate bone is absent, however, giving greater flexibility to the jaw. In some burrowers, for example *Anniella* and the amphisbaenians, as well as some surface-living forms including the geckos, the upper as well as lower temporal bar has been lost. The small burrowers have thick, tightly bound skulls, with braincases well protected by bony walls. In most lizards the front of the braincase is made up of thin cartilage and membrane. The eyes are separated by a thin vertical interorbital septum. In burrowing forms with degenerate eyes, the septum is reduced, adding to the compactness of the skull. Most lizard skulls are kinetic; *i.e.*, the upper jaw can move in relation to the rest of the cranium. Since the anterior part of the braincase is cartilaginous and elastic, the entire front end of the skull can move as a single segment on the back part, which is solidly ossified. This increases the gape of the jaws and probably assists in pulling struggling prey into the mouth.

Teeth. Most lizards are insectivores (insect eaters), with sharp, tricuspid teeth adapted for grabbing and holding. Some herbivorous (plant-eating) species, for example, iguanas, have tooth crowns expanded to a leaf shape with serrated cutting edges. The teeth of some large predators are conical and slightly recurved. Mollusk and crustacean feeders, such as the caiman lizard (*Dracaena*), may have blunt, rounded teeth in the back of the jaw designed for crushing. The venomous lizards (*Heloderma*) have a longitudinal groove or fold on the inner side of each mandibular tooth. These grooves conduct the venom. In most lizards, teeth are present along the jaw margin (on the maxilla, premaxilla, and dentary bones) and, in some forms, on the palate. In the embryo, an egg-tooth develops on the premaxilla and projects forward from the snout. It aids in piercing the shell and is lost soon after hatching. This is a true tooth, unlike the horny epidermal point in turtles and crocodilians.

The egg-tooth

The common mode of tooth implantation is pleurodonty, in which the teeth are fused to the inner side of the labial wall. In the other mode, acrodonty, teeth are fused to the tooth-bearing bone, often to the crest of the bone. Acrodont teeth are rarely replaced, once a certain growth stage is reached. The dentition of the Agamidae is usually described as acrodont, but most species have several pleurodont teeth at the front of the upper and lower jaws.

Locomotion and limb adaptations. The majority of lizards are quadrupedal, with powerful limb musculature. They are capable of rapid acceleration and possess great ability to change direction of motion rapidly. The racerunners (*Cnemidophorus*) can attain speeds of 24 kilometres (15 miles) per hour, which, in terms of their own body length, puts them in a class with fast terrestrial mammals. A tendency toward elongation of the body is found in some families, often accompanied by reduction of limb length and, not infrequently, complete loss of limbs. Such lizards propel themselves entirely by lateral undulations emanating from highly complicated ventral abdominal musculature. Limbless lizards that move quickly on the surface or through sand tend to have elongate tails (*Ophisaurus*, the glass "snake"), whereas the burrowers have extremely reduced tails (amphisbaenians). The burrowers (amphisbaenians in particular) dig by ramming the head into the substrate, then rotating the head around the head joint, compacting the substrate.

Many modifications of the toes are to be seen in lizards. Some desert geckos, the iguanid *Uma*, and the lacertid *Acanthodactylus* have fringes on the toes that provide increased surface area and prevent sinking into loose desert sand. Arboreal geckos and anoles have lamellae (fine plates) on the undersides of the toes. Each lamella is made up of brushlike setae, each double; and the tips divide several times, the final strand less than .25 micrometre (1/100,000 in.) in diameter. These fine hairlike processes greatly enhance the clinging ability of the lizards, as they are able to find purchase in the smallest irregularities of

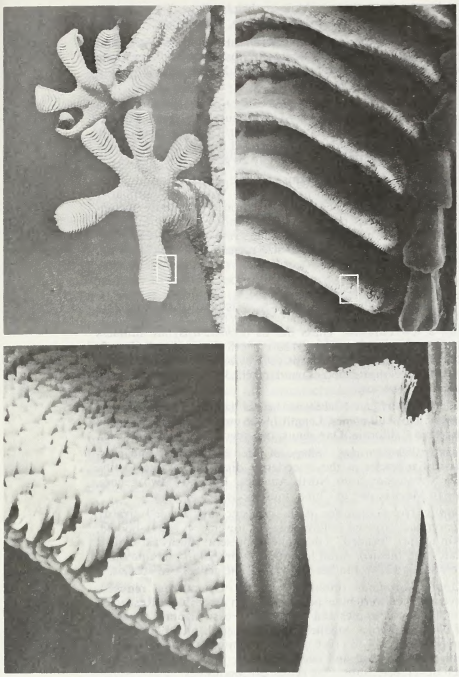


Figure 13: *Fine structure of gecko foot.*
(Top left) Foot of gecko (*Gekko gecko*) showing chevron-shaped ridges or lamellae (about life size). White rectangle indicates the area covered by the photograph at the next larger magnification. (Top right) More detailed view of six of the lamellae (magnified 25×). (Bottom left) Part of lamella revealing its surface, covered by setae (magnified 220×). (Bottom right) Seta with attached suction cups visible at fringe terminals (magnified 2,550×).
Joseph F. Gennaro, Jr.

the substrate. Geckos and anoles can easily climb vertical panes of glass.

The true chameleons (Chamaeleontidae), a predominantly arboreal family, have a different type of highly specialized limb. The digits on each foot are divided into two groups by webs of skin. On the hind limb, three of the toes are on the outside, two on the inside (a form of zygodactyl); on the forelimb the pattern is reversed. Each foot is thus divided into an outer and an inner portion, which can be opposed as the branch is gripped. In addition, these and some other lizards have prehensile tails, which aid in grasping branches.

Bipedal locomotion

Several terrestrial iguanids and agamids are able to run bipedally. Basilisks are actually able to run across water for short distances. During bipedal locomotion the tail is held out backward and upward and acts as a counterweight.

Some lizards are able to parachute or glide through the air and make soft landings. The most highly adapted of these are the agamid lizards called flying lizards (*Draco*), which have extensible lateral expansions of the skin, supported by elongate ribs.

Skin and colour change. Except for openings of nostrils, mouth, eyes, and cloaca, most lizards are completely covered in scales, which vary among species and on different parts of the body of the individual. Scales may be smooth and overlapping, form a mosaic of flat plates, or have keels or tubercles. Osteoderms, bony plates that develop in the dermis, underlie head and body scales of some lizards. The outer parts of the scales are composed of dead horny tissue made up largely of the protein keratin. The dead layer is shed at intervals and is replaced by proliferating cells in the deep part of the epidermis. Certain lizards have scale organs, with a stiff projecting seta emerging from the serrated edge of the scale and presumably responsive to tactile stimuli.

Many lizards can show some colour change. Two groups, the chameleons and the anoles, are particularly noteworthy in this ability. They can change from bright green to deep, chocolate brown, and patterns such as lines and bars may appear and disappear. The pigment cells that permit colour change are melanophores. Within these cells, pigment granules are able to migrate. In general, the animal is light when pigment is concentrated and dark when it is dispersed throughout the cells. The actual colour state, complexly controlled, is not simply matching of background but seems to be controlled by an interaction of hormones, temperature, and the nervous system.

PALEONTOLOGY

Lizards belong to the diapsid reptiles, a group characterized by the presence of an opening in the temporal bone of the skull, both above and below a bar formed by the postorbital and squamosal bones. The diapsids comprise two subclasses of reptiles: the Archosauria (crocodiles, pterosaurs, dinosaurs) and the more primitive Lepidosauria (lizards, snakes, eosuchians, and rhynchocephalians). The earliest known Lepidosaurian order (Eosuchia) appeared in the Late Permian (about 230,000,000 years ago) and almost undoubtedly was ancestral to lizards. Modern lizards differ from their ancestors in several skeletal features, the most diagnostic differences being in the construction of the upper temporal and quadrate regions of the skull. A.S. Romer divides true lizards from their ancestors on the basis of the quadrate region. Among true lizards the quadrate articulates flexibly with the squamosal above it. This occurred through loss of the lower temporal arch and reduction of squamosal development. The oldest fossil true lizards appeared in the Late Triassic (200,000,000 years ago). These were not primitive but highly specialized, having ribs that could be projected outward to provide a gliding ability. Thus it is possible that even older lizards will be found. Representatives of many modern lizard families have been found in Cretaceous deposits (from 136,000,000 to 65,000,000 years ago); and representatives of modern genera, virtually indistinguishable from living forms, were common from the Oligocene (38,000,000 to 26,000,000 years ago) on.

CLASSIFICATION

Distinguishing taxonomic features. Lizards differ from other reptilian groups in several anatomical characters, including the possession of not more than a single bony arch across the temple and paired reversible copulatory organs, the hemipenes.

Lizards and snakes distinguished

G. Underwood (1967) summarized the differences between snakes and lizards. Lizards have the eyes separated by a thin vertical interorbital septum, whereas in snakes this septum is absent and the entire width of the bony braincase separates the orbits. Other features that distinguish lizards from snakes include: (1) the pectoral (shoulder) girdle, absent even as a trace in all snakes, present in all lizards; (2) the histological structure of the adrenal glands; (3) the placement of the thymus bodies. M.R. Miller (1968) found that the internal ear of snakes is considerably different from and more primitive than that of any lizard. The families of lizards are defined primarily by osteology and dentition, as well as presence or absence of the rectus superficialis muscle. When this muscle is absent in a group, there is not the slightest tendency for limb reduction and serpentine locomotion. The limits of the families seem to be agreed upon by most systematists, although relationships among the families are not always so clear.

Annotated classification. The classification presented here is adapted from Underwood (1957), which is an updating of C.L. Camp (1923). Definitions of taxa follow

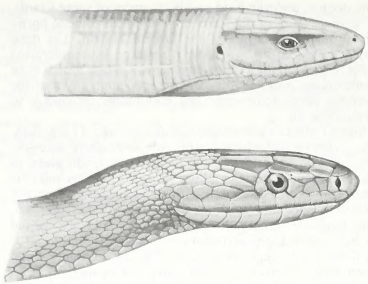


Figure 14. *Comparison of heads of a typical lizard and a typical snake.*
(Top) Legless lizard (*Ophisaurus*). (Bottom) Green rat snake (*Elaphe triaspis*).

Camp where feasible. The classification of the Gekkonidae follows A.G. Kluge (1967). The dates of earliest fossils are from R. Hoffstetter (1962).

SUBORDER SAURIA

Reptiles with a single temporal opening lying above the bar formed by postorbital and squamosal bones. Pectoral girdle and interorbital septum always present. About 3,000 species.

Infraorder Ascalabota

Rectus superficialis rarely present; usually more than four transverse rows of ventral scales found over each body segment; body is covered by scales that overlap by a wide margin ("deciduous").

Superfamily Gekkonidea

Family Gekkonidae (geckos). No supratemporal arch; well defined limbs. Jurassic to present. Eighty-two genera, 650 species. Virtually worldwide in subtropical and tropical regions. Adult total length from slightly under 4 cm to 35 cm (1.5–14 in.). Most with soft, granular skin, easily autotomized tails, no movable eyelids, and toes with fine setae.

Family Pygopodidae (flap-footed lizards). Australia and New Guinea. Some burrowers, others surface dwelling; hind limbs represented by scaly flaps. Eight genera, 15 species.

Family Dibamidae. Philippines and Vietnam, to New Guinea. Burrowers; females limbless, males with flaplike hind limbs. One genus, 3 species.

Family Anelytropsidae. Eastern Mexico. Limbless; one species.

Superfamily Iguanioidea

Family Iguanidae. Dentition pleurodont. Teeth replaced when lost. Well-developed limbs. Movable eyelids. Oligocene, possibly late Eocene to present. Dominant family of North, Central, and South America, West Indies, and Galápagos Islands. One genus on Fiji and Tonga, two genera on Madagascar. Adult lengths from 10 cm (4 in., *Uta*) to 2 m (6.6 ft, *Iguana*). A family of diverse forms including spiny desert dwellers (*Phrynosoma*), arboreal species (*Anolis*), and sea-going iguanas (*Amblyrhynchus*). Fifty genera, 600 species (200 in one genus, *Anolis*).

Family Agamidae. Most teeth acrodont. Many examples of convergent evolution with iguanids, but tooth difference absolute. Late Cretaceous to present. Old World tropics and subtropics, except Madagascar. Thirty-three genera, 300 species (60 in genus *Agama*).

Superfamily Rhiptoglossidea

Family Chamaeleontidae (Old World chameleons). Feet zygodactylous, grasping. Dentition acrodont. Tongue long, slender, extensile. Two genera, *Brookesia* 16 species, *Chamaeleo* 68 species. About half restricted to Madagascar, remainder primarily African, two species in western Asia, one in India, one reaching Mediterranean region. Most are insectivores 17–25 cm (7–10 in.); the largest, reaching 60 cm (24 in.), will eat birds.

Infraorder Autarchoglossa

Superfamily Scincomorphoidea

Family Scincidae (skinks). Skull arches present, osteoderms (dermal bone) present. Late Cretaceous to present. Worldwide (except in polar and subpolar regions), greatest diversity in Old World tropics. Typically with conical heads, cylindrical bodies, tapering tails. Less variable than the other large families. Primarily ground dwelling or burrowing, some arboreal. Limb reduction common. Majority below 12 cm (5 in.), largest reaches 60 cm (24 in.). About 50 genera, 800 species.

Family Cordylidae (girdle-tailed lizards). Femoral pores and osteoderms present. Supratemporal fossa roofed over. Late Jurassic to present. Southern Africa and Madagascar. 20–60 cm (8–24 in). About 10 genera, 50 species.

Family Lacertidae. Osteoderms absent, supratemporal fossa roofed over. Eocene (possibly Cretaceous) to present. Europe, Asia, and Africa. Morphologically uniform, conical heads, scaly bodies, movable eyelids, well developed limbs and tail. Length 15–60 cm (6–24 in.). Approximately 20 genera, 150 species.

Family Teiidae. Osteoderms absent, supratemporal fossa open. Late Cretaceous to present. New World only, primarily in tropics and subtropics. Great variation in the family, including large terrestrial predators (*Tupinambis*), large semiaquatic snail eaters (*Dracaena*), lacertid-like "racerunners" (*Cnemidophorus*), and small burrowers. Size range 7–120 cm (3–48 in.). Forty genera, 200 species.

Superfamily Anguinomorphoidea

Family Anguidae. Skull arches, osteoderms present. Six mandibular bones. Late Cretaceous to present. Most in Americas, a few Eurasian. Glass lizards (*Ophisaurus*) are limbless "grass swimmers" reaching 120 cm (48 in.). Alligator lizards (*Gherrhonotus*) and galliwasps (*Diploglossus*) have four limbs, somewhat elongate bodies, and reach 37 cm (15 in.). Seven genera, 67 species.

Family Anniellidae (California legless lizards). Limbless burrowers. No skull arches. Length 11–25 cm (4–10 in.). California and Baja California. One genus, two species.

Family Xenosauridae. Shape of interclavicle bone and presence of tubercles in the osteoderms distinguishes the family. Late Cretaceous from North America. Presently, two genera, one in Mexico, one in China; four species.

Family Helodermatidae (Gila monster and beaded lizard). Venomous; grooved hollow fangs in lower jaw; heavy-bodied. Skin texture "beaded." Oligocene to present; southwest United States and Mexico. Adult length to 50 cm (20 in.) in Gila monster, 80 cm (32 in.) in beaded lizard. One genus, two species.

Family Varanidae (monitor lizards). Osteoderms reduced, seven cervical vertebrae, postorbital arch incomplete. Restricted to Old World tropics and subtropics. Fossils from Cretaceous of North America. Smallest species is adult at 20 cm (8 in.), largest exceeds 3 m (10 ft). General uniformity of appearance with elongate head and neck, relatively heavy body, long tail, well developed limbs. One genus, 30 species.

Family Lanthanotidae (earless monitor lizard). Fossils from Recent only. No external ear. Restricted to Borneo; rare and little known. Length to 40 cm (16 in.). One genus and species.

Family Xantusiidae (night lizards). Vertebrae procoelous, supratemporal arch present. No movable eyelid. Relationships to other lizards unclear, even as to assignment within suborder. Share some features with gekkonids, others with skinks. Four genera, 12 species. California through Central America, and Cuba. All moderately small; bodies covered with small scales; ventral surfaces and head with larger platelike scales.

SUBORDER AMPHISBAENIA

Family Amphisbaenidae (amphisbaenians or worm lizards). Long cylindrical bodies with shallow grooves. No external ears, no readily visible eyes. Burrowers. Eocene to present. Primarily tropical, one species reaches Florida, and one reaches Spain. All genera lack hind limbs, and all but three lack forelimbs. Fifteen genera, about 100 species.

Critical appraisal. The boundaries of most lizard families have been stable for the past half-century, with several exceptions. The most important involves Gekkonidae, which Underwood (1954) divided into three families but which Kluge (1967) reunited into a single family with four subfamilies. C. Gans (1960) considers the Amphisbaenia so distinct that he raised them to ordinal status, equivalent in rank to lizards and snakes, and raised the two amphisbaenid subfamilies to family status. Studies by M.R. Miller (1966, 1968) on the internal ear, and A.E. Greer (1970) on cranial osteology clarify the relationships of some of the obscure burrowers and demonstrate that members of the so-called family Feylinidae are merely specialized skinks and are not closely related to the superficially similar Anelytropsidae. Perhaps of greatest significance is the finding that the cochlear duct of all snakes is more primitive than that of any lizard. This throws doubt on the theory that snakes are derived from highly advanced lizards (varanid relatives). (G.C.G.)

Serpentes (snakes)

Serpentes, one of three suborders of the order Squamata, includes only the snakes, a widely known and much misunderstood group. The closest relatives of the snakes are the lizards (suborder Sauria) and the wormlike amphisbaenians (suborder Amphisbaenia).

Arms and legs gone, no ears, only one functional lung, voiceless, eyelids missing—a human being in such condition would be institutionalized and under constant care. Snakes have not only survived these losses but have become highly successful in a great variety of ecological roles. Fearless, independent, and usually solitary, the snake has capitalized upon these apparent deficiencies to become an efficient second- and third-level predator, focussing perhaps 70 percent of its existence on tracking down, catching, and digesting its living prey. The elongate, limbless body of the snake permits it to be soundless in motion, invisible at rest, and almost unlimited in its access to hiding places, such as nests, tree holes, crevices, brier patches, gopher tunnels, and other shelters where small vertebrated animals seem to feel themselves safe from their enemies. Most of the remainder of the snake's existence is dedicated to continued survival in an antagonistic world that is usually either too hot or too cold and is full of other organisms challenging the snake's right to live. With its body temperature almost totally dependent upon the surroundings, the snake must seek out both nightly and seasonal resting places where rising temperatures eventually can be relied upon to rekindle its abilities to move, to sense, and to reproduce. Reproduction demands only a few percent of the male snake's existence, for it consists of little more than a sudden burst of highly seasonal activity followed by an immediate return to the more pressing activities of food search. The female devotes slightly more time to reproductive activities, because she must eventually seek out a suitable place to deposit the eggs or to give birth to a usually good-sized brood. This, too, however, is an act of short duration, followed immediately by return to the continuing problems of daily existence.

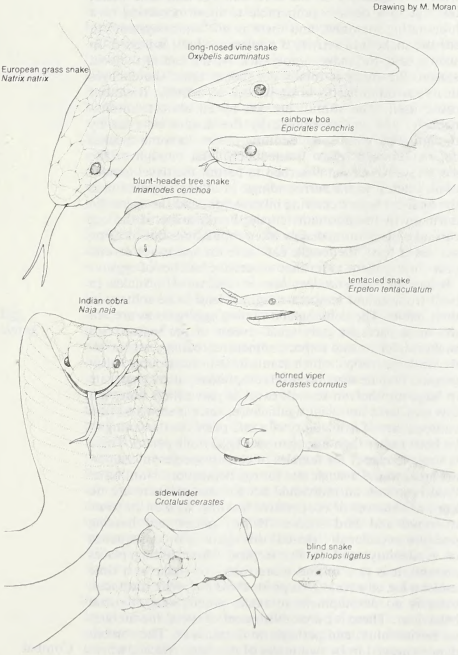


Figure 15: Variation of snake head shapes.

Snakes are found around the world in practically every kind of habitat and in all regions except near the poles. They are not successful competitors with man, and only a few diminutive species with very secretive habits find it possible to share living space with humans. When man moves in, practically all reptiles share the fate of large mammals and predatory birds—they must move out or die, because man is not and never has been a particularly good coexisting species except with those organisms that he bends to his will, such as corn or horses, or with those capable of capitalizing upon his presence, such as pigeons and rats. Of the reptiles, the snake is the most distasteful to man, and snake species are quickly wiped out. In rural areas, man pays the penalty for this eradication in an inevitable increase in the natural prey of the snakes, including rats, mice, and other rodents.

IMPORTANCE

After the unfortunate episode in the Garden of Eden, perhaps the first relationship between man and snakes that one thinks of is envenomation from poisonous snakebite.

Snakebite

The likelihood of any individual human encountering a venomous snake during his lifetime has been reduced to almost zero as a result of the tremendously increased world population, the concentration of humans in cities, where poisonous snakes cannot survive, and the reduction in overall numbers of snakes as a consequence of the invasion and destruction of their habitats by man. At the same time, the likelihood of any one poisonous snake, sometime during its life, finding itself in a position in which it must bite a human in self-defense increases geometrically, as a consequence of the enormously inflated human population, man's ever-increasing mobility, and the increased utilization of remote areas for recreation purposes of all sorts. One of these likelihoods is offset by the other, and the overall number of bites per year by venomous snakes around the world remains rather constant, at about 1,000,000 per year. Of this 1,000,000, perhaps 30,000 to 40,000 result in death, with most of the deaths occurring in areas such as tropical Africa and Asia because of the inadequate medical facilities for proper treatment. It has been estimated that in the United States about 1,000 bites occur each year, with approximately 15 resulting in death. Records for bites by the viper in England indicate that only seven deaths occurred between 1899 and 1945 and only one death between 1945 and 1960, with the latter a consequence of treatment rather than of the bite. A large percentage of the bites in the United States occur when the snakes are handled for various reasons. These reasons include care in zoos, the extraction of venoms, exhibition by showmen in carnivals and sideshows, the maintenance of snakes as pets or curiosities, and their inclusion in religious ceremonies by various sects. Practically all of the time-honoured remedies or first-aid measures for snakebite treatment are more dangerous in the hands of the inexperienced than is the bite itself. If any kind of medical care is available within an hour or less, nothing more than a light tourniquet is needed as emergency treatment.

No snake has ever regularly hunted man, but man continues to engage in seeking out the snake. The uses that man finds for snakes are not extensive, but each of them results in the decimation of wild populations. Snakes are collected as a source of rare and exceedingly expensive leather for belts, pocketbooks, shoes, and gloves. There is a phenomenal traffic in live snakes for purposes of display or as pets, including the flow to zoos, circuses, carnivals, or roadside "snake farms," as well as to the amateur snake fancier with a large collection in his basement. There are perhaps 40–50 research laboratories and antivenin institutes around the world, in which either snakes or snake venoms are used, and live snakes are needed in fairly large numbers for the work. An indication of the numbers of animals involved in this kind of traffic is given by data from the United States Department of the Interior on importations, taken from customs reports. During 1968 and 1969, more than 3,344,000 specimens of living reptiles

(including snakes) were brought into the United States. It is difficult to determine how many of these reptiles were snakes, but perhaps a quarter to a third of the total would be a reasonable estimate, which would mean that almost 1,000,000 snakes were shipped into the United States in those two years. The mortality rate among such snakes is high, but those that survive bring prices from about $5 to as much as $1,500, and, therefore, this is a multimillion-dollar business. The snake roundup, or snake rodeo, has become recognized as a tourist attraction in the United States, and such events take place in about a dozen different cities each year. The emphasis is on capturing and killing rattlesnakes or copperheads, but usually any snake will do, and the slaughter is always extensive. This phenomenon will die out as the number of available snakes decreases. Distaste for snakes has often led legislatures or monarchs to promise bounties for those killed in areas under their jurisdiction, and such bounties invariably result in the importation of animals from many miles away. In a single year, 1889, 578,415 snakes were turned in for bounties in India. Although Western cultures have never developed the taste, in some parts of Asia the snake is considered a choice delicacy and, by some, even as somewhat of an aphrodisiac.

Snakes as symbols

The serpent is a recurrent figure in religious beliefs, ceremonies, activities, and legends. The decidedly unhuman appearance of the snake, its unwinking eye, its reputed ability to deliver unexpected and sudden death, the fear, hatred, and loathing with which it is regarded by many, the folktales and legends that have surrounded it in practically every culture, all have combined to produce religious overtones in many different parts of the world. While in many religions the snake is primarily a symbol, direct worship of the snake as a godlike creature (not as an indirect representation) is not uncommon. Python worshippers are found in Africa, and the cobra cults of India are well known. Quetzalcoatl was the feathered serpent of the Aztecs. The serpent is a symbol for the snake-handling Protestant sects of the United States, the snake dancers of the Hopi Indians, and perhaps the Burmese snake charmers, who end their ceremonies with a kiss on the top of a cobra's head. Handling of snakes is done both as a gesture of belief and faith in the power of the gods and as an act of defiance of the same power. Bites and fatalities are about equally distributed in both cases.

NATURAL HISTORY

Interaction with the environment. While not exactly at the mercy of the environment, the snake is largely dependent upon it and must devote much of its day-to-day existence to coping with the vagaries and changes taking place in its surroundings. The problem varies with latitude and altitude, for the actions and reactions of a snake in temperate North America are distinct from those of one living in the American tropical lowlands but are similar to those of another living at higher altitudes in the Andes of Ecuador. No matter where they live, snakes are subjected to pressures from the biotic, or living, parts of the environment as well as from the abiotic, the physical, or nonliving, parts. But the amount or degree of challenge to the snake from different segments of the environment changes drastically depending upon the region it inhabits. The individual living in the hot, humid tropics of Africa, with comparatively constant temperatures close to optimum throughout the year and ample moisture from both rainfall and the surroundings, finds that its environmental problems involve competition with other members of its own species for food supply, the challenge of other species of snakes and perhaps other vertebrates for possession of the ecological niche, and the constant pressure of the predator that finds it a tasty morsel. The viper (*Vipera berus*) living within the Arctic Circle of Europe, on the other hand, is the only snake species present in the area and lives practically unchallenged in its niche but is faced permanently with situations in its physical environment that are substandard for survival, so that death from overheating, freezing, or dehydration is a repetitive threat. These differences among animals from different parts of the world are reflected in their life histories, and it is neither possible nor legitimate to speak of the "life history of the snake" unless one speaks only of a single region or, perhaps, of a single species.

In the tropical regions, life continues at approximately the same activity level year after year. The only break in the rhythm comes in the dry season and this only when the dry season is not just a period of slightly less rainfall. Snakes may enter a short period of dormancy at such times, which is, at least in part, a consequence of the effect the dry season has upon their prey. This dormant period is similar to the hibernation in winter by temperate-area snakes, although little is known about the physiological changes that may or may not take place in tropical dormancy. At higher latitudes and altitudes, during periods of maximum stress, which for most snakes are the cold months, the animals must seek out a place where they can be completely inactive and nonreactive, where their inability to respond to the stimulus of danger is compensated for by the absence of danger, and where the surrounding extremes of low temperature and low humidity are mitigated to the point that they remain within the limits of the tolerance capacity of snakes. Such places are few and far between, and good hibernacula (dens used for hibernation) are recognized and utilized year after year, with snakes of several different species often sharing a den. It seems likely that the snakes do not remember the location of a hibernaculum but utilize the simple expedient of following a scent trail left by other snakes randomly seeking a den, with the scent growing more and more strong as successive snakes arrive from greater and greater distances. Many of the changes that occur in the individual snake after arriving at the hibernaculum are direct results of its dependence upon the environment, and, as its body cools, the heartbeat and respiration slow almost to a stop and there is no muscular activity, little digestion, and no defecation. Physiological changes that are not correlated with or responsive to the surroundings also take place, but not to a degree comparable to those occurring in a hibernating mammal, and there is no "alarm system" to stir the snake into activity if a tolerance limit is passed. In such a case, the snake simply dies. At the end of the cold season, the snake is totally dependent upon the changes in its surroundings to bring it back to activity; it cannot rouse itself. The stimuli are felt by all almost simultaneously, and snakes emerge by the dozens or even by the hundreds from some denning places. In some species, copulation takes place immediately after enough of the sun's rays have been absorbed to permit the development of an interest in the surroundings; in others, copulation is the final act before entering hibernation, and the sperm lie dormant in the dormant female. Fertilization of the egg can take place immediately after copulation, but in some species at least, the female can store the sperm for several years, using them to fertilize successive batches of eggs.

Hibernation

Behaviour. *Interactions between individuals.* Snakes in both tropical and temperate regions tend to be solitary in their habits. The denning and mating aggregations are, for the most part, the only social events of the season. Sea snakes differ in this respect, sometimes being seen travelling in large troops, which seems to indicate an urge to aggregate. Female sea snakes (Hydrophiidae) also congregate in large numbers in seawall caves at parturition time, but this may have no social significance, since it seems to be a consequence of availability of a safe place for the young to be born rather than aggregational behaviour per se. There is some tendency for females of certain species in temperate areas to use a single site for egg deposition. Hunting of food is strictly an individual act for snakes; there are no known instances of cooperative hunting, as seen in some mammal and bird species. Hiding places and basking sites are occasionally shared; this again is a consequence of availability, and in the tropics, where hiding places abound, it is rare to find more than one snake at a time under a log or a rock. Except for these few weak instances, there is no development in snake populations of social behaviour. There is no establishment of social hierarchies, no territoriality, and perhaps no dominance. The combat dance engaged in by two males of the same species, where the bodies are entwined anteriorly and raised higher and

Combat dances

higher off the ground until finally one snake overthrows the other, certainly establishes a dominant individual, but there is no indication that awareness of this dominance is retained by either snake. A dominance that must be reestablished at every encounter does not contribute to a social structure. It has been suggested that the combat dance is essentially a homosexual encounter, with each male attempting to copulate with the other.

Reproduction. The occurrence of mating immediately after emergence from hibernation allows snakes to take advantage of the fact that the females are accessible, concentrated, and receptive. The males are equally concentrated, so pair formation and copulation is a simple matter. Males of some species have nuptial tubercles on various parts of the body, used to stroke or massage the female and, presumably, to arouse her sexually. Even when obvious tubercles are absent, the male uses a rubbing technique to stimulate the female, and in some species a muscle ripple moving along the male's body will provide a lateral caress. There are many descriptions in the literature of courtship dances done by snakes, in which the bodies are entwined and as much as one-third lifted off the ground, the coils ebbing and flowing with silent grace. Unfortunately, in many of these reports, the snakes were not captured and sexed, and the observer simply assumed that a male and female were involved. Recent work, where the snakes have been sexed, tends to indicate that the dance often, if not always, involves two males and is of little or no significance in reproduction. In any event, copulation is achieved after a comparatively brief courtship through the insertion of a hemipenis in the female's cloaca (a common urogenital chamber, lying just anterior to the anus). The hemipenis is one of a pair of mirror-image intromittent organs lying in the base of the male's tail, posterior to the anus, and strictly reserved for mating, for the urinary passages empty directly into the cloaca of the male. Either hemipenis can be used in copulation and must be everted, through a process of turning itself inside out. This is achieved primarily by engorgement of the organ with blood.

The everted organ is heavily armed with spines, spinules (minute spines), flounces, calyces, and other ornaments, all of which appear to play a role in ensuring that the male is securely attached to the female for the entire period until the sperm have been deposited. The sperm pass along a deep groove in the hemipenis, which, although open along one margin when examined in a dead snake, clearly forms a tubular passage as a result of the pressures of the engorged margins of the groove. After release, the sperm may immediately move up the oviducts and fertilize eggs just released from the ovary, or they may be stored by the female and released later to achieve fertilization. Once fertilization has occurred, the egg begins to accumulate additional layers from the shell glands in the oviduct. In some species, this continues until a firm yet pliable, leathery shell has been formed, permeable to both gases and liquids but capable of retaining much of its liquid content unless in a very dry place. The female then deposits the entire clutch of eggs in a protected, damp, warm, and usually dark place, often along with clutches from other females of the same species, for the same stimuli that lead snakes to congregate for hibernation also bring them to the same places for egg laying. Many species immediately abandon the eggs, some remain with the clutch and certainly appear to be protecting them from external danger, and a very few actually assume the role of a brood hen, maintaining a body temperature measurably higher than the surroundings, presumably assisting in incubation. In certain species, additional layers of membranous material are deposited around the embryo, but the calcareous (calcium-containing) shell does not form. Instead, the embryo is retained in the oviduct and continues its development there.

Egg laying

This is termed ovoviviparous development, since it is simply an egg retained in the oviduct, in contrast to viviparous, the condition seen in mammals, where the fetus develops in the uterus and establishes a placental connection with the uterine wall to permit exchange of materials with the maternal circulation. But, while an umbilical connection does not develop, there is now considerable evidence of exchange of materials between mother and fetus across their contiguous, highly vascularized membranous surfaces, and the distinction between the two types of development is so blurred in some species as to become specious.

Regardless of the devices used to provide it with protection, the snake fetus is always brought to term before the onslaught of environmental conditions that could result in its death. The embryonic turtle can sleep away its first winter in the egg and hatch the following spring none the worse for the experience, but there is as yet no evidence that snakes can do the same. The contrast may result from the fact that the female turtle can scoop out a hole deeply enough that freezing temperatures may not affect her brood, but the female snake is restricted, both by her limblessness and by the nature of the egg itself, to egg laying on or near the surface, where below-freezing temperatures are unavoidable. In the tropics, evidence is scanty, but it would appear that there is an endogenous (*i.e.*, controlled from within) rhythm there as well, since young are not produced throughout the year.

Early development and growth. The young snake, whether from an egg or born alive, comes equipped with a sharp cutting device on its upper lip, the egg tooth. It slashes its way out of the rubbery eggshell with this tooth or, in the case of the live-born, cuts its way out of the soft membranes and is instantly competent to cope with its surroundings. Almost invariably, the first act of a newborn snake is to extend its tongue and taste the surroundings, conveying to the Jacobson's organ (a sensitive region in the roof of the mouth) chemical information perhaps more significant than the visual cues picked up by the pair of very inexperienced eyes. Young snakes begin to feed immediately after hatching, displaying considerable ability in the capture and consumption of prey. Venomous snakes are born with functional venom glands and fangs and are capable of immediate utilization of their most formidable weapons. Some of the viperid snakes are born with a bright-green tail tip (contrasting strongly with the rest of the body colour), which they are capable of waving and shaking in a way that attracts the attention of possible prey. Within a very short time after birth, the first sloughing of the skin takes place, and the egg tooth is shed at about the same time.

The rate of growth is correlated with availability of food and temperatures high enough to permit full metabolic activity. When all factors are optimum, snakes grow surprisingly fast. A brood of California rosy boas born on October 26, 1965, more than doubled their lengths by July 26, 1966. In this nine-month period they had reached lengths only a few inches shorter than their mother, an adult close to maximum length for the species. It has been suggested that all snakes grow rapidly until they reach sexual maturity, after which time growth slows but very seldom stops completely. Snakes have indeterminate growth, which means there is no terminal point in time or size for growth in their lifetime, but they can continue to increase in length until they die. Sexual maturity is reached in about two years by many snakes. In the larger species, sexual maturity comes later, after four or five years or more.

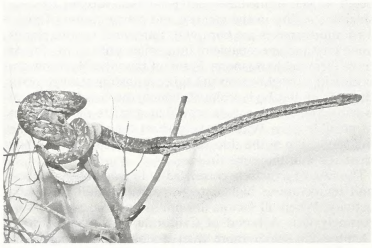
Growth rates

Molt. A regularly recurrent event during the activity period of all snakes is the shedding, or molting, of the skin. Dormant individuals do not shed, but quite often this is one of the first events to take place after the end of dormancy. The integument of all animals represents the primary buffer between internal structures and the environment, and it is constantly subject to wear, tear, and other damage. The first line of defense against damage, especially when the skin is completely broken, is the formation of a blood clot, a scab, cellular reorganization, and scar formation. The second line of defense is the constant production of new cells in the deeper layers of the skin to provide replacement cells on the surface for those lost or worn away. In snakes, the replacement procedure has been modified to a considerable degree. The replacement cells are not constantly produced independently of each other but are all on the same cycle and are cohesive into a complete unit. When this unit is functional, the old skin lying external to it becomes a threat to continued good

health. At this point, the snake's eyes become a milky blue, an indication of a physiological loosening of the skin that forms the eye cap. This loosening is duplicated all over the body, although not so obviously. Shortly, the eyes clear, and the snake rubs loose the skin around the mouth and nose and crawls out of it completely, leaving a new, functional skin resplendent in fresh, bright colours. The rattlesake sheds its skin in the same fashion as all other snakes, but the process is highly modified at the tail tip, where successive layers of keratinized, hardened epidermis are interlocked, or nested, to form the rattle, a device used to ward off death or injury to the snake caused by large mammals, including man. The use of the rattle is reasonably successful with buffalo, cattle, or horses but spectacularly unsuccessful with man, because, through this advertising, the rattling snake precipitates either its death or capture but seldom its escape.

Locomotion. The snake has overcome the handicap of absence of limbs by developing several different methods of locomotion, some of which are seen in other limbless animals, others being unique. The first method, called serpentine locomotion, is shared with almost all legless animals, such as some lizards, the caecilians, earthworms, and others. This is the way most snakes move and has been seen by any zoo visitor. The body assumes a position of a series of S-shaped horizontal loops, and each loop pushes against any resistance it can find in the environment, rocks, branches, twigs, dust, sand, pebbles, etc. The environment almost always provides sufficient resistance to make movement possible, and many snake species never use any other method of locomotion than this. Such

San Diego Zoo



In typical method of locomotion, Bahaman boa (*Epicrates striatus*) extending itself to reach branch at right (not shown).

species, when placed on a surface providing no resistance, such as smooth glass, are unable to move, whipping and thrashing around without progress. Other methods of terrestrial movement also involve at least some resistance by the environment but usually less than the first. One of these is known as "concertina" locomotion, because the snake in action resembles the opening and closing of an accordion or concertina. First the tail and the posteriormost part of the body are securely anchored, then the head and the rest of the body are extended as far forward as possible from that secure base. At the maximum extension, the head and the fore part of the body are anchored and the posterior part drawn up as close as possible in the accordion-like folds. A second cycle follows the first, and the snake progresses. Tree snakes, such as *Imantodes* and *Oxybelis,* modify this technique to move from branch to branch and have a strongly compressed body that permits surprising amounts of the body to be stiffened and extended, using a modified I-beam effect for rigidity. The third method is called "caterpillar" or "rectilinear" locomotion, because the body moves in a straight line, using a flow of muscle contractions along the sides that looks like a caterpillar in motion. The body musculature is used for sequential lifting, anchoring, and pushing against individual ventral scales, which results in an inching along. It is used by large, heavy-bodied snakes, such as the boas and some of the vipers.

The fourth method is the least dependent on friction with the substrate and, in fact, was developed because of the failure of the substrate to provide sufficient resistance for any of the previously described methods. This method, called sidewinding, characterizes snakes living in the desert, where the sand simply gives way under any kind of push. The sidewinder does not progress forward when in motion but actually goes sideways. The snake, lying extended on the sand, lifts the anterior part of the body, moves it five to six inches to the side, and rests that part on the sand, maintaining a lifted loop to the rest of the body. This lifted loop is then progressively shifted along the body to the end of the tail, at which time the entire snake has moved five to six inches to the side from its previous position. By the time the first loop reaches the end of the tail, a new loop has already lifted the head and started down the body, and the snake looks like a coiled spring rolling across the sand. There is no rolling involved, however, since the ventral surface is always in contact with a substrate and usually leaves its impression behind in the sand, like a footprint. A nonterrestrial method of locomotion is swimming. Most snakes use a somewhat frantic serpentine locomotion in water, and the water provides enough resistance to allow progress, but one would be hard put to say whether snakes are natural swimmers or cannot swim at all but simply take advantage of their natural buoyancy to crawl through the water. The true sea snakes, however, have a morphological adaptation that gives rise to a clearly distinct method of locomotion. The tail is compressed from side to side to form an oar, or paddle, and swimming efficiency is not only improved but raised almost to the level of an art by these snakes.

Sidewinding

FORM AND FUNCTION

General features. The most characteristic aspect of the snake form is the elongate body and tail and the absence of limbs. Fragments of limbs and girdles still survive in a few species of snakes, and in some snakes of the family Boidae the remnant forms a tiny, external spur on either side of the body at the anus, which is used by the male during courtship to stimulate the female. There is no snake in which the limb remnants still retain a function in locomotion. The body is usually slender, although there are some very heavily bodied species. The body shape is correlated with activity level, with the slender species moving about all the time and the heavy forms leading a sedentary life. The pit vipers, for example, while not always long snakes, are often big snakes. It seems likely that these snakes evolved in the direction of heaviness only after the development of a heat-sensitive depression, the loreal pit, located between the eye and the nostril, and the venom apparatus, which enabled them to stay in one place and wait for their prey to approach, rather than engaging in a continuous, active search for food. Similarly, some of the largest boid snakes (boas, anacondas, and pythons) have labial pits that function in the same way as the loreal pit of the vipers, so they, too, can be sedentary and grow fat. Arboreal snakes are the most elongated and slender of all, with the tail (the region posterior to the anus) often longer than the body and the body often strongly compressed laterally, which permits greater rigidity of the body frame while crawling from branch to branch. Burrowing snakes are seldom large, and the true burrowers, the Typhlopidae and Leptotyphlopidae, living all their lives like earthworms, are the tiniest snakes of all. The burrowers have almost no tail, although some of them retain a spiny tail tip, which probably serves the animal as an anchoring point when crawling through the soil. The tail in the sea snakes is flattened to form an oar, used to scull through the water. Sea snakes are almost totally helpless on land, locomoting only with the utmost of difficulty.

Body proportions

The skin. The entire body and tail in snakes are covered with scales, which are cornified folds in the epidermal layers of the skin. These scales are usually arranged in longitudinal rows, the numbers and arrangement of which are characteristic of the species. The scales may be large and shield shaped, in which case the number of rows is low, from eight to 29 or 30, or they may be very small, rounded, and occasionally with the centre raised, in which case the number of rows can be as high as 90 or 100. A sin-

gle scale may be very smooth and shiny (as in the rainbow snakes), it may have a raised ridge (keel) along its centre, it can be heavily striated, or it may even have a raised spine in the centre, as in the Javanese wart snake. The scales in some species have sensory structures on the posterior margins called apical pits, and all scales have various microornamentations, consisting of hairlike projections, holes, spinules (small spines), and other specializations visible only through an electron microscope. The scales on the ventral surface of the body are modified into broad plates in the majority of species and are used in locomotion. The sea snakes and the blind snakes lack such ventrals, the scales being small, as on the dorsal surface, and there are several other species, such as the anaconda, in which the ventrals are very reduced in size. This appears to be a secondary loss of the enlarged ventrals, correlated with life in an environment where locomotion is achieved in other ways, such as swimming or movement in burrows.

Coloration. The colours and colour patterns seen in snakes are often bright, occasionally spectacular, and in some species even beautiful. Snake colours are produced in two ways, either by pigment deposited in the skin or by differential diffraction of light as a consequence of the physical properties of the skin itself. When seen on a unicolour or uniform background, most snakes are obvious, and their colour patterns seem bold and prominent. When the animals are placed in their natural habitat, however, the significance of the colour patterns becomes obvious. The many lines running at sharp angles to the elongated lines of the body, the triangles or rectangles of colour, the blotches, spots, bands, or lozenges, all become highly disruptive to the eye, and the snake disappears into its surroundings. Blotched or spotted snakes tend to be sedentary and heavy bodied, while striped and the occasional unicolour snakes are usually active species. In both cases, the coloration is protective, since a coiled, sedentary snake has its body outline completely obscured by the overlapping patterns, while the stripes on a crawling snake eliminate the sensation of motion until they suddenly narrow at the tip of the tail and the snake disappears.

Warning coloration

Although in most snakes the colours are such that they help the animal to hide, there are some species that seem to be advertising their presence rather than trying to hide it. Their patterns are aposematic, or warning, in nature, and let a possible enemy or predator know that he runs some risk in an encounter with the snake. The warning is effective, of course, only if the intruder is knowledgeable concerning its significance and can take heed. This implies a teaching and learning sequence, with the dangerous snake as "teacher" and the predator as "student." For this reason, it has been suggested that the bright colours of the highly (and often fatally) venomous coral snakes did not evolve as warnings of the snakes' own poison but as mimics of some other venomous species, less dangerous, but still able to teach the predator the significance of the warning coloration. There is no evidence that avoidance of aposematic species is instinctive; on the contrary, naive predators readily attempt to take aposematic forms. A predator that dies in its first encounter with a dangerous species cannot act as a selective force favouring the coloration of that species. There are quite a few mildly poisonous rear-fanged snakes, brightly banded in red, black, and yellow (colours found in the coral snakes), that can make a predator suffer a sufficiently painful lesson that it will avoid contact with all similarly coloured snakes, including the fatally venomous coral snakes and the completely harmless milk snakes (*Lampropeltis*). (For a complete discussion of the evolution of mimicry, see MIMICRY.)

Skull and sense organs. Snakes rely on several senses to inform them of their surrounding. The pits, found in the region between the nostril and the eye in the pit vipers (the viperid subfamily Crotalinae) and in the scales of the lip line in some boids, are sensitive to very slight changes in temperature in their surroundings. These snakes feed almost exclusively on animals, such as birds and mammals, that maintain a constant body temperature and can therefore be located by the snake through the reception of the heat of the warm body. The heat lost by even a small rodent is sufficient to alert a waiting viper and enable it to direct a fast strike at the animal as it passes by. Death follows rapidly, and the snake follows the dying animal at a leisurely pace, perhaps in full awareness that it will not go far. The boids use the same technique for detecting warm prey, but after striking they retain the grip, killing by constriction.

Temperature detection

The eye of the snake is lidless and covered by a transparent cap of epidermis, which is shed with the rest of the skin at each molt. Animals active during the day usually have round pupils, while the nocturnal species have a vertical or slit pupil that opens up in the dark, as does that of a cat, but closes more effectively in bright light, protecting the sensitive, dark-adapted retina. The eye has been almost completely lost in the burrowing families, in which it is visible only as a black spot buried under the flesh and bone of the head. Arboreal snakes often have a bulging, laterally placed eye, permitting them to see activities directly below as well as above and around them. The structure of the eye in snakes indicates that their lizard ancestor was probably a burrower and that all aboveground activity by snakes is a secondary invasion from an ancestral life underground.

The reception of sound is entirely by bone conduction within the skull. The snake has no external ear, but it still retains a few of the vestiges of the internal ear, which are connected to other skull bones in such a way as to permit transmission of some earth-borne, and perhaps a few aerial, sound waves of low frequency.

The skull of snakes is characterized by mobility. It is light, with a reduced number of bones, and there are hinge joints at several levels that permit slight rotation or movement of one segment upon another. The only compact unit is the central braincase, with all other skull bones little more than attachments to it. The flexibility of the skull and the jawbones attached to it is a major compensatory factor to balance the loss of limbs. The snake replaces manipulation of digits with manoeuvrability of skull bones.

Teeth and fangs

The jaws of snakes are highly mobile and are usually heavily armed with teeth. The upper jaw can move to and fro on hinged joints and can also rotate slightly. In most snakes, the upper jaw is connected to the lower jaw by a joint that acts as a pivot point, and, in eating, all toothed bones on one side of the mouth move forward as a unit. In some tree snakes (the colubrid subfamilies Dipsadinae and Pareinae), the connection between the

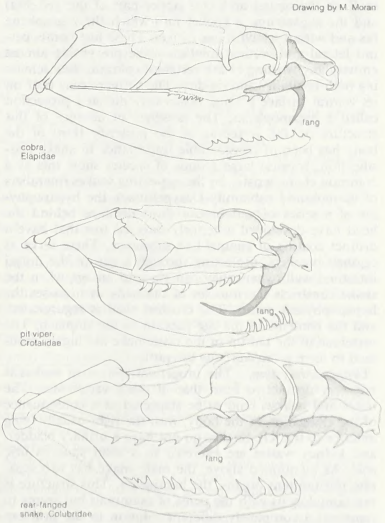


Figure 16: Skulls of representative poisonous snakes.

upper jaw and the quadrate is lost, and there are four independent units rather than two. These snakes feed exclusively on small, slippery prey, such as slugs and snails, and this jaw modification permits them to hold their food with three jaws while the fourth is advanced. The maxillary bone (the main bone of the upper jaw) of most snakes is elongated, with many teeth, but in the Viperidae (Old World vipers, New World rattlesnakes, and other pit vipers) only one functional fang remains, on a short, blunt, rotatable maxillary. The position usually occupied by the maxillary has been taken by the pterygoid bone. In the Elapidae (cobras and relatives) the maxillary bears a single fang in a fixed position, sometimes followed by a few smaller, solid teeth. In several different evolutionary lines of snakes, the posterior one or two teeth on the maxillary have enlarged and changed, usually in the direction of developing a groove or canal on the anterior edge to conduct a flow of venom. These are the rear-fanged snakes, usually (although not always) nonlethal to man. Snake teeth are usually long, slightly recurved, and needle-sharp. This facilitates swallowing and prevents loss of food, because the only direction in which a food item, which may be alive when swallowed, can go to escape the teeth is down the throat. Modifications in food habits have often been accompanied by changes in tooth structure or the loss of teeth. The egg-eating snakes, for example, have only a few peglike teeth left. The burrowing blind snakes have very reduced dentition and have often lost the teeth of one jaw entirely.

The vertebral column of snakes is highly elongated and includes many more vertebrae than almost any other kind of vertebrated animal. Since there are no limb girdles associated with the skeleton, there are no good delimiters of regions, and snakes are generally regarded as having only two kinds of vertebrae, precaudal and caudal. A pair of ribs is associated with each precaudal vertebra except for a few immediately behind the head, but, by definition, there are no ribs on caudal vertebrae. Each vertebra articulates with its neighbours at five different points; first, at the contact point between the centra (the main, central bodies of the bones), which is a ball-and-socket joint; then two points at the zygapophyses (projections from the centra), with articulating surfaces that lie dorsally and ventrally; and finally the zygosphenes and zygantra, found almost exclusively in snakes, the zygosphene being a projecting shelf on the neural arch (the upper part of the vertebra) and the zygantrum, a pocket into which the zygosphene fits and within which it can swivel. These five points permit lateral and vertical rotation while preventing almost entirely any twisting of the vertebral column, thus achieving both flexibility and rigidity. The vertebra may bear on its ventral surface a long, posteriorly directed projection called a hypapophysis. The presence or absence of this structure on the vertebrae of the posterior third of the body has been of considerable importance in snake classification, because large groups of species show this as a common characteristic. In the egg-eating snakes (members of the colubrid subfamily Dasypeltinae), the hypapophyses of a series of vertebrae a short distance behind the head have developed anteriorly directed tips that have a distinct coating of enamel-like substance. These serve as eggshell breakers, projecting through a gap in the dorsal intestinal wall, where they can rip into an egg when the snake constricts the muscles of the body as it passes the hypapophyseal points. The crushed shell is regurgitated, and the contents of the egg pass on to the stomach. The vertebrae of the tail tip in the rattlesnake are highly modified to form a "shaker" for the rattle.

The hypapophysis

Urogenital system. The urogenital system in snakes is not very distinctive from that of other vertebrates. The testes and ovaries tend to be staggered as a consequence of the elongation of the body, with the right usually lying anterior to the left. Snakes do not have a urinary bladder, and kidney wastes are excreted in a solid state as uric acid. As mentioned above, the male snake has two separate intromittent organs, the hemipenes. This structure is not homologous with the penis of mammals but seems to represent a completely different solution to the problem of internal fertilization. It is a saclike structure that must be turned inside out to be inserted in the cloaca of the female and can be removed only by turning it back to inside in, because to draw it out directly would damage the female considerably. The hemipenis is extremely variable in its overall appearance and structure; the cloaca of the female is often similarly constructed, thus preventing cross-fertilization by males of related species.

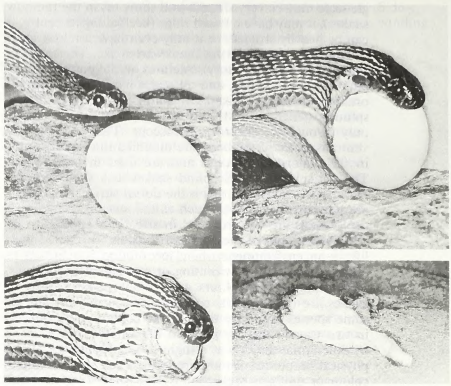


Figure 18: (Top left, top right, bottom left) Egg-eating snake (*Dasypeltis*) approaching and swallowing egg of a hen. (Bottom right) Eggshell that has been regurgitated after snake has consumed its contents.
Carl Gans

Specializations for securing food. The most essential and time-consuming activity for a snake during nondormant periods, regardless of its habitat, is the pursuit, capture, and digestion of food. There are many morphological and behavioral modifications to be observed in snakes that facilitate food gathering. Some of these changes are widespread and found in practically every kind of snake. Jacobson's organ, for example, is located in the roof of the mouth and is capable of detecting minute quantities of various chemical substances when they are picked up externally on the delicate, double-tipped tongue and thrust into the organ for analysis. It is so significant and useful that it has never been lost by any snake species. While important primarily in trailing and recognition of prey, the Jacobson's organ is also used in detection of enemies, in trailing other snakes of the same species, and in courtship.

Some snakes have specialized salivary glands that elaborate a potent poison, along with either grooved or tubular teeth to permit internal injection of the venom. This device for rapid immobilization of prey has proven successful whenever it has occurred, and several different lines of evolution can be detected both in the type of injector tooth and in the chemical composition or mode of action of the venoms. The types of injector tooth include the long, tubular fang of the viperids; the short, hollow fang on a fixed maxillary seen in the elapids; and the several kinds of grooved tooth on the posterior end of the maxillary. In the latter case, there is little doubt but that the development of grooving has taken place several different times in different places. As for the venoms, it is true that the terms neurotoxic, hemotoxic, and cardiotoxic, once used to describe the different effects, are clearly misleading and are too simple for accurate statements concerning venom composition, but it is still accurate to say that some of the components of venoms cause changes in the red blood cells, coagulation defects, and blood-vessel injury, while others produce deleterious changes in sensory and motor functions and in respiration, and still others have a direct effect on the heart. Some venoms kill very rapidly, such as that of *Bothrops insularis* of Queimada Island, off the Brazilian coast, which would lose as prey most of the birds it bites if they could fly very far away. Other venoms kill

Types of venom

more slowly, and the snake bites, retires, and waits, finally trailing the bitten prey, using the tongue and the Jacobson's organ until it finds the already stiffening body. Some snakes, in particular the rear-fanged species, bite, chew, and hold on, eventually bringing the hindmost maxillary teeth into play, which permits the injection of toxins.

Constriction

Other feeding specializations are not so widespread among species, and some are restricted to a single group. Many of the boids and some of the colubrids utilize a constriction method for killing their food. The prey is struck and held by the teeth, and a series of body coils are rapidly thrown around it. These coils tighten until respiration is impossible and suffocation results, but very seldom are bones crushed or broken. The snail-eating snakes have a series of modifications of the teeth, the toothed bones, and the lower jaw that permit the insertion of the lower jaw into a shell to pull out the snail's body. One genus of sea snakes, *Microcephalophis,* has a tiny head and a long neck with the same diameter as the head, which can be inserted deeply into very narrow holes inhabited by its prey. An Asiatic water snake, *Erpeton,* has a pair of elongate tentacles on its snout that appear to function as sighting devices to guide the strange, stiff-bodied strike at a passing fish. There is a great correlation between the difficulty in catching a particular kind of prey and the development of morphological and behavioral devices to help solve the problem. The blind snake living in a termite nest needs no more than its tongue and Jacobson's organ to permit it to recognize the soft-bodied, defenseless termites, and its eyes and most of its teeth have been lost. But the rattlesnake, seeking the most elusive prey of all in the agile, aware, sensorily acute small mammal, has a full armament of equipment, including the facial pit, the venom gland, the rotatable maxillary bone, the tubular fang, and two functional eyes.

EVOLUTION AND CLASSIFICATION

Evolutionary origins. The oldest fossil definitely identifiable as a snake is *Dinilysia,* from the Upper Cretaceous (about 80,000,000 years ago) of Argentina, considered to be most closely related to the aniliids and others in the superfamily Booidea. Some fossils from earlier deposits have been called snakes, but American paleontologist Alfred S. Romer has tentatively regarded them as lizards. Snake bones are delicate and do not tend to fossilize well. Most snakes are small, and fossil hunters in the past tended to concentrate on larger, more spectacular specimens. Modern techniques have begun to turn up a wealth of material, much of very small size, including snake skull bones and vertebrae. Much of this material is from Pliocene, Pleistocene, and older Recent deposits, and it often either is identifiable with modern genera or else shows close resemblance to them.

Relation to lizards

Snakes are, thus, a comparatively young group of animals. They arose from a lizard ancestor probably very similar to the modern varanids (monitor lizards). *Dinilysia* shows certain characters very similar to those found in the Bornean lizard *Lanthanotus,* a platynotan lizard related to the Varanidae. Several theories have been proposed as to the habitat preference of the snake ancestor, including aquatic, terrestrial in savanna-like situations, or burrowing. Strong support for the burrowing hypothesis has come from study of the eye. Losses (such as the movable eyelid) caused by ancestral adaptation to burrowing were not regained by the descendant nonburrowing snakes, and other devices to cope with vision problems were developed and refined.

The phenomenal adaptive radiation that produced the many different species of colubrid snakes (a category that includes the vast majority of familiar snakes), now found in practically every kind of habitat and efficiently capitalizing upon almost every kind of terrestrial animal as food, was a very recent event. It probably took place entirely within the Tertiary (from about 65,000,000 to 2,500,000 years ago), and represents a triumph over the handicap of limb loss, which is often a blind alley in evolution. The development of more than 2,000 species of snakes occupying practically all available niches is an achievement equivalent to the concomitant emergence and differentiation of the birds or perhaps even that of the mammals.

Distinguishing taxonomic features. Herpetologists have used a variety of characteristics in the recognition of higher taxonomic units within the snakes, but one tendency has characterized most of the classifications set up by different authors over the years. This has been to base the classification on contrasting conditions of a very few characteristics, occasionally only one. Characteristics used either by themselves or in combinations with one or two others have included the following: the manifold variations in the structure of the hemipenis; the presence or absence of various projections, knobs, folds, shelves, and other modifications of the individual vertebrae; the different types of maxillary teeth and their arrangement on the bone; arrangement and relationships of bones in the skull; presence or absence, location, origin, and insertion of muscles, primarily cranial, but occasionally those of the body; and the type of retina in the eye. If any one of these is emphasized over most others in a classification, that classification will be very different from other systems, because these characteristics are not closely associated genetically and do not tend to group themselves together but vary quite independently. Taxa at the lower levels of classification (genera and species) are usually recognized on the basis of external characteristics that permit easy identification, even though these cannot be used to indicate phylogenetic relationships. These characteristics include the shape, presence or absence, fusion, division, and relationships of the plates of the head; the numbers of rows of dorsal scales and their ornamentation with keels, striae, pits, etc.; the size, number, and condition of ventral, anal, and subcaudal scales; shape of body or head or both; and, in some cases, coloration.

Annotated classification. The classification presented here represents a conservative summary of the views expressed by G. Underwood, H.W. Parker, H.G. Dowling, A.S. Romer, A. Bellairs, and other prominent herpetologists in the recent literature. Careful analysis of the classifications offered by these authors shows that the major differences are in the level of taxonomic recognition given to a particular group rather than in the allocation of a group with regard to phylogenetic position.

SUBORDER SERPENTES

Reptiles without limbs and with the body greatly elongated; with a very high number of vertebrae in both body and tail, each individual vertebra with zygosphene and zygantrum and almost all those of the body with a pair of attached ribs. Pectoral girdle completely absent; pelvic girdle usually absent, but present as a vestige in primitive species. Outer ear and tympanum lacking, sound reception entirely by bone conduction. Bones of both jaws usually flexible, loosely attached to the skull and to each other, plentifully supplied with teeth in most species, forming poison fangs in some. The braincase is platytrabic.

Superfamily Typhlopoidea (or Scolecophidia)

Diminutive, wormlike snakes that usually retain vestiges of the pelvic girdle and 2 common carotid arteries. Skull compact, with the bones solidly united, which facilitates burrowing; dentition very reduced, with no teeth on the pterygoid, premaxillary, or palatine bones. Intercostal arteries to almost all body segments.

Family Typhlopidae (blind snakes or worm snakes)

Maxillary bone bears teeth arranged transversely and can be rotated slightly on its attachment to the skull; dentary bone usually lacks teeth (present in anomalepine species). Either there is a single pelvic bone or the girdle is absent entirely. Circumtropical in distribution; practically no fossils are known, except for a few specimens from the Tertiary of Europe.

Family Leptotyphlopidae (slender blind snakes)

Maxillary bone bears no teeth and is quite firmly united with the skull; teeth are always present on the dentary. The pelvic girdle made of 3 bones. Tropics of Latin America to southwestern United States, Africa, and southwestern Asia. No fossils are known.

Superfamily Booidea (or Henophidia)

Moderate-sized to very large snakes. Often with vestiges of the pelvis and associated bones, sometimes visible externally as a rudimentary hind limb. There are two common carotid arteries. The coronoid bone of skull usually present; parasphenoid bone does not form part of the optic foramen; supraorbital bone absent in all except some members of the Boidae. Den-

tition well developed on all jawbones, usually including the premaxillary. Intercostal arteries running to almost all body segments are present.

Family Aniliidae

The femur present as an anal spur in males, and pelvis made up of a single bone. No hypapophyses on the body vertebrae. Ventral scales only slightly larger than dorsal scales, not extending across the full width of the belly. Premaxillary bone with teeth; forms a suture (tight seam) with the maxillary; dentary immovable. One genus, *Anilius;* known only from the Recent of South America; a brightly coloured snake, with red, black, and yellow rings around the body.

Family Xenopeltidae (sunbeam snakes)

Femur absent; no trace of the pelvic girdle. Hypapophyses on the anterior vertebrae. The ventral scales enlarged. Premaxillary bone toothed; forms a suture with the maxillary. Dentary hinged on the surangular bone and movable. One genus, *Xenopeltis,* burrowing snakes from India and elsewhere in Asia. No fossils known.

Family Uropeltidae (shieldtail snakes)

Femur absent; there may or may not be pelvic rudiments. No hypapophyses on the vertebrae. Ventral scales usually not more than twice as broad as the neighbouring dorsal scales. No teeth on the premaxillary, which forms a suture with the maxillary. Dentary fixed on the surangular and, thus, immovable. A family of burrowing snakes, found only in southern Asia. No fossils known.

Family Boidae (pythons, boas, and wood snakes)

Femur often present; pelvic vestiges occuring in all but a few genera. Hypapophyses present on vertebrae. Ventral scales broad, extending across the entire belly, except in a very few groups. Premaxillary may or may not bear teeth; does not form a suture with the maxillary. Circumtropical; it includes all of the giant snakes as well as several genera of small, secretive snakes. Bones from apparent boids have been found in Tertiary deposits around the world. The family is divided into subfamilies in different ways by different authors. The subfamilies Boinae, for the large New World constrictors, and the Pythoninae, for the Old World pythons, are almost universally recognized.

Family Acrochordidae (wart snakes)

Femur and all pelvic vestiges absent; hypapophyses present throughout the vertebral column. Ventral scales cannot be differentiated from dorsals. Premaxillary bone toothless and free of maxillary bone; dentary not hinged or any part of it free. One genus, *Acrochordus;* aquatic snakes found in Australia, the East Indies, and southeastern Asia. No fossils known.

Superfamily Colubroidea (or Caenophidia)

This superfamily includes the major number of snakes, which have in common the complete absence of any vestiges of the limb girdles and have only the left common carotid artery. Skull quite flexible, with several joints and sutures that permit movement of one part on another. Coronoid bone absent. No teeth on the premaxillary; characteristically, maxillary, palatine, pterygoid, and dentary bones all are toothed; some groups have a fang, or grooved tooth, at the posterior end of the maxillary. Intercostal arteries occur at intervals of several body segments.

Family Colubridae

Maxillary bone comparatively fixed in position, cannot be rotated; usually elongate and armed with teeth throughout its length. Posteriormost maxillary teeth grooved or hollow in some groups, but these fangs are preceded by a series of normal teeth in practically every instance. Hypapophyses may be present on vertebrae throughout the body or may be absent posteriorly, and this character has been used to divide the family into smaller groups. Tail usually long and ending in a point. Found all over the world where snakes occur; species make up the majority of snakes everywhere except Australia. Most of the following subfamilies are recognized by herpetologists, and some authors raise several of them to familiar status.

Subfamily Dasypeltinae (egg-eating snakes). Perforated esophagus and a series of elongated, enamel-tipped vertebral hypapophyses, which pass through the perforations and can crush eggshells. Teeth extremely reduced and peglike. Two genera, found in Africa and Asia.

Subfamilies Pareinae and Dipsadinae. Two small subfamilies of arboreal snakes, specialized for snail eating. These adaptations, including specially modified lower jaw, have evolved separately in the Asian Pareinae and the South American Dipsadinae.

Subfamily Xenodontinae. Large, bladelike posterior maxillary teeth, a large adrenal gland, and several other characters. Two of the genera, *Lystrophis* and *Heterodon,* have pointed, recurved, digging snouts not seen in *Xenodon.*

Subfamily Homalopsinae. Aquatic, fish-eating snakes, lacking grooved teeth. Stout body with a short tail. Southeast Asia, the East Indies, and Australia.

Subfamily Colubrinae. Widespread and often abundant group; contains most of the familiar harmless snakes of Europe and North America. Highly variable in coloration, size, and structure. Many nonpoisonous; some rear-fanged (but rarely dangerous to man); many kill by constriction. About 1,500 species.

Family Viperidae (vipers, rattlesnakes, moccasins, and relatives)

The maxillary bone rides on a hinged joint and is armed with a long, sharp tooth with a canal through it for injection of poison; no other teeth on the maxillary except for replacement fangs. Hypapophyses present on the vertebrae throughout the body.

Subfamily Viperinae (Old World vipers). No heat-sensitive pit or excavation in the maxillary bone. Found throughout Europe, Asia, and Africa but not in Australia. Fossils from Miocene of Europe.

Subfamily Crotalinae (pit vipers). Heat- or infrared-sensitive pit present on the side of the head between the nostril and the eye; maxillary bone with deep excavation in its outer surface to accommodate the pit organ. Widely distributed in the New World; one of the larger groups includes rattlesnakes. A few species also known from Asia. Fossils from Pliocene and Pleistocene of North America.

Family Elapidae (cobras, mambas, coral snakes and relatives)

Maxillary bone is shortened, fixed in position; cannot be rotated. Anteriormost tooth on the maxillary a short, fixed, hollow poison fang, usually but not always followed by a few other teeth. Hypapophyses are present on all body vertebrae. Tail rounded and normal, ending in a narrow tip. Elapids are also found abundantly in Asia, Africa, Central America, and South America; especially numerous in Australia, comprising, with sea snakes, more than two-thirds of snake species; rarely enter the United States; do not occur in Europe except as possible fossils.

Family Hydrophiidae (sea snakes)

The maxillary bone much like that of the elapids. Hypapophyses are present on the vertebrae throughout the body. The tail and part of the body compressed vertically into a large, oarlike swimming device. Found in the tropical western Pacific and Indian oceans; only one genus proven to have reached the Pacific coast of the New World.

Critical appraisal. The classification of snakes cannot yet be considered to have achieved the relative stability that characterizes bird and mammal classification. Authorities continue to disagree both on the taxonomic rank to be accorded certain groups and on the relationships of many groups. Romer, for example, recognizes the Xenopeltinae and Loxoceminae as subfamilies of the Aniliidae. Underwood, on the other hand, treats the first as a family, Xenopeltidae, and considers the Loxoceminae to be a subfamily of the Boidae. The Anomalepinae are regarded by some authors as sufficiently distinct from the Typhlopinae to warrant family status.

Emphasis on one characteristic, such as the presence or absence of a pelvis, profoundly affects the classification. Some authorities remove from the Uropeltidae all genera with pelvic remnants, placing them in the Aniliidae. Others have excluded from the Boidae all genera that lack pelvic remnants.

The family Colubridae has troubled taxonomists for nearly two centuries, many attempts having been made to break it into workable groups. The conservative view still recognizes only a few small, distinct subfamilies and several large grab-bag subfamilies. Many groups have been separated from the Colubrinae as subfamilies within the Colubridae or even as distinct families.

In older classifications, the pit vipers have sometimes been separated from the Old World vipers as the family Crotalidae. (J.A.P.)

Crocodilia (crocodiles)

The crocodiles constitute an order of the vertebrate class Reptilia. They are generally large, ponderous, amphibious animals, somewhat lizardlike in appearance, and carnivorous in habit. They have powerful jaws with many conical teeth and short legs with clawed, webbed toes. The tail is long and massive and the skin thick and plated. About 20 species are recognized.

The group is of particular interest because of its evolutionary position: the crocodiles are the last living link with the dinosaur-like reptiles of prehistoric times. They are, at the same time, the nearest living relatives of the birds. A large variety of crocodile fossils have been discovered; three of the four suborders of Crocodilia are extinct. On the basis of this extensive fossil record, it has been possible to establish well-defined relationships between the crocodiles and other vertebrate groups.

GENERAL FEATURES

Size range and diversity of structure. The crocodiles are the largest and the heaviest of present-day reptiles. In former times the Nile crocodile (*Crocodilus niloticus*) and the estuarine crocodile (*Crocodilus porosus*) attained a length of almost nine metres (about 30 feet), but today, specimens rarely exceed six metres (20 feet). Other species, for example, the smooth-fronted caiman (*Paleosuchus*) and the dwarf crocodile (*Osteolaemus tetraspis*) are about 1.7 metres (six feet) in length.

All crocodiles have a relatively long snout, or muzzle, which varies considerably in proportions and shape. The large horny plates that cover most of the body generally are arranged in a regular pattern. Thick, bony plates occur on the back. The families and genera may be distinguished by anatomical features, principally those of the skull. Species are identified principally by the proportions of the snout; by the bony structures on the dorsal, or upper, surface of the snout; and by the number and the arrangement of the large neck scutes.

From Mitteilungen aus dem Zoologischen Museum in Berlin

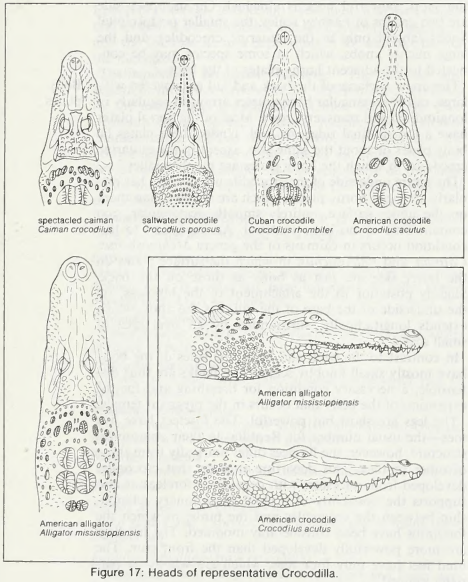


Figure 17: Heads of representative Crocodilia.

Distribution and abundance. The habitat of the crocodile is mainly the tropics and subtropics of the northern and southern hemispheres. The Mississippi, or American, alligator (*Alligator mississippiensis*), the American crocodile (*Crocodylus acutus*), and the Chinese alligator (*Alligator sinensis*) are the only species found outside the tropics.

The true crocodiles (family Crocodylidae) occur in most of Africa south of the Sahara, Madagascar, India, Sri Lanka, Southeast Asia, the East Indies, northern Australia, Mexico and Central America, the West Indies, and most of South America east of the Andes and north of the mouth of the Río de la Plata. The caimans are confined to South America. The gavial occurs in India.

Declining populations

As recently as several decades ago, crocodiles were plentiful in much of the tropics and subtropics, but today they are almost extinct in many areas, as the inroads of civilization continue to limit their natural habitat. Thousands are killed annually by man, either for sport or for their valuable skins, which provide leather for handbags, luggage, shoes, belts, and other articles. In some localities they are killed because of their depredation of domestic cattle. The governments of some countries have sought by legislation to protect the remaining crocodile populations from hunters.

The disappearance of the Nile crocodile from parts of Africa has resulted in an overabundance of the catfish *Clarias*, which in turn has greatly diminished the supply of popular food fishes. In an attempt to restore the original balance, crocodiles are bred and raised on farms. When they have grown to a length of about 1.5 metres (five feet) the crocodiles are released to their natural habitat.

NATURAL HISTORY

Life cycle. The young crocodile emerges from the egg with a length of 20–25 centimetres (eight to 10 inches). At first it remains concealed at the edge of its water habitat in order to avoid various predators. Principal among these are fishes and birds, but larger crocodiles also prey upon the young. During the first three to four years, the young increase in length by about 30 centimetres (about one foot) per year. The growth rate then gradually decreases, but growth can continue throughout life. Sexual maturity occurs at about ten years of age and at a body length of about 1.5 metres (five feet).

Of the little information available on the longevity of crocodiles, most has been gained from observation of animals in captivity. Captive animals seldom live more than 40 years, possibly because of lack of proper exercise and diet. Statistics on the growth of the Nile crocodile suggest, however, that specimens about six metres (20 feet) long are probably much more than 100 years old.

Behaviour. Crocodiles are predators, mostly nocturnal (*i.e.*, active at night), and spend most of their time in the water; they are also known to take rather long journeys over land. In their first weeks of life crocodiles eat mostly worms and water insects, then frogs and tadpoles; finally, their main diet is fish. Older crocodiles are more apt to prey upon waterfowl and on mammals, and occasionally a member of one of the larger species eats a human. This happens so infrequently, however, that crocodiles cannot be generally regarded as maneaters.

Diet

Crocodiles capture water animals in their jaws with a sideways movement of the muzzle. To catch land animals they remain motionless at the edge of a water hole from which the prey habitually drink, or they float passively in the water, resembling a drifting log. With a swift blow of the tail, they knock unsuspecting prey into the water. A number of crocodile species grip the legs of the victim in their jaws, then rotate themselves rapidly in the water, thus tearing the prey apart. When a crocodile cannot consume all of a victim at one time, it drags the carcass into its burrow.

The burrow is dug more or less horizontally at or just above the waterline and may extend for several metres, eventually ending in a chamber. The Mississippi alligator and the China alligator enter a state of inactivity (*i.e.*, hibernate) in these chambers during cold periods.

During the day, crocodiles often lie at the water's edge to sun themselves, frequently in large numbers. Otherwise, crocodiles live as lone individuals and establish individual territories. The extent of the territory is apparently defined for the animal's neighbours by its loud, vibrant roar, answered in kind by the crocodiles in adjacent territories.

When it roars, the crocodile tenses the musculature of its body so that the head and tail rise high out of the water; the flanks may vibrate so violently that water is sprayed high into the air from each side. Roaring can be provoked by similar noises, such as a deep note from a trumpet or the sound of artillery or supersonic aircraft. Crocodiles are also capable of deep grunting sounds, which apparently

play a role during courtship, and a warning hiss. The young make squeaking sounds.

Intelligence

Of all reptile brains, the crocodile brain is the most highly developed. Crocodiles often show curiosity and are capable of being developed into tame animals—both attributes being measures of intelligence. If kept in captivity from birth, individuals of some species are known to recognize their keepers, show neither fear nor aggressiveness, beg for food, and permit themselves to be petted. In other words, they learn to adapt their behaviour to the unnatural captive environment.

Locomotion. The principal style of locomotion is that of swimming, in which the crocodile places its legs back against the sides of the body and moves forward by means of lateral wavelike motions of the tail. In walking on land, crocodiles hold themselves high on all four legs, and the body moves in waves, with a lateral swinging motion. When startled, crocodiles are able to run short distances, the two front legs moving forward and backward together as do also the two hindlegs, somewhat in the manner of a hopping rabbit. When moving quickly into the water from a bank, crocodiles slide on their bellies, pushing with the feet.

Reproduction. The sexes are outwardly different only in the Indian gavial (*Gavialis gangeticus*), in which the males, during the reproductive period, have a bulbous knob at the tip of the muzzle. Copulation occurs in the water and lasts about ten minutes. It is preceded by a courtship in which the animals rub their muzzles against each other and over the neck of the partner. The male then mounts the back of the female, and both animals rotate their tails so that the respective bodily openings are brought into contact. The distensible male reproductive organ is then inserted into the female.

All crocodiles lay hard-shelled eggs, which may number more than 100, depending upon the age and size of the female. The female builds a nest as a shelter for the eggs. The Nile crocodile female digs a trench, which it refills with dirt after laying the eggs. The female estuarine crocodile builds a mound of mud and decaying plant material, in the centre of which are the eggs. With her tail the female splashes water onto the nest. This promotes the heat-generating process of vegetative decay. The trench and mound types of nest are extremes, between which intermediate types are made by the various other species. In every case the female remains close to the nest and protects the eggs from predators until the eggs hatch.

The nest

After two or three months the young are fully developed and ready to hatch. While still in the egg they utter squeaking sounds, perhaps signalling that they are ready to emerge. The female then removes the dirt or other debris from the eggs but thereafter provides no further care for her offspring.

Ecology. Crocodiles are mainly inhabitants of swamps, lakes, and rivers, although some species make their way to brackish water or to the sea. The estuarine (or saltwater) crocodile (*C. porosus*), which lives almost entirely in the ocean, may swim miles out to sea. Such crocodiles venture upriver only as far as the limit of the salt water. The smooth-fronted caiman (*Paleosuchus*) in South America prefers rocky, fast-flowing rivers. In West Africa the dwarf crocodile (*Osteolaemus tetraspis*) is found principally in the rivers of the forest regions.

FORM AND FUNCTION

The form of the crocodile is adapted to its amphibious way of life. The elongated body with its long, muscular paddle tail is well suited to rapid swimming. Other features, enumerated below, are also adapted to the animal's amphibious habit.

Sensory structures

The external nostril openings, the eyes, and the ear openings are the highest parts of the upper side of the head. These important sense organs thus remain above the water surface even when the rest of the head is submerged. The two nostril openings are close together on a raised portion of the point of the muzzle and may be closed by membranous flaps so that no water can enter when the animal dives. A long nasal passage enclosed in bone leads from the exterior nostril openings to the interior nostril openings, or choanae, located at the extreme posterior end of the palate; a membranous flap in front of the choanae constitutes the posterior closure of the mouth cavity. Thus, the crocodile can breathe even if its mouth is open under water.

Like many nocturnal animals, crocodiles have eyes with vertical, slit-shaped pupils; these narrow in bright light and widen in darkness, thus controlling the amount of light that enters. On the back wall of the eye the so-called tapetum lucidum reflects the incoming light, thus utilizing small amounts of light to the best advantage. In addition to the protection provided by the upper and lower eyelids, the nictitating membrane, a thin, translucent structure, may be drawn over the eye from the inner corner while the lids are open. The delicate eyeball surface is thus protected under water, but a certain degree of vision is still possible.

Unlike the ears of other modern Reptilia, those of the crocodile have a movable, external membranous flap. By means of this structure the crocodile is able to protect its ears from the water.

The crocodile's relatively flat snout is usually quite long; in some species it is extremely elongated. The outer margin of the jaws in most species is jagged. Each jaw carries a row of sharp teeth, which may number more than 100 in species with very long muzzles. The teeth are held in sockets and replaced continuously, new ones growing from below and old ones being forced out. On the floor of the mouth is the thick, fleshy tongue, firmly attached and therefore almost immobile.

The posterior portion of the head forms a flat plate. Here the short, powerful neck is attached. On its upper side are two groups of knobby scales: the smaller postoccipital knobs (absent only in the estuarine crocodile); and the large nuchal knobs, which in some species may be connected to the adjacent horny plates of the back.

Body plates

The upper surfaces of the back and tail are covered with large, rather rectangular horny plates arranged regularly in longitudinal and transverse rows. Most of the dorsal plates have a longitudinal ridge, or keel. Under these plates lie bony plates of about the same size, except in the estuarine crocodile, in which the bony plates are much smaller.

The entire underside of the crocodile usually also has regularly arranged horny plates, which are smaller than those on the upper surface, entirely smooth, rectangular, and contain little or no bone material. An exception to this condition occurs in caimans of the genera *Melanosuchus, Caiman,* and *Paleosuchus,* in which the surface plates on the lower side are just as bony as those on the back. Slightly posterior to the attachment of the hindlegs, on the underside of the base of the tail, lies the anus, which extends longitudinally, surrounded by an oval area of small scales.

In contrast to the back and belly, the sides of the body have mostly small knobby scales. The flanks are thus distensible, a necessary condition for breathing and for the expansion of the body that occurs in the pregnant female.

The legs are short but powerful. The forefeet have five toes—the usual number for Reptilia. In their anatomical structure, however, the forelegs differ markedly from those of other reptiles. This departure suggests that crocodiles developed from ancestors with degenerate forelegs; it also supports the possibility of a close evolutionary relationship between the crocodiles and the birds, in which the forelimbs have been considerably modified. The hindlegs are more powerfully developed than the front pair. The hind feet have only four toes, which are wholly or partially webbed.

On its flat upper side the tail carries a comb which, in the anterior portion, is double, consisting of high leaf-shaped scales. Near the middle of the tail the two combs merge.

The heart of the crocodile is markedly different in structure from that of other reptiles: the two auricles and the two ventricles are completely separate.

Arterial-venous connection

There exists, nevertheless, a connection between the arterial and venous circulation by way of the so-called Panizzae foramen, which opens between the two vessels leading separately from the ventricles. This connection is necessary for equalization of pressure differences between

the arterial and venous circulatory systems that arise during long dives in deep water.

EVOLUTION AND CLASSIFICATION

Paleontology. The crocodile skull exhibits distinctly developed upper and lower temporal (*i.e.*, behind the eye sockets) openings; the teeth arise from sockets, and the roof of the skull lacks an opening for the parietal organ, a median, dorsal outgrowth of the brain. The crocodiles thus show the most important characteristics of the group that includes the dinosaurs (Archosauria). Within the Archosauria the crocodiles are a separate order, since they have developed a secondary bony palate, which encloses the nasal passage from the exterior nasal openings to the choanae (internal nostrils).

These features occur even in the most primitive representatives of the crocodile group, namely the Protosuchia of the Upper Triassic Period (about 190,000,000–200,000,000 years ago); but their muzzles were very short, and the choanae were relatively far forward on the palate. As the crocodiles continued to evolve, the openings of the choanae tended to move further back. In the Mesosuchia of the Jurassic (136,000,000–190,000,000 years ago) and Cretaceous periods (65,000,000–136,000,000 years ago)—to which the long-snouted ocean crocodiles also belong—the choanae had already moved to the posterior part of two bones of the skull (palatines). In the true crocodiles (Eusuchia), which appear in the Upper Jurassic, the choanae are entirely enclosed by the pterygoids. In modern species they have moved to the posterior border of the palate. In the Sebecosuchia of the Upper Cretaceous to the Miocene Epoch (7,000,000–26,000,000 years ago), a branch collateral to that of the crocodiles, the skull is laterally flattened, and the choanae lie in a depression in the anterior part of the pterygoids.

Distinguishing taxonomic features. The families and genera of the order Crocodilia are differentiated primarily by the anatomical peculiarities of their skulls. The classification of the species is based mainly upon external characteristics, such as the proportions of the snout, the bony structures on the dorsal side of the snout, the number of teeth, the number and arrangement of the large knobs on the nape of the neck, and the characteristics of the dorsal plates.

Annotated classification. Extinct groups represented only by fossils are indicated by a dagger (†).

ORDER CROCODILIA

Heavy cylindrical body; large, triangular head; legs short, toes webbed; long, muscular tail; large flat plates on belly, keeled ones on back; heart 4-chambered.

†**Suborder Protosuchia**

Upper Triassic; muzzle very short; choanae (internal nostrils) in region of palatine bones.

†**Suborder Mesosuchia**

Jurassic to Upper Cretaceous; choanae in posterior part of palatine bones.

†**Suborder Sebecosuchia**

Upper Cretaceous to Miocene; skull laterally flattened; choanae in depression in anterior part of pterygoids.

Suborder Eusuchia

Upper Jurassic to Recent; choanae entirely enclosed by pterygoids.

Family Alligatoridae (alligators)

Four genera and 7 species; teeth of lower jaw fit inside those of upper jaw.

Family Crocodilidae (true crocodiles)

Three genera and 13 species; teeth of upper and lower jaws form one interdigitating row when mouth is closed.

Family Gavialidae (gavial)

One genus and 1 species; extremely long snout, more than 22 teeth in each jaw; nasal bones separated from premaxillaries.

Critical appraisal. One authority has separated the order Crocodilomorpha into two suborders, Crocodilia and Paracrocodilia. According to this scheme the Crocodilia include as infra-orders those groups given above as suborders. This scheme also contains a suborder, Thalattosuchia.

Widely different views prevail concerning the classification of the living groups of Eusuchia—*i.e.*, the alligators, the true crocodiles, and the gavials. Of these, the alligators and the true crocodiles, without doubt more closely related to each other than to the gavials, are sometimes regarded as constituting two subfamilies of the family Crocodilidae. Some authors regard the gavials as a third subfamily. Others give the false gavial, or Sunda gavial, a special position with respect to the true crocodiles. For conciseness, these three groups have been treated here as distinct families.

(H.F.We.)

BIBLIOGRAPHY

General works: ANGUS D'A. BELLAIRS, *The Life of Reptiles* (1970); ROGER CONANT, *A Field Guide to Reptiles and Amphibians (Eastern U.S. and Canada)* (1958); P.J. DARLINGTON, *Zoogeography* (1957); CARL GANS *et al.* (eds.), *Biology of the Reptilia*, 3 vol. (1968–70) C.J. and O.B. GOIN, *Introduction to Herpetology*, 2nd ed. (1971); ARTHUR LOVERIDGE, *Reptiles of the Pacific World* (1945); ROBERT MERTENS, *La Vie des amphibiens et reptiles* (1959; Eng. trans., *The World of Amphibians and Reptiles*, 1960); J.A. OLIVER, *The Natural History of North American Amphibians and Reptiles* (1955); J.A. PETERS, *Dictionary of Herpetology* (1964); C.H. POPE, *The Reptile World* (1955); C.L. PROSSER and F.A. BROWN, JR., *Comparative Animal Physiology*, 2nd ed. (1961); K.P. SCHMIDT and R.F. INGER, *Living Reptiles of the World* (1957); M.A. SMITH, *The British Amphibians and Reptiles* (1951); R.C. STEBBINS, *A Field Guide to Western Reptiles and Amphibians* (1966); BERNARD S. MARTOF *et al.*, *Amphibians and Reptiles of the Carolinas and Virginia* (1980), a guide to 159 herpetozoan species.

Paleontology: E.H. COLBERT, *The Dinosaur Book* (1945) and *The Age of Reptiles* (1965); A.S. ROMER, *Vertebrate Paleontology*, 3rd ed. (1966).

Chelonia: G.A. BOULENGER, *Catalogue of the Chelonians, Rhynchocephalians, and Crocodiles in the British Museum (Natural History)*, new ed. (1889), an early standard description of the turtles of the world; A.F. CARR, *Handbook of Turtles: The Turtles of the United States, Canada, and Baja California* (1952), a detailed account including copious notes on habits; *So Excellent a Fishe: A Natural History of Sea Turtles* (1967), a readable account of Carr's study of the living green turtle; E.H. COLBERT, *Evolution of the Vertebrates: A History of the Backboned Animals Through Time* (1955), an authoritative work by a specialist who writes simply and clearly; R. CONANT, *A Field Guide to Reptiles and Amphibians of the United States and Canada East of the 100th Meridian* (1958), a useful field guide with illustrations of virtually all species and subspecies, as well as 248 distribution maps; P.E.P. DERANIYAGALA, *The Tetrapod Reptiles of Ceylon*, vol. 1, *Testudinates Marine and Terrestrial and Crocodilians* (1939), the biology of many Asian turtles; J.A. OLIVER, *The Natural History of North American Amphibians and Reptiles* (1955), a good book for the general reader; J.J. PARSONS, *The Green Turtle and Man* (1962), a fascinating history of the exploitation of the green turtle; C.H. POPE, *The Reptiles of China* (1935), descriptions of many Asiatic forms; *Turtles of the United States and Canada* (1939), profusely illustrated with excellent photographs; P.C.H. PRITCHARD, *Living Turtles of the World* (1967), a brief but complete guide with many illustrations in colour; A.S. ROMER, *Osteology of the Reptiles* (1956), an authoritative classification of reptiles; *Vertebrate Paleontology*, 3rd ed., (1966), contains a good summary of chelonian fossil history; N. de ROOIJ, *The Reptiles of the Indo-Australian Archipelago*, vol. 1, *Lacertilia, Chelonia, Emydosauria* (1915), a classic work on Australasian reptiles; H.M. SMITH and E.H. TAYLOR, *Herpetology of Mexico: Annotated Checklists and Keys to the Amphibians and Reptiles* (1966), identification of Mexican reptiles; M.A. SMITH, *Reptilia and Amphibia*, vol. 1, *Loricata, Testudines* in the "Fauna of British India, Ceylon and Burma" (1931), account of a rich turtle fauna by a leading specialist; J. de C. SOWERBY and E. LEAR, *Tortoises, Terrapins, and Turtles* (1872), an old work with many magnificent coloured illustrations made from life; H. WERMUTH and R. MERTENS, *Schildkröten, Krokodile, Brückenechsen* (1961), with descriptions of living turtles and many illustrations; E. WORRELL, *Reptiles of Australia* (1963), contains accounts of the habits of many Australian turtles; JACK J. RUDLOE, *Time of the Turtle* (1979), a study of sea turtles with emphasis on their conservation.

Rhynchocephalia: W.H. DAWBIN, "The Tuatara in Its Natural Habitat," *Endeavour*, 21:16–24 (1962), a review of distribution, life cycle, and ecology; A. GUNTHER, "Contribution to the Anatomy of *Hatteria* (*Rhynchocephalus*, Owen.)," *Phil. Trans. R. Soc.*, 157:595–629 (1867), the original detailed description of skeleton, muscles, externals, and classification; A.S. ROMER, *Osteology of the Reptiles* (1956), sections on musculature, the brain, and the circulatory system, as well as osteology and relationships; RICHARD SHARELL, *The Tuatara, Lizards and Frogs of New Zealand* (1966), a popular account of the natural history of the tuatara.

Sauria: General surveys of the species of lizards and their life histories, with many photographs, are found in K.P. SCHMIDT and R.F. INGER, *Living Reptiles of the World* (1957); A.F. CARR, *The Reptiles* (1963); R. MERTENS, *La vie des amphibiens et reptiles* (1959; Eng. trans., *The World of Amphibians and Reptiles,* (1960); and H.M. SMITH, *Handbook of Lizards: Lizards of the United States and Canada* (1946). V.A. HARRIS, *The Anatomy of the Rainbow Lizard* (1963), provides a dissection manual for students, moderately technical but very readable. Of a more technical nature is a work by A.S. ROMER: *Osteology of the Reptiles* (1956). W.W. MILSTEAD (ed.), *Lizard Ecology: A Symposium* (1967), contains summaries of research in population ecology, physiological ecology, and social behaviour, with panel discussions of the reports; an excellent source for understanding trends in research.

Important taxonomic books and papers, not easily read by the layman, but fundamental to an understanding of the classification are: C.L. CAMP, "Classification of the Lizards," *Bull. Am. Mus. Nat. Hist.,* 48:289–480 (1923); R. HOFFSTETTER, "Revue des récentes acquisitions concernant l'histoire et la systématique des squamates," in *Problèms actuels de paléontologie (évolutions des vertébrés),* pp. 243–279 (1962); A.G. KLUGE, "Higher Taxonomic Categories of Gekkonid Lizards, and Their Evolution," *Bull. Am. Mus. Nat. Hist.,* 135:1–59 (1967); S.B. MCDOWELL and C.M. BOGERT, "The Systematic Position of *Lanthanotus* and the Affinities of the Anguinomorphan Lizards," *ibid.,* 105:1–142 (1954); M.R. MILLER, "The Cochlear Duct of Lizards," *Proc. Calif. Acad. Sci., 4th series,* 33:255–359 (1966), and "The Cochlear Duct of Snakes," *ibid.,* 35:425–476 (1968); G. UNDERWOOD, "On the Classification and Evolution of Geckoes," *Proc. Zool. Soc., Lond.,* 124:469–492 (1954), "On Lizards of the Family Pygopodidae, a Contribution to the Morphology and Phylogeny of the Squamata," *J. Morph.,* 100:207–268 (1957), and "A Contribution to the Classification of Snakes," *Publs. Br. Mus. Nat. Hist.,* no. 635 (1967); A.E. GREER, "A Subfamilial Classification of Scincid Lizards," *Bull. Mus. Comp. Zool. Harv.,* 139:151–183 (1970).

Serpentes: CARL GANS *et al.* (eds.), *Biology of the Reptilia* (1969–), a multivolume work, made up of contributions by specialists, that will eventually summarize current knowledge about the class Reptilia, with emphasis on morphology, embryology and physiology, and ecology and behaviour; GEORGE A. BOULENGER, *Catalogue of the Snakes in the British Museum, Natural History,* 3 vol. (1893–96), a classic work, the only attempt ever made in the English language to list, describe, and help identify all the snakes of the world; GARTH UNDERWOOD, *A Contribution to the Classification of Snakes* (1967), a highly technical discussion of relationships among the various kinds of snakes; JAMES A. PETERS, *Dictionary of Herpetology* (1964), contains extensive information about snakes and other reptiles, in a readily accessible form; ROGER CONANT, *A Field Guide to Reptiles and Amphibians of the United States and Canada East of the 100th Meridian* (1958), and ROBERT C. STEBBINS, *A Field Guide to Western Reptiles and Amphibians* (1966), extensively illustrated manuals for field identification of snakes, as well as other reptiles; CLIFFORD H. POPE, *The Giant Snakes* (1961), an entertaining discussion of the very large snakes, and *Snakes Alive and How They Live* (1937, reprinted 1958), although somewhat outdated, still a readable account of snake natural history; ALBERT H. and ANNA A. WRIGHT, *Handbook of Snakes of the United States and Canada,* 2 vol. (1957), detailed descriptions of all the kinds of snakes known on the North American continent, with many photographs; COLEMAN J. and OLIVE B. GOIN, *Introduction to Herpetology* (1962), a textbook designed for undergraduate classes at the college level; KARL P. SCHMIDT and ROBERT F. INGER, *Living Reptiles of the World,* pp. 175–279 (1957), a popular account, aimed at the general public, profusely illustrated with colour photographs; RAMONA and DESMOND MORRIS, *Men and Snakes* (1965), a thorough review of man's relationships with snakes, including culture, mythology, medicine, and exploitation; H.W. PARKER, *Natural History of Snakes* (1965), one of the handbook guides to the British Museum, and *Snakes* (1963), a factual account of the general biology of snakes; LAURENCE M. KLAUBER, *Rattlesnakes: Their Habits, Life Histories, and Influence on Mankind,* 2 vol. (1956; abridged ed., 1982), a thorough summary, not only of rattlesnakes but of all snakes, their biology and relationships with man; HOBART M. SMITH, *Snakes As Pets* (1965), a guide to the care and feeding of snakes in captivity; JAMES A. OLIVER, *Snakes in Fact and Fiction* (1958), a popular account of the snake, with both true and imaginative stories about size, food, numbers, and other bits of natural history; WOLFGANG BUCHERL, E.E. BUCKLEY, and V. DEULOFEU (eds.), *Venomous Animals and Their Venoms,* 2 vol. (1968–71), a discussion of the classification, distribution, and biology of venomous snakes, with detailed discussions of thir venoms as well as treatment of snakebite; SHERMAN A. and MADGE R. MINTON, *Venomous Reptiles* (1969), an entertaining and often anecdotal account of venoms, venomous reptiles, their bites and the treatment of bites, and the intricate relationships of snakes with man and his cultures; TONY PHELPS, *Poisonous Snakes* (1981), a treatment of the rear-fanged colubrids, the elapids, and the vipers.

Crocodilia: M.M. COHEN and C. GANS, "The Chromosomes of the Order Crocodilia," *Cytogenetics,* 9:81–105 (1970), pictures of karyotypes of all recent species of crocodiles and an attempt to interpret them from the viewpoint of evolution; H.B. COTT, "Scientific Results of an Inquiry into the Ecology and Economic Status of the Nile Crocodile (*Crocodilus niloticus*) in Uganda and Northern Rhodesia," *Trans. Zool. Soc. Lond.,* 29:211–356 (1961), one of the most comprehensive accounts of the behaviour and territorialism of the Nile crocodile; O. KUHN, *Die vorzeitlichen Krokodile* (1968), a survey of all recent and fossil crocodiles and a proposal of a common classification for both; F.J. MEDEM, "The Crocodilian Genus *Paleosuchus*," *Fieldiana, Zool.,* 39: 227–247 (1958), the first comprehensive field work on smooth-fronted caimans, especially on their ecology; W.T. NEILL, *The Last of the Ruling Reptiles: Alligators, Crocodiles, and Their Kin* (1971), a richly illustrated and extremely careful compendium of the biology of all recent crocodiles; A.S. ROMER, *Vertebrate Paleontology,* 3rd ed. (1966), the classic work on the classification and phylogeny of the vertebrates and their mutual relations; A.D. WALKER, "A Revision of the Jurassic Reptile *Hallopus victor* (Marsh), with Remarks on the Classification of Crocodiles," *Phil. Trans. R. Soc.,* Series B, 257:323–372 (1970), a discussion of the division of crocodiles (Crocodilomorpha) in the suborders Crocodilia and Paracrocodilia based on anatomical examination of some fossil forms; H. WERMUTH, "Systematik der rezenten Krokodile," *Mitt. Zool. Mus. Berl.,* 29:375–514 (1950), a review of recent crocodiles and a discussion of the taxonomic importance of their characteristics, with figures of heads and skulls of all species; "Farbwechsel und Lernfähigkeit bei Krokodilen," *Dt. Aquar.-Terrar.-Z.,* 16:90–92 (1963), observations on the intelligence of a tame, seven-foot, spectacled caiman; O. VON WETTSTEIN, "Crocodilia," *Handb. Zool.,* 7:236–424 (1931), a complete monograph of the recent crocodiles, with an extensive bibliography.

Respiration and Respiratory Systems

Respiration is the process by which animal organisms take up oxygen and discharge carbon dioxide in order to satisfy their energy requirements. In the living organism, energy is liberated, along with carbon dioxide, through the oxidation of molecules containing carbon. The term respiration also denotes the exchange of the respiratory gases (oxygen and carbon dioxide) between the organism and the medium in which it lives and between the cells of the body and the tissue fluid that bathes them.

With the exception of energy used by animal life in the deep ocean, all energy used by animals is ultimately derived from the energy of sunlight. The carbon dioxide in the atmosphere in conjunction with the energy of sunlight is used by plants to synthesize sugars and other components. Animals consume plants or other organic material to obtain chemical compounds, which are then oxidized to sustain vital processes.

This article first considers the gaseous components of air and water, the natural respiratory habitats of animals, and the basic types of respiratory structures that facilitate gas exchange in these environments. After this general survey, the human respiratory system and human respiration are examined in detail, followed by an overview of diseases and disorders affecting respiration. For a depiction of some of the structures that make up the human respiratory system, shown in relation to other parts of the gross anatomy, see the colour Trans-Vision in the *Propædia*, Part Four, Section 421.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 336, 421, and 423.

The article is divided into the following sections:

General features of the respiratory process 725
- The gases in the environment 725
- Basic types of respiratory structures 726
  - Respiratory organs of invertebrates
  - Respiratory organs of vertebrates
- Dynamics of vertebrate respiratory mechanisms 730
  - Fishes
  - Amphibians
  - Reptiles
  - Birds
  - Mammals

Human respiration 733
- The design of the respiratory system 733
  - Morphology of the upper airways
  - Morphology of the lower airways
  - Structural design of the airway tree
  - The lungs
- Control of breathing 738
  - Central organization of respiratory neurons
  - Chemoreceptors
  - Variations in breathing
- The mechanics of breathing 740
  - The lung–chest system
  - The role of muscles
  - The respiratory pump and its performance
- Gas exchange 741
  - Transport of oxygen
  - Transport of carbon dioxide
  - Gas exchange in the lung
  - Abnormal gas exchange
- Interplay of respiration, circulation, and metabolism 743
- Adaptations 744
  - High altitudes
  - Swimming and diving

Diseases and disorders of respiration 746
- Signs and symptoms 746
- The defenses of the lung 747
- Methods of investigation 747
- Morphological classification of respiratory disease 747
- Viral infections 748
- Bacterial pneumonia 749
- Allergic lung diseases 750
  - Asthma
  - Hay fever
  - Hypersensitivity pneumonitis
- Bronchitis and bronchiolitis 751
  - Acute bronchitis
  - Chronic bronchitis
  - Pulmonary emphysema
  - Bronchiectasis
- Occupational lung disease 752
- Lung cancer 753
- Miscellaneous pulmonary conditions 754
- Lung transplantation 756

Bibliography 756

General features of the respiratory process

Although the acquisition of oxygen and the elimination of carbon dioxide are essential requirements for all animals, the rate and amount of gaseous exchange vary according to the kind of animal and its state of activity. In the Table the oxygen consumption of various animals is expressed in terms of millilitres of oxygen per kilogram of body weight per hour, reflecting the gas demands of different species at rest and in motion. A change in the chemical composition of the body fluids elicits a response from the central nervous system, which then excites or depresses the machinery of external respiration.

THE GASES IN THE ENVIRONMENT

The range of respiratory problems faced by aquatic and terrestrial animals can be seen from the varying composition and physical characteristics of water and air. Air contains about 20 times the amount of oxygen found in air-saturated water. In order to extract an equivalent amount of oxygen as an air breather, an aquatic animal may find it necessary to pass across the respiratory surfaces a relatively larger volume of the external medium. Moreover, the diffusion rate of oxygen is much lower in water than in air. The problem is further compounded by the higher density (1,000 times air) and viscosity (100 times air) of water, which impose on the machinery of aquatic respiration a much greater work load. Thus, fish may expend about 20 percent of their total oxygen consumption in

Oxygen Consumption of Various Animals and Its Variation with Rest and Activity

	weight (grams)	oxygen consumption (millilitres per kilogram of weight per hour)
Paramecium	0.000001	500
Mussel (*Mytilus*)	25	22
Crayfish (*Astacus*)	32	47
Butterfly (*Vanessa*)	0.3	
resting		600
flying		100,000
Carp (*Cyprinus*)	200	100
Pike (*Esox*)	200	350
Mouse	20	
resting		2,500
running		20,000
Human	70,000	
resting		200
maximal work		4,000

Source: A. Krogh, *The Comparative Physiology of Respiratory Mechanisms* (1959).

running the respiratory pump, as compared with about 1 to 2 percent in mammals, including humans.

The carbon dioxide content of most natural waters is low compared with air, often almost nil. In contrast to oxygen, carbon dioxide is extremely soluble in water and diffuses rapidly. Most of the carbon dioxide entering water combines either with the water (to form carbonic acid) or with other substances (to form carbonates or bicarbonates). This buffering capacity maintains a low level of free carbon dioxide and facilitates the maintenance of a favourable diffusion gradient for carbon dioxide exchange by water breathers. In general, oxygen exchange, which is strongly dependent on the oxygen content of the water, is more critically limiting for aquatic forms than is the exchange of carbon dioxide.

Temperature exerts a profound effect on the solubility of gases in water. A change from 5° to 35° C (41° to 95° F) reduces the oxygen content of fresh water by nearly half. At the same time, a rise in body temperature produces an increase in oxygen consumption among animals that do not closely regulate their body temperatures (so-called cold-blooded animals). A fish experiencing both rising water and body temperatures is under a double handicap: more water must be pumped across its gill surfaces to extract the same amount of oxygen as was needed at the lower temperature; and the increased metabolism requires greater quantities of oxygen.

The amount of oxygen available in natural waters is also limited by the amount of dissolved salts. This factor is a determinant of oxygen availability in transitional zones between sea and fresh water. Pure water, when equilibrated with oxygen at 0° C, for example, contains about 50 millilitres of oxygen per litre; under the same conditions, a solution containing 2.9 percent of sodium chloride contains only 40 millilitres of oxygen per litre. Bodies of water may have oxygen-poor zones. Such zones are especially evident in swamps and at the lower levels of deep lakes. Many animals are excluded from such zones; others have become remarkably adapted to living in them.

The Earth's atmosphere extends to a height of many miles. It is composed of a mixture of gases held in an envelope around the globe by gravitational attraction. The atmosphere exerts a pressure proportional to the weight of a column of air above the surface of the Earth extending to the limit of the atmosphere: atmospheric pressure at sea level is on average sufficient to support a column of mercury 760 millimetres in height (abbreviated as 760 mm Hg—the latter being the chemical symbol for mercury). Dry air is composed chiefly of nitrogen and inert gases (79.02 percent), oxygen (20.94 percent), and carbon dioxide (0.03 percent), each contributing proportionately to the total pressure. These percentages are relatively constant to about 80.5 kilometres in altitude. At sea level and a barometric pressure of 760 millimetres of mercury, the partial pressure of nitrogen is 79.02 percent of 760 millimetres of mercury, or 600.55 millimetres of mercury; that of oxygen is 159.16 millimetres of mercury; and that of carbon dioxide is 0.20 millimetres of mercury.

Composition of air

The existence of water vapour in a gas mixture reduces the partial pressures of the other component gases but does not alter the total pressure of the mixture. The importance of water-vapour pressure to gas composition can be appreciated from the fact that at the body temperature of humans (37° C, or 98.6° F) the atmospheric air drawn into the lungs becomes saturated with water vapour. The water-vapour pressure at 37° C is 47 millimetres of mercury. To calculate the partial pressures of the respiratory gases, this value must be subtracted from the atmospheric pressure. For oxygen, 760 (the atmospheric pressure) − 47 = 713 millimetres of mercury, and 713 × 0.209 (the percentage of oxygen in the atmosphere) = 149 millimetres of mercury; this amounts to some 10 millimetres of mercury lower than the partial pressure of oxygen in dry air at 760 millimetres of mercury total pressure.

Atmospheric pressures fall at higher altitudes, but the composition of the atmosphere remains unchanged. At 7,600 metres (25,000 feet) the atmospheric pressure is 282 millimetres of mercury and the partial pressure of oxygen is about 59 millimetres of mercury. Oxygen continues to constitute only 20.94 percent of the total gas present. The rarefaction of the air at high altitudes not only limits the availability of oxygen for the air breather, it also limits its availability for aquatic forms, since the amount of dissolved gas in water decreases in parallel with the decline in atmospheric pressure. Lake Titicaca in Peru is at an altitude of about 3,810 metres; one litre of lake water at this altitude (and at 20° C, or 68° F) holds four millilitres of oxygen in solution; at sea level, it would hold 6.4.

The variations in the characteristics of air and water suggest the many problems with which the respiratory systems of animals must cope in procuring enough oxygen to sustain life.

BASIC TYPES OF RESPIRATORY STRUCTURES

Adaptation of respiratory structures

Respiratory structures are tailored to the need for oxygen. Minute life-forms, such as protozoans, exchange oxygen and carbon dioxide across their entire surfaces. Multicellular organisms, in which diffusion distances are longer, generally resort to other strategies. Aquatic worms, for example, lengthen and flatten their bodies to refresh the external medium at their surfaces. Sessile sponges rely on the ebb and flow of ambient water. By contrast, the jellyfish, which can be quite large, has a low oxygen need because its content of organic matter is less than 1 percent and its metabolizing cells are located just beneath the surface, so that diffusing distances are small.

Organisms too large to satisfy their oxygen needs from the environment by diffusion are equipped with special respiratory structures in the form of gills, lungs, specialized areas of the intestine or pharynx (in certain fishes), or tracheae (air tubes penetrating the body wall, as in insects).

Typical respiratory structures

Respiratory structures typically have an attenuated shape and a semipermeable surface that is large in relation to the volume of the structure. Within them there is usually a circulation of body fluids (blood through the lungs, for example). Two sorts of pumping mechanisms are frequently encountered: one to renew the external oxygen-containing medium, the other to ensure circulation of the body fluids through the respiratory structure. In air-breathing vertebrates, alternately contracting sets of muscles create the pressure differences needed to expand or deflate the lungs, while the heart pumps blood through the respiratory surfaces within the lungs. Oxygenated blood returning to the heart is then pumped through the vascular system to the various tissues where the oxygen is consumed.

Respiratory organs of invertebrates. Two common respiratory organs of invertebrates are trachea and gills. Diffusion lungs, as contrasted with ventilation lungs of vertebrates, are confined to small animals, such as pulmonate snails and scorpions.

Trachea. This respiratory organ is a hallmark of insects (Figure 1). It is made up of a system of branching tubes that deliver oxygen to, and remove carbon dioxide from, the tissues, thereby obviating the need for a circulatory system to transport the respiratory gases (although the circulatory system does serve other vital functions, such as the delivery of energy-containing molecules derived from food). The pores to the outside, called spiracles, are typically paired structures, two in the thorax and eight in the abdomen. Periodic opening and closing of the spiracles prevents water loss by evaporation, a serious threat to insects that live in dry environments. Muscular pumping motions of the abdomen, especially in large animals, may promote ventilation of the tracheal system.

Bubble breathers

Although tracheal systems are primarily designed for life in air, in some insects modifications enable the tracheae to serve for gas exchange under water (Figure 2). Of special interest are the insects that might be termed bubble breathers, which, as in the case of the water beetle *Dytiscus*, take on a gas supply in the form of an air bubble under their wing surfaces next to the spiracles before they submerge. Tracheal gas exchange continues after the beetle submerges and anchors beneath the surface. As oxygen is consumed from the bubble, the partial pressure of oxygen within the bubble falls below that in the water; consequently oxygen diffuses from the water into the bubble to replace that consumed. The carbon dioxide produced by the insect diffuses through the tracheal system into

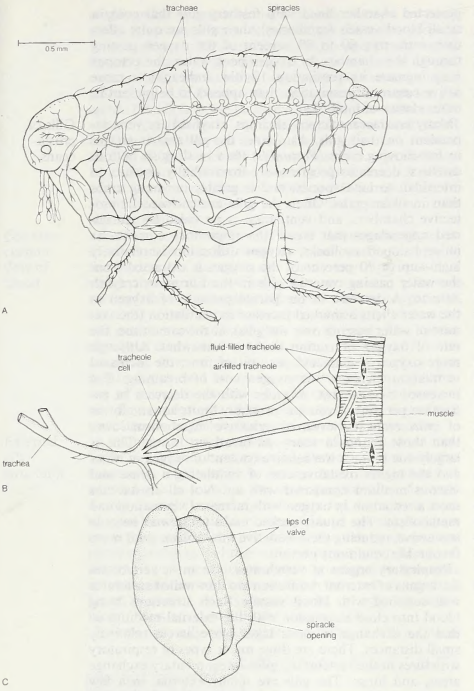


Figure 1: *Insect respiratory apparatus.*
(A) The main branches of the tracheal system of the flea; the spiracles are shown as a row of small circles along the side of the body. (B) Minute air capillaries (tracheoles) indenting the surface of muscle cells. (C) Enlarged view of a spiracle showing the valve lips and opening.

From (A) V.B. Wigglesworth, *The Principles of Insect Physiology*, 7th ed. (1972), Chapman & Hall, London; (B) V.B. Wigglesworth, "A Theory of Tracheal Respiration in Insects," *Proceedings of the Royal Society of London*, series B, vol. 106, pp. 229–250 (1930); (C) after R.E. Snodgrass in R. Buchsbaum, *Animals Without Backbones*, 2nd ed. (1948), The University of Chicago Press

the bubble and thence into the water. The bubble thus behaves like a gill. There is one major limitation to this adaptation: As oxygen is removed from the bubble, the partial pressure of the nitrogen rises, and this gas then diffuses outward into the water. The consequence of outward nitrogen diffusion is that the bubble shrinks and its oxygen content must be replenished by another trip to the surface. A partial solution to the problem of bubble renewal has been found by small aquatic beetles of the family Elmidae (*e.g., Elmis, Riolus*), which capture bubbles containing oxygen produced by algae and incorporate this gas into the bubble gill. Several species of aquatic beetles also augment gas exchange by stirring the surrounding water with their posterior legs.

Plastron

An elegant solution to the problem of bubble exhaustion during submergence has been found by certain beetles that have a high density of cuticular hair over much of the surface of the abdomen and thorax. The hair pile is so dense that it resists wetting, and an air space forms below it, creating a plastron, or air shell, into which the tracheae open. As respiration proceeds, the outward diffusion of nitrogen and consequent shrinkage of the gas space are prevented by the surface tension—a condition manifested by properties that resemble those of an elastic skin under tension—between the closely packed hairs and the water. The plastron becomes "permanent" in the sense that further bubble trapping at the surface is no longer necessary, and the beetles may remain submerged indefinitely. Since the plastron hairs tend to resist deformation, the beetles can live at considerable depths without compression of the plastron gas.

One extraordinary strategy used by the hemipteran insects *Buenoa* and *Anisops* is an internal oxygen store that enables them to lurk for minutes without resurfacing while awaiting food in relatively predator-free but oxygen-poor mid-water zones. The internal oxygen store is in the form of hemoglobin-filled cells that constitute the first line of oxygen delivery to actively metabolizing cells, sparing the small air mass in the tracheal system while the hemoglobin store is being depleted.

The "book lungs" of spiders

The respiratory structures of spiders consist of peculiar "book lungs," leaflike plates over which air circulates through slits on the abdomen. The book lungs contain blood vessels that bring the blood into close contact with the surface exposed to the air and where gas exchange between blood and air occurs. In addition to these structures, there may also be abdominal spiracles and a tracheal system like that of insects.

Since spiders are air breathers, they are mostly restricted to terrestrial situations, although some of them regularly hunt aquatic creatures at stream or pond edges and may actually travel about on the surface film as easily as on land. The water spider *Argyroneta aquatica*, the frogman of the spider world, utilizes the water-beetle method of capturing air bubbles at the surface. The bubble is pressed against the respiratory openings on the abdomen, but, because there is no permanent plastron, trips must be made to the surface for bubble renewal. Most of the life cycle of the water spider, however, including courtship and breeding, prey capture and feeding, and the development of eggs and embryos, occurs below the water surface. Many of these activities take place in a kind of diving bell formed by silk. The spider weaves an inverted basketlike web that is anchored to underwater plants or other objects. Bubbles captured at the surface are ejected into the interior, inflating the underwater house with air. The combination of bubble trapping and web building has given *Argyroneta aquatica* access to an environment denied most of its relatives.

Many immature insects have special adaptations for an aquatic existence. Thin-walled protrusions of the integument, containing tracheal networks, form a series of gills (tracheal gills) that bring water into close contact with the closed tracheal tubes (see Figure 2). The nymphs of mayflies and dragonflies have external tracheal gills attached to their abdominal segments, and certain of the gill plates may move in a way that sets up water currents over the exchange surfaces. Dragonfly nymphs possess a series of tracheal gills enclosed within the rectum. Periodic pumping of the rectal chamber serves to renew water flow over the gills. Removing the gills or plugging the rectum results in lower oxygen consumption. Considerable gas exchange also occurs across the general body surface in immature aquatic insects.

The insect tracheal system has inherent limitations. Gases diffuse slowly in long narrow tubes, and effective gas transport can occur only if the tubes do not exceed a certain length. It is generally thought that this has imposed a size limit upon insects.

Varieties of gill-like structures

Gills of invertebrates. Gills are evaginations of the body surface. Some open directly to the environment; others, as in fishes, are enclosed in a cavity. In contrast, lungs represent invaginations of the body surface. Many invertebrates use gills as a major means of gas exchange; a few, such as the pulmonate land snail, use lungs. Almost any thin-walled extension of the body surface that comes in contact with the environmental medium and across which gas exchange occurs can be viewed as a gill. Gills usually have a large surface area in relation to their mass; pumping devices are often employed to renew the external medium. Although gills are generally used for water breathing and lungs for air breathing, this association is not invariable, as exemplified by the water lungs of sea cucumbers.

The marine polychaete worms use not only the general body surface for gas exchange but also a variety of gill-like structures: segmental flaplike parapodia (in *Nereis*) or

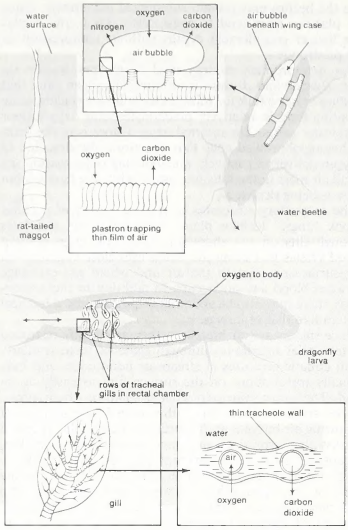


Figure 2: Ways by which aquatic insects obtain oxygen.
This figure first appeared in *New Scientist*, the weekly review of science and technology, 128 Long Acre, London W.C. 2

elaborate branchial tufts (among the families Terebellidae and Sabellidae). The tufts, used to create both feeding and respiratory currents, offer a large surface area for gas exchange.

In echinoderms (starfish, sea urchins, brittle stars), most of the respiratory exchange occurs across tube feet (a series of suction-cup extensions used for locomotion). However, this exchange is supplemented by extensions of the coelomic, or body-fluid, cavity into thin-walled "gills" or dermal branchiae that bring the coelomic fluid into close contact with seawater. Sea cucumbers (Holothuroidea), soft-bodied, sausage-shaped echinoderms that carry on some respiration through their oral tentacles, which correspond to tube feet, also have an elaborate "respiratory tree" consisting of branched hollow outpouchings off the cloaca (hindgut). Water is pumped in and out of this system by the action of the muscular cloaca, and it is probable that a large fraction of the animals' respiratory gas is exchanged across this system.

The gills of mollusks have a relatively elaborate blood supply, although respiration also occurs across the mantle, or general epidermis. Clams possess gills across which water circulates, impelled by the movements of millions of microscopic whips called cilia. In the few forms studied, the extraction of oxygen from the water has been found to be low, on the order of 2 to 10 percent. The currents produced by cilial movement, which constitute ventilation, are also utilized for bringing in and extracting food. At low tide or during a dry period, clams and mussels close their shells and thus prevent dehydration. Metabolism then shifts from oxygen-consuming (aerobic) pathways to oxygen-free (anaerobic) pathways, which causes acid products to accumulate; when normal conditions are restored, the animals increase their ventilation and oxygen extraction in order to rid themselves of the acid products. In snails, the feeding mechanism is independent of the respiratory surface. A portion of the mantle cavity in the form of a gill or "lung" serves as a gas-exchange site. In air-breathing snails, the "lung" may be protected from drying out through contact with the air by having only a pore in the mantle as an opening to the outside. Cephalopod mollusks, such as squid and octopus, actively ventilate a protected chamber lined with feathery gills that contain small blood vessels (capillaries); their gills are quite effective, extracting 60 to 80 percent of the oxygen passing through the chamber. In oxygen-poor water, the octopus may increase its ventilation 10-fold, indicating a more active control of respiration than appears to be present in other classes of mollusks.

Gills of crustaceans

Many crustaceans (crabs, shrimps, crayfish) are very dependent on their gills. As a rule, the gill area is greater in fast-moving crabs (*Portunids*) than in sluggish bottom dwellers; decreases progressively from wholly aquatic, to intertidal, to land species; and is greater in young crabs than in older crabs. Often the gills are enclosed in protective chambers, and ventilation is provided by specialized appendages that create the respiratory current. As in cephalopod mollusks, oxygen utilization is relatively high—up to 70 percent of the oxygen is extracted from the water passing over the gills in the European crayfish *Astacus*. A decrease in the partial pressure of oxygen in the water elicits a marked increase in ventilation (the volume of water passing over the gills); at the same time, the rate of oxygen utilization declines somewhat. Although more oxygen is extracted per unit of time, the increased ventilation increases the oxygen cost of breathing. The increased oxygen cost, together with the decrease in extraction per unit of volume, probably limits aquatic forms of crustaceans to levels of oxidative metabolism lower than those found in many air-breathing forms. This is largely due to the lower relative content of oxygen in water and the higher oxidative cost of ventilating a dense and viscous medium compared with air. Not all crustaceans meet a reduction in oxygen with increased ventilation and metabolism. The square-backed crabs (*Sesarma*) become less active, reducing their oxidative metabolism until more favourable conditions prevail.

Respiratory organs of vertebrates. In most vertebrates the organs of external respiration are thin-walled structures well supplied with blood vessels. Such structures bring blood into close association with the external medium so that the exchange of gases takes place across relatively small distances. There are three major types of respiratory structures in the vertebrates: gills, integumentary exchange areas, and lungs. The gills are totally external in a few forms (as in *Necturus*, a neotenic salamander), but in most they are composed of filamentous leaflets protected by bony plates (as in fish). Some fishes and numerous amphibians also use the body integument, or skin, as a gas-exchange structure. Both gills and lungs are formed from outpouchings of the gut wall during embryogenesis. Such structures have the advantage of a protected internal location, but this requires some sort of pumping mechanism to move the external gas-containing medium in and out.

Ventilation volume

The quantity of air or water passing through the lungs or gills each minute is known as the ventilation volume. The rate or depth of respiration may be altered to bring about adjustments in ventilation volume. The ventilation volume of humans at rest is approximately six litres per minute. This may increase to more than 100 litres per minute with increases in the rate of respiration and the quantity of air breathed in during each respiratory cycle (tidal volume). Certain portions of the airways (trachea, bronchi, bronchioles) do not participate in respiratory exchange, and the gas that fills these structures occupies an anatomical dead space of about 150 millilitres in volume. Of a tidal volume of 500 millilitres, only 350 millilitres ventilate the gas-exchange sites.

The maximum capacity of human lungs is about six litres. During normal quiet respiration, a tidal volume of about 500 millilitres is inspired and expired during every respiratory cycle. The lungs are not collapsed at the close of expiration; a certain volume of gas remains within them. At the close of the expiratory act, a normal subject may, by additional effort, expel another 1,200 millilitres of gas. Even after the most forceful expiratory effort, however, there remains a residual volume of approximately 1,200 millilitres. By the same token, at the end of a normal inspiration, further effort may succeed in drawing into the lungs an additional 3,000 millilitres.

Fish gills

The gills. The gills of fishes are supported by a series

of gill arches encased within a chamber formed by bony plates (the operculum). A pair of gill filaments projects from each arch; between the dorsal (upper) and ventral (lower) surfaces of the filaments, there is a series of secondary folds, the lamellae, where the gas exchange takes place (Figure 3). The blood vessels passing through the gill arches branch into the filaments and then into still smaller vessels (capillaries) in the lamellae. Deoxygenated blood from the heart flows in the lamellae in a direction counter to that of the water flow across the exchange surfaces. In a number of fishes the water-to-blood distance across which gases must diffuse is 0.0003 to 0.003 millimetre, or about the same distance as the air-to-blood pathway in the mammalian lung.

Countercurrent flow of blood

The countercurrent flow of blood through the lamellae in relation to external water flow has much to do with the efficiency of gas exchange. Laboratory experiments in which the direction of water flow across fish gills was reversed showed that about 80 percent of the oxygen was extracted in the normal situation, while only 10 percent was extracted when water flow was reversed. The uptake of oxygen from water to blood is thus facilitated by countercurrent flow; in this way, greater efficiency of oxygen uptake is achieved by an anatomical arrangement that is free of energy expenditure by the organism. Countercurrent flow is a feature of elasmobranchs (sharks, skates) and cyclostomes (hagfishes, lampreys) as well as bony fishes.

External gill structures

A number of vertebrates use externalized gill structures. Some larval fishes have external gills that are lost with the appearance of the adult structures. A curious example of external gills is found in the male lungfish (*Lepidosiren*). At the time the male begins to care for the nest, a mass of vascular filaments (a system of blood vessels) develops as an outgrowth of the pelvic fins. The fish meets its own needs by refilling its lungs with air during periodic excursions to the water surface. When it returns to the nest, its pelvic-gill filaments are perfused with well-oxygenated blood, providing an oxygen supply for the eggs, which are more or less enveloped by the gill filaments.

It is theoretically possible for a skin that is well supplied with blood vessels to serve as a major or even the only respiratory surface. This requires a thin, moist, and heavily vascularized skin, which increases the animal's vulnerability to enemies. In terrestrial animals a moist integument also provides a major avenue of water loss. A number of fishes and amphibians rely on the skin for much of their respiratory exchange; hibernating frogs utilize the skin for practically all their gas exchanges.

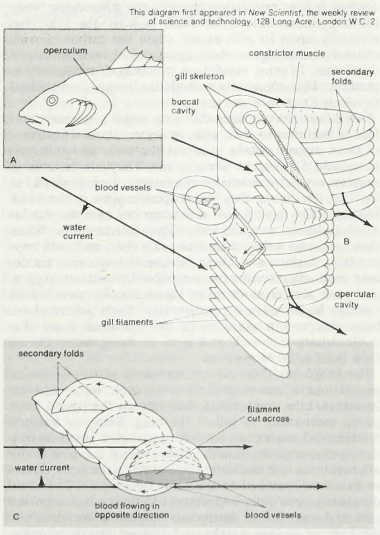


Figure 3: (A) Position of four gill arches beneath operculum on left side of fish. (B) Part of two gill arches with filaments of adjacent rows touching at their tips and the blood vessels that carry the blood before and after its passage over the gills. (C) Part of a single filament with three secondary folds on each side.

The lung. The lungs of vertebrates range from simple saclike structures found in the Dipnoi (lungfishes) to the complexly subdivided organs of mammals and birds. An increasing subdivision of the airways and the development of greater surface area at the exchange surfaces appear to be the general evolutionary trend among the higher vertebrates.

In the embryo, lungs develop as an outgrowth of the forward portion of the gut. The lung proper is connected to the outside through a series of tubes; the main tube, known as the trachea (windpipe), exits in the throat through a controllable orifice, the glottis. At the other end the trachea subdivides into secondary tubes (bronchi), in varying degree among different vertebrate groups.

The trachea of amphibians is not divided into secondary tubes but ends abruptly at the lungs. The relatively simple lungs of frogs are subdivided by incomplete walls (septa), and between the larger septa are secondary septa that surround the air spaces where gas exchange occurs. The diameter of these air spaces (alveoli) in lower vertebrates is larger than in mammals: The alveolus in the frog is about 10 times the diameter of the human alveolus. The smaller alveoli in mammals are associated with a greater surface for gas exchange: although the respiratory surface of the frog (*Rana*) is about 20 square centimetres per cubic centimetre (50.8 square inches per one cubic inch) of air, that of humans is about 300 square centimetres.

Characteristics of lungs

An important characteristic of lungs is their elasticity. An elastic material is one that tends to return to its initial state after the removal of a deforming force. Elastic tissues behave like springs. As the lungs are inflated, there is an accompanying increase in the energy stored within the elastic tissues of the lungs, just as energy is stored in a stretched rubber band. The conversion of this stored, or potential, energy into kinetic, or active, energy during the deflation process supplies part of the force needed for the expulsion of gases. A portion of the energy put into expansion is thus recovered during deflation. The elastic properties of the lungs have been studied by inflating them with air or liquid and measuring the resulting pressures. Muscular effort supplies the motive power for expanding the lungs, and this is translated into the pressure required to produce lung inflation. It must be great enough to overcome (1) the elasticity of the lung and its surface lining; (2) the frictional resistance of the lungs; (3) the elasticity of the thorax or thoraco-abdominal cavity; (4) frictional resistance in the body-wall structures; (5) resistance inherent in the contracting muscles; and (6) the airway resistance. The laboured breathing of the asthmatic is an example of the added muscular effort necessary to achieve adequate lung inflation when airway resistance is high, owing to narrowing of the tubes of the airways.

Studies of the pressure–volume relationship of lungs filled with salt solution or air have shown that the pressure required to inflate the lungs to a given volume is less when the lungs are filled with liquid than when they are filled with air. The differences in the two circumstances have been thought to result from the nature of the environment-alveolar interface, that interface being liquid–liquid in the fluid-filled lung and gas–liquid in the air-filled lung. In the case of the latter, the pressure–volume relationship represents the combined effects of the elastic properties of the lung wall plus the surface tension of the film, or surface coating, lining the lungs. Surface tension is the property, resulting from molecular forces, that exists in the surface film of all liquids and tends to contract the volume into a form with the least surface area; the particles in the surface are inwardly attracted, thus resulting in tension. Surface tension is nearly zero in the fluid-filled lung.

Elasticity of alveoli

The alveoli of the lungs are elastic bodies of nonuniform size. If their surfaces had a uniform surface tension, small

alveoli would tend to collapse into large ones. The result in the lungs would be an unstable condition in which some alveoli would collapse and others would overexpand. This does not normally occur in the lung because of the properties of its surface coating (surfactant), a complex substance composed of lipid and protein. Surfactant causes the surface tension to change in a nonlinear way with changes in surface area. As a result, when the lungs fill with air, the surface tensions of the inflated alveoli are less than those of the relatively undistended alveoli. This results in a stabilization of alveoli of differing sizes and prevents the emptying of small alveoli into larger ones. It has been suggested that compression wrinkles of the surface coating and attractive forces between adjacent wrinkles inhibit expansion. Surfactants have been reported to be present in the lungs of birds, reptiles, and amphibians.

DYNAMICS OF VERTEBRATE RESPIRATORY MECHANISMS

Organs in primitive vertebrates

Fishes. Among the most primitive of present-day vertebrates are the cyclostomes (lampreys and hagfishes), the gill structures of which are in the form of pouches that connect internally with the pharynx (throat) and open outward through slits, either by a fusion of the excurrent gill ducts into a single tube (in *Myxine*) or individually by separate gill slits (in *Petromyzon*). The gill lamellae of cyclostomes form a ring around the margins of the gill sac, and the series of sacs is supported in a flexible branchial skeleton. The number of paired pouches varies in different forms from six to 14. The pharynx of lampreys divides into an esophagus above and a blind tube below, from which the gill pouches arise. The upper pharynx of hagfishes communicates to the exterior through a nostril, a structure absent in lampreys. When the parasitic lampreys are embedded in the flesh of fish, upon which they live, they maintain a flow of water through the gills by alternate contractions of the gill pouches. When the gill-pouch muscles relax, the pouches expand, and water is sucked in. The water is forced out through the gills by muscular contraction; the branchial musculature apparently prevents reflux of the water into the pharynx while the head of the lamprey is embedded in the flesh of its prey.

In the hagfish *Myxine glutinosa*, the major oxygen supply is derived from water drawn in through the nostril that opens into the pharynx. A peculiar respiratory structure, the velum, just behind the nostril opening, dangles from the upper midline of the pharynx, resembling an inverted T. Membranous scrolls attached to this horizontal bar can extend downward and then roll upward like window shades. A combination of velar and gill-pouch contractions directs the flow of water through the gill pouches. Foreign material entering the nostril is expelled from both the mouth and nostril by a violent "sneeze." This reaction probably protects the respiratory surfaces, since the animals have common respiratory and alimentary ducts. Blood flow in the gills of cyclostomes, as in those of bony fishes, is in a direction counter to that of water flow—an arrangement that increases the efficiency of gas exchange across the respiratory surface (Figure 4).

Cartilaginous fishes (sharks and rays) and bony fishes employ a double-pumping mechanism to maintain a relatively constant flow of water over the gill exchange surfaces. In sharks and rays a small forward gill slit, the spiracle, also provides a channel for water flow into the gill chamber. Bottom-dwelling forms (*e.g.*, skates) have relatively larger spiracles, and the major portion of the water flow passes through them rather than through the downward-oriented mouth.

The pumping mechanism is not the only method of ventilation; sharks have been observed to keep both mouth and gill flaps open while swimming, ensuring a constant water flow across the gill surfaces. When they slow down or settle to the bottom, the pumping activity is resumed. Tunas and mackerel cannot stop swimming: They have no active respiratory mechanism and are dependent for their gill ventilation on the current that results from their forward motion through the water.

Aerial respiration in fishes

A number of fishes depend in varying degree on aerial respiration. The ability to breathe air enables them to live in places where the oxygen content of water may be low or

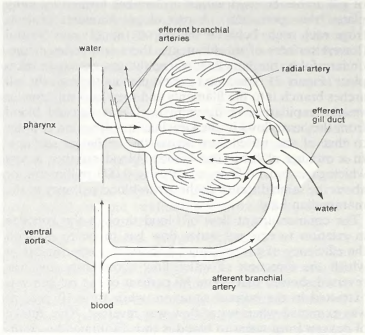


Figure 4: Blood and water flow through gill body of hagfish. Note that water flow is counter to blood flow, an arrangement facilitating gas exchange.

Reprinted with permission of the Macmillan Company from *Animal Function*, by M.S. Gordon. Copyright © 1968 by Malcolm S. Gordon

nil. Two general means of acquiring oxygen are employed. Some fishes stay near the surface of the water, where the oxygen pressure resulting from surface diffusion is highest. Others have developed ancillary respiratory structures in the pharynx or the stomach; the gulping of air at the surface is a means of charging these respiratory surfaces (such as the pharyngeal epithelium in *Electrophorus* or the stomach in *Plecostomus*). The frequency with which these fishes rise to the surface to gulp air corresponds to their current need for oxygen.

The swamp-dwelling *Erythrinus* of Guyana uses both aquatic and aerial respiration, varying them according to the gaseous composition of the water. When the oxygen content is low, respiration through the gills ceases; when the oxygen content of the water is high, the fish relies primarily upon its gills except when the carbon dioxide content is also high—when, again, aerial respiration predominates. In other conditions it uses both modes of respiration. This apparently extends the range of conditions in which *Erythrinus* can survive.

Eels and lungfishes

Eels (*Anguilla*) use their skin as a major respiratory surface in addition to their gills. In water, about 15 percent of their oxygen uptake is across the skin, and this rises to around 50 percent when in air. They are capable of making extensive overland migrations during which, in the first few hours, they draw upon oxygen in the swim bladder. Like most fishes, eels when out of water exhibit a reduced heart rate and less oxygen consumption. When they return to water, their heart rate rises, and both oxygen consumption and blood lactic-acid levels rise. Lactic-acid production results from metabolism without oxygen, and such acid products must themselves be metabolized through higher oxygen consumption. Such patterns have been observed in grunions of the California coast that come ashore to breed, and even in flying fishes during their brief aerial excursions.

The lungfishes (Dipnoi) are remnants of the Devonian period and a transitional form between water and air breathers. Like amphibians, they rely on the buccal force pump mechanism to inflate the lung. They are adapted for bimodal respiration so that oxygenated blood leaving the air-exchange organ, either gills or lung, can pass to the afferent branchial circulation and then to the body tissues or can be dispatched to the lungs (Figure 5). The three dipnoan genera differ with respect to their reliance on the gills or lungs. In the Australian lungfish (*Neoceratodus*), the bulk of oxygen uptake and carbon dioxide elimination is by way of well-developed gills; in the African lungfish (*Protopterus*) and the South American lungfish (*Lepidosiren*), the gills are reduced and ventilation depends heavily on the lungs. In the latter two, which rely primarily on lung ventilation, separation of oxygenated

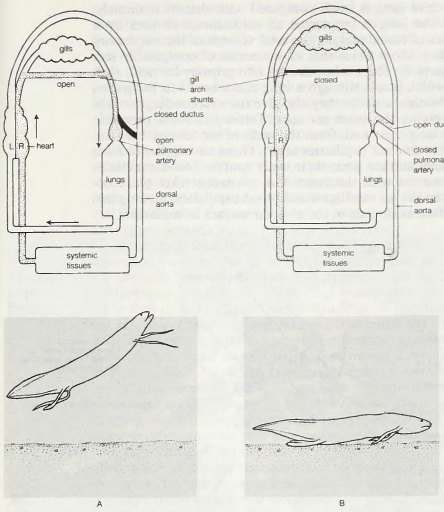


Figure 5: *Rearrangements of the circulation in a fish during (A) air breathing and (B) prolonged submersion.*
(A) During air breathing, the gill shunts are open, the ductus closed, and the pulmonary arterial vasomotor segment open. This arrangement clearly favours blood flow to the lungs.
(B) During prolonged submersion, the gill shunts are closed, the ductus open, and the pulmonary arterial vasomotor segment closed. This arrangement diverts blood flow from the pulmonary to systemic circulations.

Modified after A.P. Fishman, R.G. DeLaney, P. Laurent, and J.P. Szidon in K. Johansen and W.W. Burggren (eds.), *Cardiovascular Shunts: Phylogenetic, Ontogenetic, and Clinical Aspects*, p. 89; © 1985 Munksgaard International Publishers Ltd., Copenhagen, Denmark

and deoxygenated blood is much more complete than in the Australian lungfish.

Adaptation to drought

During long periods of drought, both *Protopterus* and *Lepidosiren* build a subterranean cocoon that opens to the surface via a thin tunnel. They then enter into a state of estivation in which metabolism, respiration, and heart rate fall to low levels. This state of diminished oxygen requirement enables the lungfish to remain viable without food or water for months or years, until the waters return.

The bowfin, *Amia calva,* has both gills and an air bladder that may be used for respiration. It is almost exclusively a water breather at 10° C (50° F), a temperature at which it shows low physical activity. Its air-breathing rate increases with temperature and activity, and, at around 30° C (86° F), it draws about three times as much oxygen from air as from water. As in lungfishes, carbon dioxide elimination is predominantly across the gills. The bowfin's air-breathing frequency varies inversely with the oxygen content of the water; when oxygen tensions in water decline below 40 or 50 millimetres of mercury at 20° C (68° F), air breathing largely replaces water breathing. When an exchange surface (gill or air bladder) is not being utilized as the primary oxygen-exchange site, there is a tendency for blood to bypass it.

The so-called electric eel of South America (*Electrophorus electricus*) inhabits muddy streams that may become severely oxygen deficient. It is an obligatory air breather that depends upon the exchange of oxygen across the membranes of its mouth, expelling the air through its gill slits. Its blood has a high percentage of red corpuscles, is high in hemoglobin, and has an oxygen-absorbing capacity similar to that of mammals. Carbon dioxide elimination is primarily across the skin and, to a lesser extent, through the vestigial gills. (F.N.W./A.P.F.)

Amphibians. The living amphibians (frogs, toads, salamanders, and caecilians) depend on aquatic respiration to a degree that varies with species, stage of development, temperature, and season. With the exception of a few frog species that lay eggs on land, all amphibians begin life as completely aquatic larvae. Respiratory gas exchange is conducted through the thin, gas-permeable skin and the gills. In addition to these structures, frog tadpoles use their large tail fins for respiration; the tail fins contain blood vessels and are important respiratory structures because of their large surface area. As amphibian larvae develop, the gills (and in frogs, the tail fin) degenerate, paired lungs develop, and the metamorphosing larvae begin making excursions to the water surface to take air breaths.

The lungs of amphibians (Figure 6) are simple saclike structures that internally lack the complex spongy appearance of the lungs of birds and mammals. The lungs of most amphibians receive a large proportion of the total blood flow from the heart. Even though the amphibian ventricle is undivided, there is surprisingly little mixture of blood from the left and right atrial chambers within the single ventricle. As a consequence, the lungs are perfused primarily with deoxygenated blood from the systemic tissues.

Variation from higher vertebrates

By the time the larva has reached adult form, the lungs have assumed the respiratory function of the larval gills. A few species of salamanders (for example, the axolotl) never metamorphose to the adult stage, and although they may develop lungs for air breathing, they retain external gills throughout life. Another exception to the usual pattern of respiratory development is seen in the Plethodontidae family of salamanders, which lose their gills upon metamorphosis but never develop lungs as adults; instead, gas exchange is conducted entirely across the skin. In almost all amphibian species, the skin in adults continues to play an important role in gas exchange.

The relative contributions of lungs and skin, and even local areas of skin, to gas exchange differ in different species and in the same species may change seasonally. In frogs, the skin of the back and thighs (the areas exposed to air) contains a richer capillary network than the skin of the underparts and therefore contributes more to gas exchange. The aquatic newt *Triton* utilizes both lung and skin respiration, the skin containing about 75 percent of the respiratory capillaries. At the other extreme, the tree frog *Hyla arborea* is much less aquatic, and its lungs contain over 75 percent of the respiratory capillary surface area. Similar differences are found even in closely related forms: In the relatively more terrestrial frog *Rana temporaria,* uptake of oxygen across the lung is about three times greater than across the skin; in *R. esculenta,* which is more restricted to water, the lungs and skin function about equally in the uptake of oxygen. Carbon dioxide is eliminated mainly through the skin in both these species; in fact, the skin appears to be a major avenue for carbon dioxide exchange in amphibians generally.

Seasonal changes in gas exchange

In temperate climates, as winter approaches, the colder environmental temperature (and thus lower body temperature) induces a marked lowering of the metabolic rate in amphibians. Terrestrial forms (*e.g.,* toads and some

From W. Burggren in S.C. Wood and C. Lenfant (eds.), *Comparative Pulmonary Physiology* (1988); Marcel Dekker, Inc., New York City, reprinted by courtesy of Marcel Dekker, Inc.

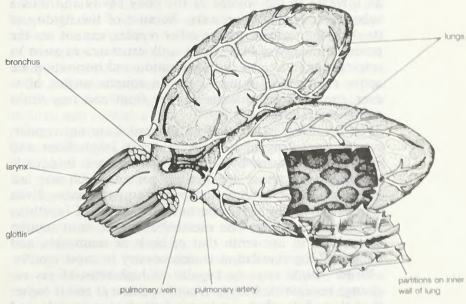


Figure 6: The lungs of a typical frog, with a portion of the lung wall removed to show the hollow, saclike structure.

salamanders) may burrow into the ground to overwinter. Aquatic species burrow into the mud at the bottom of lakes or ponds. Because their metabolic rate is much lower during winter, adequate gas exchange can be provided entirely by the skin in either terrestrial or aquatic habitats.

The mechanism of lung inflation in amphibians is the buccal cavity (mouth-throat) pumping mechanism that also functions in air-breathing fishes. To produce inspiration, the floor of the mouth is depressed, causing air to be drawn into the buccal cavity through the nostrils. The nostrils are then closed, and the floor of the mouth is elevated. This creates a positive pressure in the mouth cavity and drives air into the lungs through the open glottis. Expiration is produced by contraction of the muscles of the body wall and the elastic recoil of the lungs, both acting to drive gas out of the lungs through the open glottis. In aquatic amphibians the pressure of water on the body wall can also assist expiration. Many amphibians show rhythmic oscillations of the floor of the mouth between periods of lung inflation; these oscillations are thought to be involved in olfaction by producing a flow of gas over the olfactory epithelial surfaces.

Adaptation to land

Reptiles. To survive on land, the reptiles had to develop a skin relatively impermeable to water, so as to prevent dessication, and hence not well suited for respiration. Thus, while a few specialized reptiles (for example, sea snakes) can acquire nearly half of their oxygen supply through their skin, most reptiles depend almost entirely on the lungs for gas exchange. Reptilian lungs are considerably more complex than those of amphibians, showing much more internal partitioning to provide additional surface area for gas exchange between lung gas and blood. The most complex reptilian lungs are found in sea turtles such as *Chelonia mydas,* the green turtle. This species can develop a high metabolic rate associated with its prolific swimming ability. Its lungs are suited to providing a high rate of gas exchange, with extensive branching of the airways leading to the numerous gas sacs of the lungs.

The mechanism for lung inflation in reptiles is an aspiration (suction) pump, which is the same in general principle as the lung inflation mechanism in birds and mammals. In most reptiles inspiration is produced by muscular expansion of the rib cage and body wall, creating a subatmospheric pressure within the lungs that causes air to flow in. Crocodiles and alligators have a specialized muscle attached to the posterior surface of the liver; the anterior surface of the liver in turn is attached to the posterior surface of the lungs. Contraction of this muscle pulls on the liver and results in expansion of the lungs.

Respiration in turtles

The adoption of a rigid shell by turtles and tortoises necessitated the development of highly specialized skeletal muscles to inflate the lungs. In the tortoise *Testudo graeca,* lung ventilation is achieved by changing the volume of the body cavity. Expiration is brought about by the activity of muscles that draw the shoulder girdle back into the shell, compressing the abdominal viscera. The increased pressure in the body cavity is transmitted to the lungs. Inspiration involves opposite muscular actions that produce an increase in the volume of the body cavity and thus a subatmospheric lung pressure. Because of the rigidity of its shell, the tortoise, unlike other reptiles, cannot use the potential energy of abdominal wall structures to assist in respiration, and hence both expiration and inspiration are active energy-consuming events. In aquatic turtles, however, the pressure of water on the front and rear limbs assists expiration.

The breathing patterns of most reptiles are not regular, usually consisting of a series of active inspirations and expirations followed by relatively long pauses. In aquatic reptiles diving occurs during these pauses, which may last an hour or more in some turtles and aquatic snakes. Even terrestrial reptiles show intermittent periods of breathing and breath holding. The metabolic rate of most reptiles is one-fifth to one-tenth that of birds or mammals, and constant lung ventilation is unnecessary in most reptiles.

Birds. Birds must be capable of high rates of gas exchange because their oxygen consumption at rest is higher than that of all other vertebrates, including mammals, and it increases many times during flight. The gas volume of the bird lung is small compared with that of mammals, but the lung is connected to voluminous air sacs by a series of tubes, making the total volume of the respiratory system about twice that of mammals of comparable size (Figure 7). The trachea divides into primary bronchi, each of which passes through a lung and onward to the paired abdominal air sacs; they also give rise to secondary bronchi supplying the other air sacs. Tertiary bronchi penetrate the lung mass and, from the walls of the tertiary bronchi, rather fine air capillaries arise. These air capillaries have a large surface area; their walls contain blood capillaries connected with the heart. Gas exchange takes place between the air capillaries and blood capillaries, making this surface analogous to the alveolar surface in mammals.

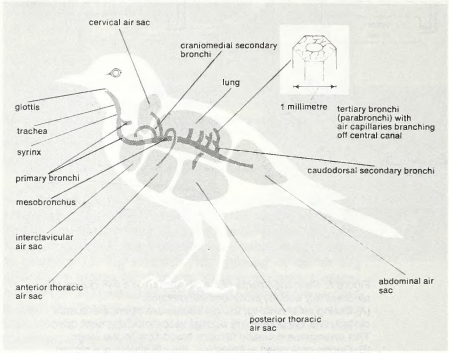


Figure 7: Respiratory system of a bird.

There are several important differences in the mechanism and pattern of lung ventilation in birds compared with other vertebrates with lungs. The lungs of birds do not inflate and deflate but rather retain a constant volume. Also, the lungs are unidirectionally ventilated rather than having a tidal, bidirectional flow, as in other vertebrates with lungs. To achieve this unidirectional flow, the various air sacs are inflated and deflated in a complex sequence, like a series of interconnected bellows. The lungs, which are located midway between air sacs in terms of the flow of gas, are continuously ventilated in a single direction with freshly inspired air during both inspiration and expiration at the nostrils. Aspiration into the air sacs is produced by expansion of the chest and abdominal cavity. The sternum (breastbone) swings forward and downward, while the ribs and chest wall move laterally. Expiration is caused by compression of the air sacs by skeletal muscle.

Unidirectional airflow in birds

Efficiency of bird ventilation

As a consequence of the continual, unidirectional airflow, the lungs of birds are more completely ventilated than the lungs of mammals. The flow of gas and blood within the bird lung is carefully arranged to maximize gas exchange, which is far more efficient than in the mammalian lung: Himalayan geese have been observed not only to fly over human climbers struggling to reach the top of Mount Everest, but to honk as they do so. The ventilation of pigeons increases around 20-fold during flight, brought about by more rapid breathing and not by taking in more air at a breath. There is a precise synchrony between breathing and wing motion: the peak of expiration occurs at the downstroke of the wingbeat. The pigeon's in-flight ventilation is about two and one-half times that needed to support metabolism; around 17 percent of the heat production during flight is lost through evaporative cooling, suggesting that the excess ventilation is for regulating body heat. Studies of evening grosbeaks and ring-billed gulls show that their ventilation, in contrast to that of pigeons, increases in proportion to oxygen consumption. The increased ventilation in these birds is brought about by deeper as well as by more rapid breathing.

The respiratory system of birds is also used for commu-

nication through song. The "voice box" is the syrinx, a membranous structure at the lower end of the trachea. Sound is produced only when air flows outward across the syrinx. In canaries, notes or pulses are synchronous with chest movements; the trills, however, are made with a series of shallow breaths. The song of many small birds is of long duration relative to their breathing frequencies.

Mammals. To provide the gas exchange necessary to support the elevated metabolic rate of mammals, mammalian lungs are subdivided internally. The repetitive subdivisions of the lung airways provide gas to the tiny alveoli (gas sacs) that form the functional gas-exchange surface area of the lungs. Human lungs have an estimated 300,000,000 alveoli, providing in an adult a total surface area approximately equivalent to a tennis court.

Ventilatory pump in humans

Inspiration in mammals, as in reptiles, is powered by an aspiration (suction) pump. Expansion of the chest lowers the pressure between the lungs and the chest wall, as well as the pressure within the lungs. This causes atmospheric air to flow into the lungs. The chief muscles of inspiration are the diaphragm and the external intercostal muscles. The diaphragm is a domelike sheet of muscle separating the abdominal and chest cavities that moves downward as it contracts. The downward motion enlarges the chest cavity and depresses the organs below. As the external intercostal muscles contract, the ribs rotate upward and laterally, increasing the chest circumference. During severe exercise other muscles may also be used. Inspiration ends with the closing of the glottis.

In expiration, the glottis opens, and the inspiratory muscles relax; the stored energy of the chest wall and lungs generates the motive power for expiration. During exercise or when respiration is laboured, the internal intercostal muscles and the abdominal muscles are activated. The internal intercostals produce a depression of the rib cage and a decrease in chest circumference. (F.N.W./W.W.Bu.)

Human respiration

THE DESIGN OF THE RESPIRATORY SYSTEM

The human gas exchanging organ, the lung, is located in the thorax, where its delicate tissues are protected by the bony and muscular thoracic cage. The lung provides the organism with a continuous flow of oxygen and clears the blood of the gaseous waste product, carbon dioxide. Atmospheric air is pumped in and out regularly through a system of pipes, called conducting airways, which join the gas exchange region with the outside of the body (Figure 8). The airways can be divided into upper and lower airway systems. The transition between the two systems is located where the pathways of the respiratory and digestive systems cross, just at the top of the larynx.

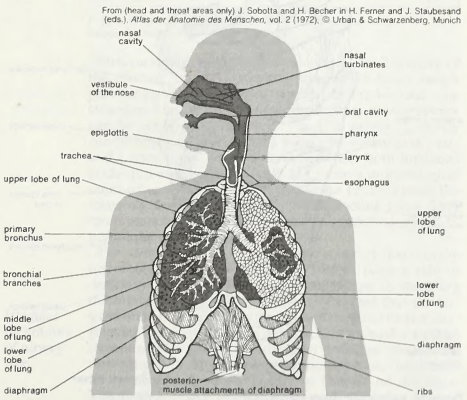


Figure 8: The respiratory system and its location.

Upper and lower airways

The upper airway system comprises the nose and the paranasal cavities, called sinuses, the pharynx, or throat, and partly also the oral cavity, since it may be used for breathing. The lower airway system consists of the larynx, the trachea, the stem bronchi, and all the airways ramifying intensively within the lungs, such as the intrapulmonary bronchi, the bronchioles, and the alveolar ducts. For respiration, the collaboration of other organ systems is clearly essential. The diaphragm, as the main respiratory muscle, and the intercostal muscles of the chest wall play an essential role by generating, under the control of the central nervous system, the pumping action on the lung. The muscles expand and contract the internal space of the thorax, whose bony framework is formed by the ribs and the thoracic vertebrae. The contribution of the lung and chest wall (ribs and muscles) to respiration is described below in *The mechanics of breathing.* The blood, as a carrier for the gases, and the circulatory system (*i.e.,* the heart and the blood vessels) are mandatory elements of a working respiratory system (*see* BLOOD and CIRCULATION AND CIRCULATORY SYSTEMS).

Morphology of the upper airways. *The nose.* The nose is the external protuberance of an internal space, the nasal cavity (Figure 9). It is subdivided into a left and right canal by a thin medial cartilaginous and bony wall, the nasal septum. Each canal opens to the face by a nostril and into the pharynx by the choana. The floor of the nasal cavity is formed by the palate, which also forms the roof of the oral cavity. The complex shape of the nasal cavity is due to projections of bony ridges, the superior, middle, and inferior turbinate bones (or conchae), from the lateral wall. The passageways thus formed below each ridge are called the superior, middle, and inferior nasal meatuses.

On each side, the intranasal space communicates with a series of neighbouring air-filled cavities within the skull (the paranasal sinuses) and also, via the nasolacrimal duct, with the lacrimal apparatus in the corner of the eye. The duct drains the lacrimal fluid into the nasal cavity. This fact explains why nasal respiration can be rapidly impaired or even impeded during weeping: the lacrimal fluid is not only overflowing into tears, it is also flooding the nasal cavity.

Paranasal sinuses

The paranasal sinuses are sets of paired single or multiple cavities of variable size. Most of their development takes place after birth, and they reach their final size toward the age of 20 years. The sinuses are located in four different skull bones—the maxilla, the frontal, the ethmoid, and the sphenoid bones. Correspondingly, they are called the maxillary sinus, which is the largest cavity; the frontal sinus; the ethmoid sinuses; and the sphenoid sinus, which is located in the upper posterior wall of the nasal cavity. The sinuses have two principal functions: because they are filled with air, they help keep the weight of the skull within reasonable limits, and they serve as resonance chambers for the human voice.

The nasal cavity with its adjacent spaces is lined by a respiratory mucosa. Typically, the mucosa of the nose contains mucus-secreting glands and venous plexuses; its top cell layer, the epithelium, consists principally of two cell types, ciliated and secreting cells. This structural design reflects the particular ancillary functions of the nose and of the upper airways in general with respect to respiration. They clean, moisten, and warm the inspired air, preparing it for intimate contact with the delicate tissues of the gas-exchange area. During expiration through the nose, the air is dried and cooled, a process that saves water and energy.

Two regions of the nasal cavity have a different lining. The vestibule, at the entrance of the nose, is lined by skin that bears short thick hairs called vibrissae. In the roof of the nose, the olfactory organ with its sensory epithelium checks the quality of the inspired air. About two dozen olfactory nerves convey the sensation of smell from the olfactory cells through the bony roof of the nasal cavity to the central nervous system.

The pharynx. For the anatomical description, the pharynx can be divided into three floors (see Figure 9). The upper floor, the nasopharynx, is primarily a passageway for air and secretions from the nose to the oral pharynx. It is also connected to the tympanic cavity of the middle ear

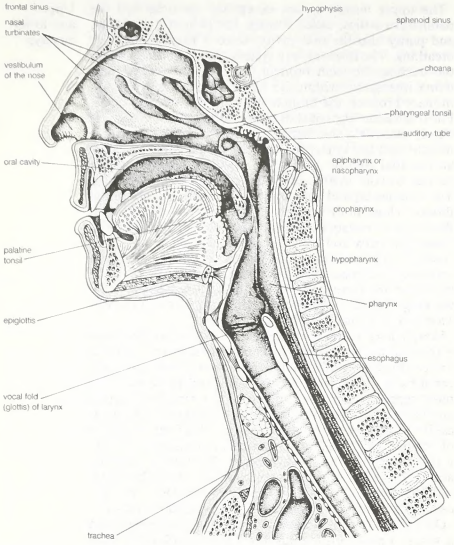


Figure 9: Upper airways with nasal and oral cavities, pharynx, and larynx.

From E. Pernkopf in H. Ferner (ed.): *Atlas der Topographischen und angewandten Anatomie des Menschen*, vol. 1 (1960). © Urban & Schwarzenberg, Munich

through the auditory tubes that open on both lateral walls. The act of swallowing opens briefly the normally collapsed auditory tubes and allows the middle ears to be aerated and pressure differences to be equalized. In the posterior wall of the nasopharynx is located a lymphatic organ, the pharyngeal tonsil. When it is enlarged (as in tonsil hypertrophy or adenoid vegetation), it may interfere with nasal respiration and alter the resonance pattern of the voice.

Middle and lower floors

The middle floor of the pharynx connects anteriorly to the mouth and is therefore called the oral pharynx or oropharynx. It is delimited from the nasopharynx by the soft palate, which roofs the posterior part of the oral cavity.

The lower floor of the pharynx is called the hypopharynx. Its anterior wall is formed by the posterior part of the tongue. Lying directly above the larynx, it represents the site where the pathways of air and food cross each other: Air from the nasal cavity flows into the larynx, and food from the oral cavity is routed to the esophagus directly behind the larynx. The epiglottis, a cartilaginous, leaf-shaped flap, functions as a lid to the larynx and, during the act of swallowing, controls the traffic of air and food.

Morphology of the lower airways. *The larynx.* The larynx is an organ of complex structure that serves a dual function: as an air canal to the lungs and a controller of its access, and as the organ of phonation (see Figure 10). Sound is produced by forcing air through a sagittal slit formed by the vocal cords, the glottis. This causes not only the vocal cords but also the column of air above them to vibrate. As evidenced by trained singers, this function can be closely controlled and finely tuned. Control is achieved by a number of muscles innervated by the laryngeal nerves. For the precise function of the muscular apparatus, the muscles must be anchored to a stabilizing framework (Figure 10). The laryngeal skeleton consists of almost a dozen pieces of cartilage, most of them very small, interconnected by ligaments and membranes. The largest cartilage of the larynx, the thyroid cartilage, is made of two plates fused anteriorly in the midline. At the upper end of the fusion line is an incision, the thyroid notch; below it is a forward projection, the laryngeal prominence. Both of these structures are easily felt through the skin. The angle between the two cartilage plates is sharper and the prominence more marked in men than in women, which has given this structure the common name of Adam's apple. Behind the shieldlike thyroid cartilage, the vocal cords span the laryngeal lumen. They correspond to elastic ligaments attached anteriorly in the angle of the thyroid shield and posteriorly to a pair of small pyramidal pieces of cartilage, the arytenoid cartilages. The vocal ligaments are part of a tube, resembling an organ pipe, made of elastic tissue. Just above the vocal cords, the epiglottis is also attached to the back of the thyroid plate by its stalk. The cricoid, another large cartilaginous piece of the laryngeal skeleton, has a signet-ring shape. The broad plate of the ring lies in the posterior wall of the larynx and the narrow arch in the anterior wall. The cricoid is located below the thyroid cartilage, to which it is joined in an articulation reinforced by ligaments. The transverse axis of the joint allows a hingelike rotation between the two cartilages. This movement tilts the cricoid plate with respect to the shield of the thyroid cartilage and hence alters the distance between them. Because the arytenoid cartilages rest upright on the cricoid plate, they follow its tilting movement. This mechanism plays an important role in altering length and tension of the vocal cords. The arytenoid cartilages articulate with the cricoid plate and hence are able to rotate and slide to close and open the glottis.

Adam's apple

Viewed frontally, the lumen of the laryngeal tube has an hourglass shape, with its narrowest width at the glottis. Just above the vocal cords there is an additional pair of mucosal folds called the false vocal cords or the vestibular folds. Like the true vocal cords, they are also formed by the free end of a fibroelastic membrane. Between the vestibular folds and the vocal cords, the laryngeal space enlarges and forms lateral pockets extending upward. This space is called the ventricle of the larynx. Because the gap between the vestibular folds is always larger than the gap between the vocal cords, the latter can easily be seen from above with the laryngoscope, an instrument designed for visual inspection of the interior of the larynx.

Muscular apparatus

The muscular apparatus of the larynx comprises two functionally distinct groups. The intrinsic muscles act directly or indirectly on the shape, length, and tension of the vocal cords. The extrinsic muscles act on the larynx as a whole, moving it upward (*e.g.*, during high-pitched phonation or swallowing) or downward. The intrinsic muscles attach to the skeletal components of the larynx itself; the

From J.W. Rohen and C. Yokochi: *Anatomie des Menschen*, vol. 1 (1982). © F.K. Schattauer Verlag, Stuttgart, and Igaku-Shoin Ltd., Tokyo

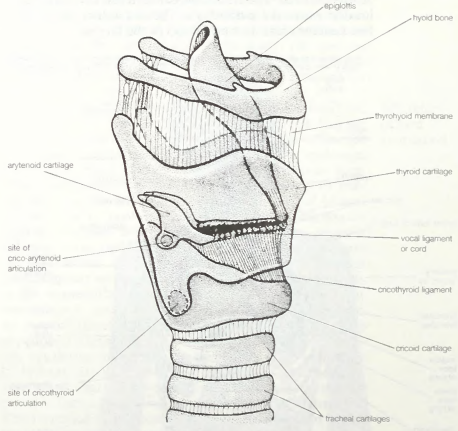


Figure 10: Larynx with major cartilages, hyoid bone, and vocal cords.

extrinsic muscles join the laryngeal skeleton cranially to the hyoid bone or to the pharynx and caudally to the sternum (breastbone).

Size and shape

The trachea and the stem bronchi. Below the larynx lies the trachea, a tube about 10 to 12 centimetres long and two centimetres wide. Its wall is stiffened by 16 to 20 characteristic horseshoe-shaped, incomplete cartilage rings that open toward the back and are embedded in a dense connective tissue. The dorsal wall contains a strong layer of transverse smooth muscle fibres that spans the gap of the cartilage. The interior of the trachea is lined by the typical respiratory epithelium. The mucosal layer contains mucous glands.

At its lower end, the trachea divides in an inverted Y into the two stem (or main) bronchi, one each for the left and right lung. The right main bronchus has a larger diameter, is oriented more vertically, and is shorter than the left main bronchus. The practical consequence of this arrangement is that foreign bodies passing beyond the larynx will usually slip into the right lung. The structure of the stem bronchi closely matches that of the trachea.

After E.R. Weibel in A.P. Fishman *Pulmonary Diseases and Disorders*, copyright © 1980 by McGraw-Hill, Inc.: all rights reserved

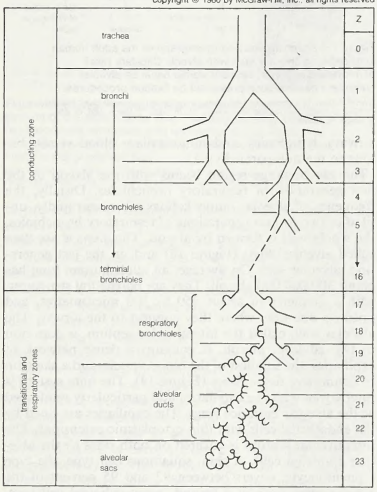


Figure 11: Model of airway branching in the human lung, assuming regular dichotomy (see text).

Structural design of the airway tree. The hierarchy of the dividing airways, and partly also of the blood vessels penetrating the lung, largely determines the internal lung structure. Functionally the intrapulmonary airway system can be subdivided into three zones, a proximal, purely conducting zone, a peripheral, purely gas-exchanging zone, and a transitional zone in between, where both functions grade into one another (Figure 11). From a morphological point of view, however, it makes sense to distinguish the relatively thick-walled, purely air-conducting tubes from those branches of the airway tree structurally designed to permit gas exchange.

The structural design of the airway tree is functionally important because the branching pattern plays a role in determining air flow and particle deposition. In modeling the human airway tree, it is generally agreed that the airways branch according to the rules of irregular dichotomy.

Irregular dichotomy of airway tree

Regular dichotomy means that each branch of a treelike structure gives rise to two daughter branches of identical dimensions. In irregular dichotomy, however, the daughter branches may differ greatly in length and diameter. The models calculate the average path from the trachea to the lung periphery as consisting of about 24–25 generations of branches. Individual paths, however, may range from 11 to 30 generations. The transition between the conductive and the respiratory portions of an airway lies on average at the end of the 16th generation, if the trachea is counted as generation 0. The conducting airways comprise the trachea, the two stem bronchi, the bronchi, and the bronchioles. Their function is to further warm, moisten, and clean the inspired air and distribute it to the gas-exchanging zone of the lung. They are lined by the typical respiratory epithelium with ciliated cells and numerous interspersed mucus-secreting goblet cells. Ciliated cells are present far down in the airway tree, their height decreasing with the narrowing of the tubes, as does the frequency of goblet cells. In bronchioles the goblet cells are completely replaced by another type of secretory cells named Clara cells. The epithelium is covered by a layer of low-viscosity fluid, within which the cilia exert a synchronized, rhythmic beat directed outward. In larger airways, this fluid layer is topped by a blanket of mucus of high viscosity. The mucus layer is dragged along by the ciliary action and carries the intercepted particles toward the pharynx, where they are swallowed. This design can be compared to a conveyor belt for particles, and indeed the mechanism is referred to as the mucociliary escalator.

Conducting airways

Whereas cartilage rings or plates provide support for the walls of the trachea and bronchi, the walls of the bronchioles, devoid of cartilage, gain their stability from their structural integration into the gas-exchanging tissues. The last purely conductive airway generations in the lung are the terminal bronchioles. Distally, the airway structure is greatly altered by the appearance of cuplike outpouchings from the walls (see Figure 11). These form minute air chambers and represent the first gas-exchanging alveoli on the airway path. In the alveoli, the respiratory epithelium gives way to a very flat lining layer that permits the formation of a thin air–blood barrier. After several generations (Z) of such respiratory bronchioles, the alveoli are so densely packed along the airway that an airway wall proper is missing; the airway consists of alveolar ducts. The final generations of the airway tree end blindly in the alveolar sacs.

The lungs. *Gross anatomy.* The organ lung is parted into two slightly unequal portions, a left lung and a right lung, which occupy most of the intrathoracic space (see Figure 8). The space between them is filled by the mediastinum, which corresponds to a connective tissue space containing the heart, major blood vessels, the trachea with the stem bronchi, the esophagus, and the thymus gland. The right lung represents 56 percent of the total lung volume and is composed of three lobes, a superior, middle, and inferior lobe, separated from each other by a deep horizontal and an oblique fissure. The left lung, smaller in volume because of the asymmetrical position of the heart, has only two lobes separated by an oblique fissure. In the thorax, the two lungs rest with their bases on the diaphragm, while their apexes extend above the first rib. Medially, they are connected with the mediastinum at the hilum, a circumscribed area where airways, blood and lymphatic vessels, and nerves enter or leave the lungs. The inside of the thoracic cavities and the lung surface are covered with serous membranes, respectively the parietal pleura and the visceral pleura, which are in direct continuity at the hilum. Depending on the subjacent structures, the parietal pleura can be subdivided into three portions: the mediastinal, costal, and diaphragmatic pleurae. The lung surfaces facing these pleural areas are named accordingly, since the shape of the lungs is determined by the shape of the pleural cavities. Because of the presence of pleural recesses, which form a kind of reserve space, the pleural cavity is larger than the lung volume (Figure 12).

Right and left lung volumes

During inspiration, the recesses are partly opened by the expanding lung, thus allowing the lung to increase in volume. Although the hilum is the only place where the lungs are secured to surrounding structures, the lungs are maintained in close apposition to the thoracic wall by a negative pressure between visceral and parietal pleurae. A thin film of extracellular fluid between the pleurae enables the lungs to move smoothly along the walls of the cavity during breathing. If the serous membranes become in-

flamed (pleurisy), respiratory movements can be painful. If air enters a pleural cavity (pneumothorax), the lung immediately collapses owing to its inherent elastic properties, and breathing is abolished on this side.

Pulmonary segments. The lung lobes are subdivided into smaller units, the pulmonary segments. There are 10 segments in the right lung and, depending on the classification, eight to 10 segments in the left lung. Unlike the lobes, the pulmonary segments are not delimited from each other by fissures but by thin membranes of connective tissue containing veins and lymphatics; the arterial supply follows the segmental bronchi. These anatomical features are important because pathological processes may be limited to discrete units, and the surgeon can remove single diseased segments instead of whole lobes.

Structure

The intrapulmonary conducting airways: bronchi and bronchioles. In the intrapulmonary bronchi, the cartilage rings of the stem bronchi are replaced by irregular cartilage plates; furthermore, a layer of smooth muscle is added between the mucosa and the fibrocartilaginous tunic. The bronchi are ensheathed by a layer of loose connective tissue that is continuous with the other connective tissue elements of the lung and hence is part of the fibrous skeleton spanning the lung from the hilum to the pleural sac. This outer fibrous layer contains, besides lymphatics and nerves, small bronchial vessels to supply the bronchial wall with blood from the systemic circulation. Bronchioles are small conducting airways ranging in diameter from three to less than one millimetre. The walls of the bronchioles lack cartilage and seromucous glands. Their lumen is lined by a simple cuboidal epithelium with ciliated cells and Clara cells, which produce a chemically ill-defined secretion. The bronchiolar wall also contains a well-developed layer of smooth muscle cells, capable of narrowing the airway. Abnormal spasms of this musculature cause the clinical symptoms of bronchial asthma.

The gas-exchange region. The gas-exchange region comprises three compartments: air, blood, and tissue. Whereas air and blood are continuously replenished, the function of the tissue compartment is twofold: it provides the stable supporting framework for the air and blood compartments, and it allows them to come into close contact with each other (thereby facilitating gas exchange) while keeping them strictly confined. The respiratory gases diffuse from air to blood, and vice versa, through the 140 square metres of internal surface area of the tissue compartment.

Parenchyma

The gas-exchange tissue proper is called the pulmonary parenchyma, while the supplying structures, conductive airways, lymphatics, and non-capillary blood vessels belong to the non-parenchyma.

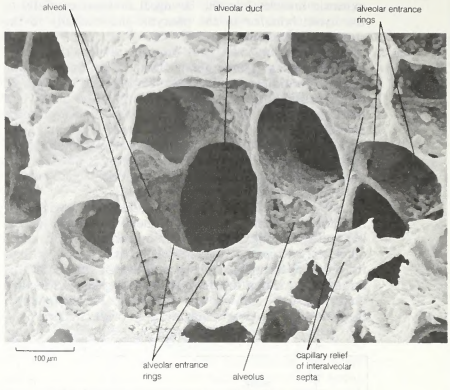


Figure 13: Scanning electron micrograph of the adult human lung showing alveolar duct with alveoli. Capillary relief of interalveolar septa is clearly visible because alveolar surfactant has not been preserved by fixation procedures.

From P.H. Burri, "Morphology and Respiratory Function of the Alveolar Unit," *International Archives of Allergy and Applied Immunology*, no. 76, suppl. 1, March 1985. © 1985, S. Karger AG, Basel

The gas-exchange region begins with the alveoli of the first generation of respiratory bronchioles. Distally, the frequency of alveolar outpocketings increases rapidly, until after two to four generations of respiratory bronchioles, the whole wall is formed by alveoli. The airways are then called alveolar ducts (Figure 13) and, in the last generation, alveolar sacs. On average, an adult human lung has about 300,000,000 alveoli. They are polyhedral structures, with a diameter of about 250 to 300 micrometres, and open on one side, where they connect to the airway. The alveolar wall, called the interalveolar septum, is common to two adjacent alveoli. It contains a dense network of capillaries, the smallest of the blood vessels, and a skeleton of connective tissue fibres (Figure 14). The fibre system is interwoven with the capillaries and particularly reinforced at the alveolar entrance rings. The capillaries are lined by flat endothelial cells with thin cytoplasmic extensions. The interalveolar septum is covered on both sides by the alveolar epithelial cells. A thin, squamous cell type, the type I pneumocyte, covers between 92 and 95 percent of the gas-exchange surface; a second, more cuboidal cell type, the type II pneumocyte, covers the remaining surface. The type I cells form, together with the endothelial cells, the thin air–blood barrier for gas exchange; the type II cells are secretory cells. Type II pneumocytes produce a surface-tension-reducing material, the pulmonary surfactant, which spreads on the alveolar surface and prevents the tiny alveolar spaces from collapsing. Before it is released into the airspaces, pulmonary surfactant is stored in the type II cells in the form of lamellar bodies. These granules are the conspicuous ultrastructural features of this cell type. On top of the epithelium, alveolar macrophages creep around within the surfactant fluid. They are large cells, and their cell bodies abound in granules of various content, partly foreign material that may have reached the alveoli, or cell debris originating from cell damage or normal cell death. Ultimately, the alveolar macrophages are derived from the bone marrow, and their task is to keep the air–blood barrier clean and unobstructed. The tissue space between the endothelium of the capillaries and the epithelial lining is occupied by the interstitium. It contains connective tissue and interstitial fluid. The connective tissue comprises a system of fibres, amorphous ground substance, and cells (mainly fibroblasts), which seem to be endowed with contractile properties. The fibroblasts are thought to control capillary blood flow or, alternatively,

Pulmonary surfactant

After E.R. Weibel in A.P. Fishman, *Pulmonary Diseases and Disorders*, copyright © 1980 by McGraw-Hill, Inc., all rights reserved

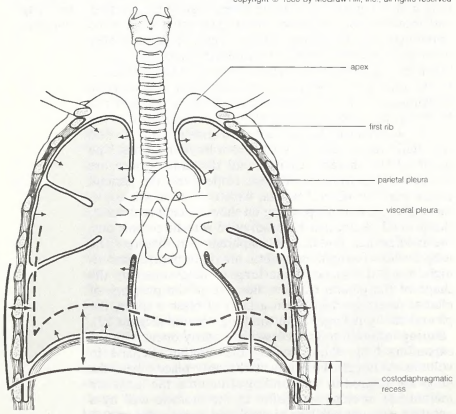


Figure 12: *Frontal section of chest.*
Pleural space and lobar fissures are shown. Single arrows represent retractive forces. Double arrows indicate the expansion of the lung into recesses during deep inspiration.

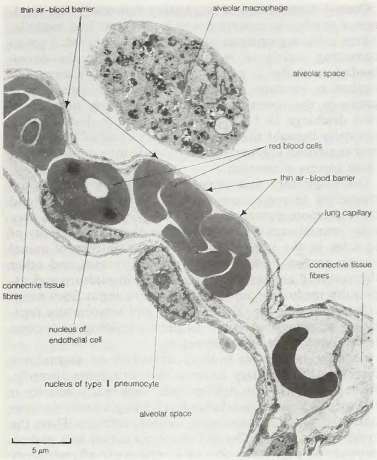


Figure 14: Electron micrograph of part of the interalveolar septum of the adult human lung. The lung capillary contains red blood cells interlaced with connective tissue fibres. The alveolar macrophage normally lies within the surfactant layer, which is not preserved here.

From P.H. Burri, "Lung Development and Growth," in A.P. Fishman and A.B. Fisher (eds.), *The Respiratory System*, vol. 1. © copyright 1985, American Physiological Society

to prevent the accumulation of extracellular fluid in the interalveolar septa. If for some reason the delicate fluid balance of the pulmonary tissues is impaired, an excess of fluid accumulates in the lung tissue and within the airspaces. This pathological condition is called pulmonary edema. As a consequence, the respiratory gases must diffuse across longer distances, and proper functioning of the lung is severely jeopardized.

Blood vessels, lymphatic vessels, and nerves. With respect to blood circulation, the lung is a complex organ. It has two distinct though not completely separate vascular systems: a low-pressure pulmonary system and a high-pressure bronchial system. The pulmonary (or lesser) circulation is responsible for the oxygen supply of the organism. Blood, low in oxygen content but laden with carbon dioxide, is carried from the right heart through the pulmonary arteries to the lungs. On each side, the pulmonary artery enters the lung in the company of the stem bronchus and then divides rapidly, following relatively closely the course of the dividing airway tree. After numerous divisions, small arteries accompany the alveolar ducts and split up into the alveolar capillary networks. Because intravascular pressure determines the arterial wall structure, the pulmonary arteries, which have on average a pressure five times lower than systemic arteries, are much flimsier than systemic arteries of corresponding size. The oxygenated blood from the capillaries is collected by venules and drained into small veins. These do not accompany the airways and arteries but run separately in narrow strips of connective tissue delimiting small lobules. The interlobular veins then converge on the intersegmental septa. Finally, near the hilum the veins merge into large venous vessels that follow the course of the bronchi. Generally, four pulmonary veins drain blood from the lung and deliver it to the left atrium of the heart.

Lung vascular systems

The bronchial circulation has a nutritional function for the walls of the larger airways and pulmonary vessels. The bronchial arteries originate from the aorta or from an intercostal artery. They are small vessels and generally do not reach as far into the periphery as the conducting airways. With a few exceptions, they end several generations short of the terminal bronchioles. They split up into capillaries surrounding the walls of bronchi and vessels and also supply adjacent airspaces. Most of their blood is naturally collected by pulmonary veins. Small bronchial veins exist, however; they originate from the peribronchial venous plexuses and drain the blood through the hilum into the azygos and hemiazygos veins of the posterior thoracic wall.

The lymph is drained from the lung through two distinct but interconnected sets of lymphatic vessels. The superficial, subpleural lymphatic network collects the lymph from the peripheral mantle of lung tissue and drains it partly along the veins toward the hilum. The deep lymphatic system originates around the conductive airways and arteries and converges into vessels that mostly follow the bronchi and arterial vessels into the mediastinum.

Filtering action of lymph nodes

Within the lung and the mediastinum, lymph nodes exert their filtering action on the lymph before it is returned into the blood through the major lymphatic vessels, called bronchomediastinal trunks. Lymph drainage paths from the lung are complex. The precise knowledge of their course is clinically relevant, because malignant tumours of the lung spread via the lymphatics.

The pleurae, the airways, and the vessels are innervated by afferent and efferent fibres of the autonomic nervous system. Parasympathetic nerve fibres from the vagus nerve (10th cranial nerve) and sympathetic branches of the sympathetic nerve trunk meet around the stem bronchi to form the pulmonary autonomic nerve plexus, which penetrates into the lung along the bronchial and vascular walls. The sympathetic fibres mediate a vasoconstrictive action in the pulmonary vascular bed and a secretomotor activity in the bronchial glands. The parasympathetic fibres stimulate bronchial constriction. Afferent fibres to the vagus nerve transmit information from stretch receptors, and those to the sympathetic centres carry sensory information (*e.g.*, pain) from the bronchial mucosa.

Lung development. After early embryogenesis, during which the lung primordium is laid down, the developing human lung undergoes four consecutive stages of development, ending after birth. The names of the stages describe the actual morphology of the prospective airways. The pseudoglandular stage exists from five to 17 weeks; the canalicular stage, from 16 to 26 weeks; the saccular stage, from 24 to 38 weeks; and finally the alveolar stage, from 36 weeks of fetal age to about 1½ to two years after birth.

First development

The lung appears around the 26th day of intrauterine life as a ventral bud of the prospective esophagus. The bud separates distally from the gut, divides, and starts to grow into the surrounding mesenchyme. The epithelial components of the lung are thus derived from the gut (*i.e.*, they are of endodermal origin), and the surrounding tissues and the blood vessels are derivatives of the mesoderm.

Following rapid successive dichotomous divisions, the lung begins to look like a gland, giving the first stage of development (pseudoglandular) its name. At the same time the vascular connections also develop and form a capillary plexus around the lung tubules. Toward week 17, all the conducting airways of the lung are preformed, and it is assumed that, at the outermost periphery, the tips of the tubules represent the first structures of the prospective gas-exchange region.

During the canalicular stage, the future lung periphery develops further. The prospective airspaces enlarge at the expense of the intervening mesenchyme, and their cuboidal epithelium differentiates into type I and type II epithelial cells or pneumocytes. Toward the end of this stage, areas with a thin prospective air–blood barrier have developed, and surfactant production has started. These structural and functional developments give a prematurely born fetus a small chance to survive at this stage.

During the saccular stage, further generations of airways are formed. The tremendous expansion of the prospective respiratory airspaces causes the formation of saccules and a marked decrease in the interstitial tissue mass. The lung looks more and more "aerated," although it is filled with fluid originating from the lungs and from the amniotic fluid surrounding the fetus. Some weeks before birth, alveolar formation begins by a septation process that subdivides the saccules into alveoli. At this stage of lung development, the infant is born.

Changes at birth

At birth the intrapulmonary fluid is rapidly evacuated and the lung fills with air with the first breaths. Simultaneously, the pulmonary circulation, which before was practically bypassed and very little perfused, opens up to accept the full cardiac output.

The newborn lung is far from being a miniaturized version of the adult lung. It has only about 20,000,000 to 50,000,000 alveoli, or 6 to 15 percent of the full adult complement. Therefore, alveolar formation is completed in the early postnatal period. Although it was previously thought that alveolar formation could continue to the age of eight years and beyond, it is now accepted that the bulk of alveolar formation is concluded much earlier, probably before the age of two years. Even with complete alveolar formation, the lung is not yet mature. The newly formed interalveolar septa still contain a double capillary network instead of the single one of the adult lungs. This means that the pulmonary capillary bed must be completely reorganized during and after alveolar formation; it has to mature. Only after full microvascular maturation, which is terminated sometime between the ages of two and five years, is the lung development completed, and the lung can enter a phase of normal growth. (P.H.B.)

CONTROL OF BREATHING

Breathing is an automatic and rhythmic act produced by networks of neurons in the hindbrain (the pons and medulla). The neural networks direct muscles that form the walls of the thorax and abdomen and produce pressure gradients that move air into and out of the lungs. The respiratory rhythm and the length of each phase of respiration are set by reciprocal stimulatory and inhibitory interconnection of these brain-stem neurons.

An important characteristic of the human respiratory system is its ability to adjust breathing patterns to changes in both the internal milieu and the external environment. Ventilation increases and decreases in proportion to swings in carbon dioxide production and oxygen consumption caused by changes in metabolic rate. The respiratory system is also able to compensate for disturbances that affect the mechanics of breathing, such as the airway narrowing that occurs in an asthmatic attack. Breathing also undergoes appropriate adjustments when the mechanical advantage of the respiratory muscles is altered by postural changes or by movement.

This flexibility in breathing patterns in large part arises from sensors distributed throughout the body that send signals to the respiratory neuronal networks in the brain. Chemoreceptors detect changes in blood oxygen levels and change the acidity of the blood and brain. Mechanoreceptors monitor the expansion of the lung, the size of the airway, the force of respiratory muscle contraction, and the extent of muscle shortening.

Although the diaphragm is the major muscle of breathing, its respiratory action is assisted and augmented by a complex assembly of other muscle groups. Intercostal muscles inserting on the ribs, the abdominal muscles, and muscles such as the scalene and sternocleidomastoid that attach both to the ribs and to the cervical spine at the base of the skull also play an important role in the exchange of air between the atmosphere and the lungs. In addition, laryngeal muscles and muscles in the oral and nasal pharynx adjust the resistance of movement of gases through the upper airways during both inspiration and expiration. Although the use of these different muscle groups adds considerably to the flexibility of the breathing act, they also complicate the regulation of breathing. These same muscles are used to perform a number of other functions, such as speaking, chewing and swallowing, and maintaining posture. Perhaps because the "respiratory" muscles are employed in performing nonrespiratory functions, breathing can be influenced by higher brain centres and even controlled voluntarily to a substantial degree. An outstanding example of voluntary control is the ability to suspend breathing by holding one's breath. Input into the respiratory control system from higher brain centres may help optimize breathing so that not only are metabolic demands satisfied by breathing but ventilation also is accomplished with minimal use of energy.

Non-respiratory functions of respiratory muscles

Central organization of respiratory neurons. The respiratory rhythm is generated within the pons and medulla. Three main aggregations of neurons are involved: a group consisting mainly of inspiratory neurons in the dorsomedial medulla, a group made up of inspiratory and expiratory neurons in the ventrolateral medulla, and a group in the rostral pons consisting mostly of neurons that discharge in both inspiration and expiration. It is currently thought that the respiratory cycle of inspiration and expiration is generated by synaptic interactions within these groups of neurons.

The inspiratory and expiratory medullary neurons are connected to projections from higher brain centres and from chemoreceptors and mechanoreceptors; in turn they drive cranial motor neurons, which govern the activity of muscles in the upper airways and the activity of spinal motor neurons, which supply the diaphragm and other thoracic and abdominal muscles. The inspiratory and expiratory medullary neurons also receive input from nerve cells responsible for cardiovascular and temperature regulation, allowing the activity of these physiological systems to be coordinated with respiration.

Neurally, inspiration is characterized by an augmenting discharge of medullary neurons that terminates abruptly. After a gap of a few milliseconds, inspiratory activity is restarted, but at a much lower level, and gradually declines until the onset of expiratory neuron activity. Then the cycle begins again. The full development of this pattern depends on the interaction of several types of respiratory neurons: inspiratory, early inspiratory, off-switch, post-inspiratory, and expiratory.

Early inspiratory neurons trigger the augmenting discharge of inspiratory neurons. This increase in activity, which produces lung expansion, is caused by self-excitation of the inspiratory neurons and perhaps by the activity of an as yet undiscovered upstream pattern generator. Off-switch neurons in the medulla terminate inspiration, but pontine neurons and input from stretch receptors in the lung help control the length of inspiration. When the vagus nerves are sectioned or pontine centres are destroyed, breathing is characterized by prolonged inspiratory activity that may last for several minutes. This type of breathing, which occasionally occurs in persons with diseases of the brain stem, is called apneustic breathing.

Post-inspiratory neurons

Post-inspiratory neurons are responsible for the declining discharge of the inspiratory muscles that occurs at the beginning of expiration. Mechanically, this discharge aids in slowing expiratory flow rates and probably assists the efficiency of gas exchange. It is believed by some that these post-inspiratory neurons have inhibitory effects on both inspiratory and expiratory neurons and therefore play a significant role in determining the length of the respiratory cycle and the different phases of respiration.

As the activity of the post-inspiratory neurons subsides, expiratory neurons discharge and inspiratory neurons are strongly inhibited. There may be no peripheral manifestation of expiratory neuron discharge except for the absence of inspiratory muscle activity, although in upright humans the lower expiratory intercostal muscles and the abdominal muscles may be active even during quiet breathing. Moreover, as the demand to breathe increases (for example, with exercise), more expiratory intercostal and abdominal muscles contract. As expiration proceeds, the inhibition of the inspiratory muscles gradually diminishes and inspiratory neurons resume their activity.

Chemoreceptors. One way in which breathing is controlled is through feedback by chemoreceptors. There are two kinds of respiratory chemoreceptors: arterial chemoreceptors, which monitor and respond to changes in the partial pressure of oxygen and carbon dioxide in the arterial blood, and central chemoreceptors in the brain, which respond to changes in the partial pressure of carbon dioxide in their immediate environment. Ventilation levels behave as if they were regulated to maintain a constant level of carbon dioxide partial pressure and to ensure adequate oxygen levels in the arterial blood. Increased activity of chemoreceptors caused by hypoxia or an increase in the partial pressure of carbon dioxide augments both the rate and depth of breathing, which restores partial pressures

of oxygen and carbon dioxide to their usual levels. On the other hand, too much ventilation depresses the partial pressure of carbon dioxide, which leads to a reduction in chemoreceptor activity and a diminution of ventilation. During sleep and anesthesia, lowering carbon dioxide levels three to four millimetres of mercury below values occurring during wakefulness can cause a total cessation of breathing (apnea).

Peripheral chemoreceptors. Hypoxia, or the reduction of oxygen supply to tissues to below physiological levels (produced, for example, by a trip to high altitudes), stimulates the carotid and aortic bodies, the principal arterial chemoreceptors. The two carotid bodies are small organs located in the neck at the bifurcation of each of the two common carotid arteries into the internal and external carotid arteries. This organ is extraordinarily well perfused and responds to changes in the partial pressure of oxygen in the arterial blood flowing through it rather than to the oxygen content of that blood (the amount of oxygen chemically combined with hemoglobin). The sensory nerve from the carotid body increases its firing rate hyperbolically as the partial pressure of oxygen falls. In addition to responding to hypoxia, the carotid body increases its activity linearly as the partial pressure of carbon dioxide in arterial blood is raised. This arterial blood parameter rises and falls as air enters and leaves the lungs, and the carotid body senses these fluctuations, responding more to rapid than to slow changes in the partial pressure of carbon dioxide. Larger oscillations in the partial pressure of carbon dioxide occur with breathing as metabolic rate is increased. The amplitude of these fluctuations, as reflected in the size of carotid body signals, may be used by the brain to detect changes in the metabolic rate and to produce appropriate adjustment in ventilation.

The carotid bodies

The carotid body communicates with medullary respiratory neurons through sensory fibres that travel with the carotid sinus nerve, a branch of the glossopharyngeal nerve. Microscopically, the carotid body consists of two different types of cells. The type I cells are arranged in groups and are surrounded by type II cells. The type II cells are generally not believed to have a direct role in chemoreception. Fine sensory nerve fibres are found in juxtaposition to type I cells, which, unlike type II cells, contain electron-dense vesicles. Acetylcholine, catecholamines, and neuropeptides such as enkephalins, vasoactive intestinal peptide, and substance P, are located within the vesicles. It is believed that hypoxia and hypercapnia (excessive carbon dioxide in the blood) cause the release of one or more of these neuroactive substances from the type I cells, which then act on the sensory nerve. It is possible to interfere independently with the responses of the carotid body to carbon dioxide and oxygen, which suggests that the same mechanisms are not used to sense or transmit changes in oxygen or carbon dioxide. The aortic bodies located near the arch of the aorta also respond to acute changes in the partial pressure of oxygen, but less well than the carotid body responds to changes in the partial pressure of carbon dioxide. The aortic bodies are responsible for many of the cardiovascular effects of hypoxia.

The aortic bodies

Central chemoreceptors. Carbon dioxide is one of the most powerful stimulants of breathing. As the partial pressure of carbon dioxide in arterial blood rises, ventilation increases nearly linearly. Ventilation normally increases by two to four litres per minute with each one millimetre of mercury increase in the partial pressure of carbon dioxide. Carbon dioxide increases the acidity of the fluid surrounding the cells but also easily passes into cells and thus can make the interior of cells more acid. It is not clear whether the receptors respond to the intracellular or extracellular effects of carbon dioxide or acidity.

Even if both the carotid and aortic bodies are removed, inhaling gases that contain carbon dioxide stimulates breathing. This observation shows that there must be additional receptors that respond to changes in the partial pressure of carbon dioxide. Current thinking places these receptors near the undersurface (ventral part) of the medulla. However, microscopic examination has not conclusively identified specific chemoreceptor cells in this region. The same areas of the ventral medulla also contain vasomotor neurons that are concerned with the regulation of blood pressure. Some investigators argue that respiratory responses produced at the ventral medullary surface are direct and are caused by interference with excitatory and inhibitory inputs to respiration from these vasomotor neurons. They believe that respiratory chemoreceptors that respond to carbon dioxide are more diffusely distributed in the brain.

Muscle and lung receptors. Receptors in the respiratory muscles and in the lung can also affect breathing patterns. These receptors are particularly important when lung function is impaired, since they can help maintain tidal volume and ventilation at normal levels.

Changes in the length of a muscle affect the force it can produce when stimulated. Generally there is a length at which the force generated is maximal. Receptors, called spindles, in the respiratory muscles measure muscle length and increase motor discharge to the diaphragm and intercostal muscles when increased stiffness of the lung or resistance to the movement of air caused by disease impedes muscle shortening. Tendon organs, another receptor in muscles, monitor changes in the force produced by muscle contraction. Too much force stimulates tendon organs and causes decreasing motor discharge to the respiratory muscles and may prevent the muscles from damaging themselves.

Inflation of the lungs in animals stops breathing by a reflex described by the German physiologist Ewald Hering and the Austrian physiologist Josef Breuer. The Hering-Breuer reflex is initiated by lung expansion, which excites stretch receptors in the airways. Stimulation of these receptors, which send signals to the medulla by the vagus nerve, shortens inspiratory times as tidal volume (the volume of air inspired) increases, accelerating the frequency of breathing. When lung inflation is prevented, the reflex allows inspiratory time to be lengthened, helping to preserve tidal volume.

The Hering-Breuer reflex

There are also receptors in the airways and in the alveoli that are excited by rapid lung inflations and by chemicals such as histamine, bradykinin, and prostaglandins. The most important function of these receptors, however, may be to defend the lung against noxious material in the atmosphere. When stimulated, these receptors constrict the airways and cause rapid shallow breathing, which inhibits the penetration of injurious agents into the bronchial tree. These receptors are supplied, like the stretch receptors, by the vagus nerve. Some of these receptors (called irritant receptors) are innervated by myelinated nerve fibres, others (the J receptors) by unmyelinated fibres. Stimulation of irritant receptors also causes coughing.

Variations in breathing. *Exercise.* One of the remarkable features of the respiratory control system is that ventilation increases sufficiently to keep the partial pressure of carbon dioxide in arterial blood nearly unchanged despite the large increases in metabolic rate that can occur with exercise, thus preserving acid–base homeostasis. A number of signals arise during exercise that can augment ventilation. Sources of these signals include mechanoreceptors in the exercising limbs; the arterial chemoreceptors, which can sense breath-by-breath oscillations in the partial pressure of carbon dioxide; and thermal receptors, because body temperature rises as metabolism increases. The brain also seems to anticipate changes in the metabolic rate caused by exercise, because parallel increases occur in the output from the motor cortex to the exercising limbs and to respiratory neurons. Changes in the concentration of potassium and lactic acid in the exercising muscles acting on unmyelinated nerve fibres may be another mechanism for stimulation of breathing during exercise. It remains unclear, however, how these various mechanisms are adjusted to maintain acid–base balance.

Sleep. During sleep, body metabolism is reduced, but there is an even greater decline in ventilation so that the partial pressure of carbon dioxide in arterial blood rises slightly and arterial partial pressure of oxygen falls. The effects on ventilatory pattern vary with sleep stage. In slow-wave sleep, breathing is diminished but remains regular, while in rapid eye movement sleep, breathing can become quite erratic. Ventilatory responses to inhaled car-

bon dioxide and to hypoxia are less in all sleep stages than during wakefulness. Sufficiently large decreases in the partial pressure of oxygen or increases in the partial pressure of carbon dioxide will cause arousal and terminate sleep.

During sleep, ventilation may swing between periods when the amplitude and frequency of breathing are high and periods in which there is little attempt to breathe, or even apnea (cessation of breathing). This rhythmic waxing and waning of breathing, with intermittent periods of apnea, is called Cheyne-Stokes breathing, after the physicians who first described it. The mechanism that produces the Cheyne-Stokes ventilation pattern is still argued, but it may entail unstable feedback regulation of breathing. Similar swings in ventilation sometimes occur in persons with heart failure or with central nervous system disease.

In addition, ventilation during sleep may intermittently fall to low levels or cease entirely because of partial or complete blockage of the upper airways. In some individuals, this intermittent obstruction occurs repeatedly during the night, leading to severe drops in the levels of blood oxygenation. The condition, termed sleep apnea syndrome, occurs most commonly in the elderly, in the newborn, in males, and in the obese. Because arousal is often associated with the termination of episodes of obstruction, sleep is of poor quality, and complaints of excessive daytime drowsiness are common. Snoring and disturbed behaviour during sleep may also occur.

Sleep apnea syndrome

In some persons with sleep apnea syndrome, portions of the larynx and pharynx may be narrowed by fat deposits or by enlarged tonsils and adenoids, which increase the likelihood of obstruction. Others, however, have normal upper airway anatomy, and obstruction may occur because of discoordinated activity of upper airway and chest wall muscles. Many of the upper airway muscles, like the tongue and laryngeal adductors, undergo phasic changes in their electrical activity synchronous with respiration, and the reduced activity of these muscles during sleep may lead to upper airway closure. (N.S.C.)

THE MECHANICS OF BREATHING

Air moves in and out of the lungs in response to differences in pressure. When the air pressure within the alveolar spaces falls below atmospheric pressure, air enters the lungs (inspiration), provided the larynx is open; when the air pressure within the alveoli exceeds atmospheric pressure, air is blown from the lungs (expiration). The flow of air is rapid or slow in proportion to the magnitude of the pressure difference. Because atmospheric pressure remains relatively constant, flow is determined by how much above or below atmospheric pressure the pressure within the lungs rises or falls.

Alveolar pressure fluctuations are caused by expansion and contraction of the lungs resulting from tensing and relaxing of the muscles of the chest and abdomen. Each small increment of expansion transiently increases the space enclosing lung air. There is, therefore, less air per unit of volume in the lungs and pressure falls. A difference in air pressure between atmosphere and lungs is created, and air flows in until equilibrium with atmospheric pressure is restored at a higher lung volume. When the muscles of inspiration relax, the volume of chest and lungs decreases, lung air becomes transiently compressed, its pressure rises above atmospheric pressure, and flow into the atmosphere results until pressure equilibrium is reached at the original lung volume. This, then, is the sequence of events during each normal respiratory cycle: lung volume change leading to pressure difference, resulting in flow of air into or out of the lung and establishment of a new lung volume.

The lung–chest system. The forces that normally cause changes in volume of the chest and lungs stem not only from muscle contraction but from the elastic properties of both the lung and the chest. A lung is similar to a balloon in that it resists stretch, tending to collapse almost totally unless held inflated by a pressure difference between its inside and outside. This tendency of the lung to collapse or pull away from the chest is measurable by carefully placing a blunt needle between the outside of the lung and the inside of the chest wall, thereby allowing the lung to separate from the chest at this particular spot. The pressure measured in the small pleural space so created is substantially below atmospheric pressure at a time when the pressure within the lung itself equals atmospheric pressure. This negative (below-atmospheric) pressure is a measure, therefore, of the force required to keep the lung distended. The force increases (pleural pressure becomes more negative) as the lung is stretched and its volume increases during inspiration. The force also increases in proportion to the rapidity with which air is drawn into the lung and decreases in proportion to the force with which air is expelled from the lungs. In summary, the pleural pressure reflects primarily two forces: (1) the force required to keep the lung inflated against its elastic recoil and (2) the force required to cause airflow in and out of the lung. Because the pleural pressure is below atmospheric pressure, air is sucked into the chest and the lung collapses (pneumothorax) when the chest wall is perforated, as by a wound or by a surgical incision.

Elastic properties of lung

The force required to maintain inflation of the lung and to cause airflow is provided by the chest and diaphragm (the muscular partition between chest and abdomen), which are in turn stretched inward by the pull of the lungs. The lung–chest system thus acts as two opposed coiled springs, the length of each of which is affected by the other. Were it not for the outward traction of the chest on the lungs, these would collapse; and were it not for the inward traction of the lungs on the chest and diaphragm, the chest would expand to a larger size and the diaphragm would fall from its dome-shaped position within the chest.

The role of muscles. The respiratory muscles displace the equilibrium of elastic forces in the lung and chest in one direction or the other by adding muscular contraction. During inspiration, muscle contraction is added to the outward elastic force of the chest to increase the traction on the lung required for its additional stretch. When these muscles relax, the additional retraction of lung returns the system to its equilibrium position.

Contraction of the abdominal muscles displaces the equilibrium in the opposite direction by adding increased abdominal pressure to the retraction of lungs, thereby further raising the diaphragm and causing forceful expiration. This additional muscular force is removed on relaxation and the original lung volume is restored. During ordinary breathing, muscular contraction occurs only on inspiration, expiration being accomplished "passively" by elastic recoil of the lung.

At total relaxation of the muscles of inspiration and expiration, the lung is distended to a volume—called the functional residual capacity—of about 40 percent of its maximum volume at the end of full inspiration. Further reduction of the lung volume results from maximal contraction of the expiratory muscles of chest and abdomen. The volume in these circumstances is known as the residual volume; it is about 20 percent of the volume at the end of full inspiration (known as the total lung capacity). Additional collapse of the lung to its "minimal air" can be accomplished only by opening the chest wall and creating a pneumothorax.

Functional residual capacity

The membranes of the surface of the lung (visceral pleura) and on the inside of the chest (parietal pleura) are normally kept in close proximity (despite the pull of lung and chest in opposite directions) by surface tension of the thin layer of fluid covering these surfaces. The strength of this bond can be appreciated by the attempt to pull apart two smooth surfaces, such as pieces of glass, separated by a film of water.

The respiratory pump and its performance. The energy expended on breathing is used primarily in stretching the lung–chest system and thus causing airflow. It normally amounts to 1 percent of the basal energy requirements of the body but rises substantially during exercise or illness. The respiratory pump is versatile, capable of increasing its output 25 times, from a normal resting level of about six litres (366 cubic inches) per minute to 150 litres per minute in adults. Pressures within the lungs can be raised to 130 centimetres of water (about 1.8 pounds per square inch) by the so-called Valsalva maneuver—*i.e.*, a forceful contraction of the chest and abdominal muscles against a closed glottis (*i.e.*, with no space between the vocal

Airflow velocity

cords). Airflow velocity, normally reaching 30 litres per minute in quiet breathing, can be raised voluntarily to 400 litres per minute. Cough is accomplished by suddenly opening the larynx during a brief Valsalva maneuver. The resultant high-speed jet of air is an effective means of clearing the airways of excessive secretions or foreign particles. The beating of cilia (hairline projections) from cells lining the airways normally maintains a steady flow of secretions toward the nose, cough resulting only when this action cannot keep pace with the rate at which secretions are produced.

An infant takes 33 breaths per minute with a tidal volume (the amount of air breathed in and out in one cycle) of 15 millilitres, totaling about 0.5 litre—approximately one pint—per minute as compared to adult values of 14 breaths, 500 millilitres, and seven litres, respectively.

If the force of surface tension is responsible for the adherence of parietal and visceral pleurae, it is reasonable to question what keeps the lungs' alveolar walls (also fluid-covered) from sticking together and thus eliminating alveolar airspaces. In fact, such adherence occasionally does occur and is one of the dreaded complications of premature births. Normal lungs, however, contain a substance—a phospholipid surfactant—that reduces surface tension and keeps alveolar walls separated. (A.A.S.)

GAS EXCHANGE

Respiratory gases—oxygen and carbon dioxide—move between the air and the blood across the respiratory exchange surfaces in the lungs. The structure of the human lung provides an immense internal surface that facilitates gas exchange between the alveoli and the blood in the pulmonary capillaries. The area of the alveolar surface in the adult human is about 100 square metres. Gas exchange across the membranous barrier between the alveoli and capillaries is enhanced by the thin nature of the membrane, about 0.5 micrometre, or $^1/_{100}$ of the diameter of a human hair.

Convection and diffusion

Respiratory gases move between the environment and the respiring tissues by two principal mechanisms, convection and diffusion. Convection, or mass flow, is responsible for movement of air from the environment into the lungs and for movement of blood between the lungs and the tissues. Respiratory gases also move by diffusion across tissue barriers such as membranes. Diffusion is the primary mode of transport of gases between air and blood in the lungs and between blood and respiring tissues in the body. The process of diffusion is driven by the difference in partial pressures of a gas between two locales. In a mixture of gases, the partial pressure of each gas is directly proportional to its concentration. The partial pressure of a gas in fluid is a measure of its tendency to leave the fluid when exposed to a gas or fluid that does not contain that gas. A gas will diffuse from an area of greater partial pressure to an area of lower partial pressure regardless of the distribution of the partial pressures of other gases. There are large changes in the partial pressures of oxygen and carbon dioxide as these gases move between air and the respiring tissues. The partial pressure of carbon dioxide in this pathway is lower than the partial pressure of oxygen, due to differing modes of transport in the blood, but almost equal quantities of the two gases are involved in metabolism and gas exchange.

Oxygen and carbon dioxide are transported between tissue cells and the lungs by the blood. The quantity transported is determined both by the rapidity with which the blood circulates and the concentrations of gases in blood. The rapidity of circulation is determined by the output of the heart, which in turn is responsive to overall body requirements. Local flows can be increased selectively, as occurs, for example, in the flow through skeletal muscles during exercise. The performance of the heart and circulatory regulation are, therefore, important determinants of gas transport.

Oxygen and carbon dioxide are too poorly soluble in blood to be adequately transported in solution. Specialized systems for each gas have evolved to increase the quantities of those gases that can be transported in blood. These systems are present mainly in the red cells, which make up 40 to 50 percent of the blood volume in most mammals. Plasma, the cell-free, liquid portion of blood, plays little role in oxygen exchange but is essential to carbon dioxide exchange.

Transport of oxygen. Oxygen is poorly soluble in plasma, so that less than 2 percent of oxygen is transported dissolved in plasma. The vast majority of oxygen is bound to hemoglobin, a protein contained within red cells. Hemoglobin is composed of four iron-containing ring structures (hemes) chemically bonded to a large protein (globin). Each iron atom can bind and then release an oxygen molecule. Enough hemoglobin is present in normal human blood to permit transport of about 0.2 millilitre of oxygen per millilitre of blood. The quantity of oxygen bound to hemoglobin is dependent on the partial pressure of oxygen in the lung to which blood is exposed. The curve representing the content of oxygen in blood at various partial pressures of oxygen, called the oxygen-dissociation curve (Figure 15), is a characteristic S-shape because binding of oxygen to one iron atom influences the ability of oxygen to bind to other iron sites. In alveoli at sea level, the partial pressure of oxygen is sufficient to bind oxygen to essentially all available iron sites on the hemoglobin molecule.

Hemoglobin and oxygen

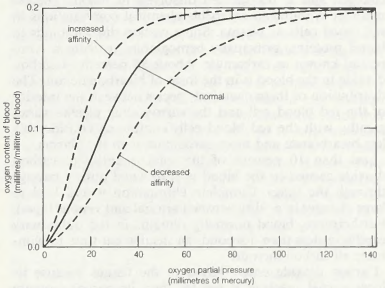


Figure 15: The oxygen-dissociation curve, which shows the relationship between the oxygen content of blood (ml O_2/ml blood) and the partial pressure of oxygen. An increase in the temperature or in the concentrations of 2,3-diphosphoglycerate (2,3-DPG), carbon dioxide (CO_2), or hydrogen ions will decrease the affinity of hemoglobin for oxygen and shift the curve to the right. Conversely, a decrease in the above variables will result in increased affinity and therefore shift the curve to the left.

Not all of the oxygen transported in the blood is transferred to the tissue cells. The amount of oxygen extracted by the cells depends on their rate of energy expenditure. At rest, venous blood returning to the lungs still contains 70 to 75 percent of the oxygen that was present in arterial blood; this reserve is available to meet increased oxygen demands. During extreme exercise the quantity of oxygen remaining in venous blood decreases to 10 to 25 percent. At the steepest part of the oxygen-dissociation curve (the portion between 10 and 40 millimetres of mercury partial pressure), a relatively small decline in the partial pressure of oxygen in the blood is associated with a relatively large release of bound oxygen.

Hemoglobin and other substances

Hemoglobin binds not only to oxygen but to other substances such as hydrogen ions (which determine the acidity, or pH, of the blood), carbon dioxide, and 2,3-diphosphoglycerate (2,3-DPG; a salt in the red blood cells that plays a role in liberating oxygen from hemoglobin in the peripheral circulation). These substances do not bind to hemoglobin at the oxygen-binding sites; however, with the binding of oxygen, changes in the structure of the hemoglobin molecule occur that affect its ability to bind other gases or substances. Conversely, binding of these substances to hemoglobin affects the affinity of hemoglobin for oxygen. (Affinity denotes the tendency of molecules of different species to bind to one another.) Increases in hydrogen ions, carbon dioxide, or 2,3-DPG decrease the affinity of hemoglobin for oxygen, and the oxygen-disso-

ciation curve shifts to the right. Because of this decreased affinity, an increased partial pressure of oxygen is required to bind a given amount of oxygen to hemoglobin. A rightward shift of the curve is thought to be of benefit in releasing oxygen to the tissues when needs are great in relation to oxygen delivery, as occurs with anemia or extreme exercise. Reductions in normal concentrations of hydrogen ions, carbon dioxide, and 2,3-DPG result in an increased affinity of hemoglobin for oxygen, and the curve is shifted to the left. This displacement increases oxygen binding to hemoglobin at any given partial pressure of oxygen and is thought to be beneficial if the availability of oxygen is reduced, as occurs at extreme altitude.

Effect of temperature

Temperature changes affect the oxygen-dissociation curve similarly. An increase in temperature shifts the curve to the right (decreased affinity; enhanced release of oxygen); a decrease in temperature shifts the curve to the left (increased affinity). The range of body temperature usually encountered in humans is relatively narrow, so that temperature-associated changes in oxygen affinity have little physiological importance.

Transport of carbon dioxide. Transport of carbon dioxide in the blood is considerably more complex. A small portion of carbon dioxide, about 5 percent, remains unchanged and is transported dissolved in blood. The remainder is found in reversible chemical combinations in red blood cells or plasma. Some carbon dioxide binds to blood proteins, principally hemoglobin, to form a compound known as carbamate. About 88 percent of carbon dioxide in the blood is in the form of bicarbonate ion. The distribution of these chemical species between the interior of the red blood cell and the surrounding plasma varies greatly, with the red blood cells containing considerably less bicarbonate and more carbamate than the plasma.

Less than 10 percent of the total quantity of carbon dioxide carried in the blood is eliminated during passage through the lungs. Complete elimination would lead to large changes in acidity between arterial and venous blood. Furthermore, blood normally remains in the pulmonary capillaries less than a second, an insufficient time to eliminate all carbon dioxide.

Carbon dioxide enters blood in the tissues because its local partial pressure is greater than its partial pressure in blood flowing through the tissues. As carbon dioxide enters the blood, it combines with water to form carbonic acid (H_2CO_3), a relatively weak acid, which dissociates into hydrogen ions (H^+) and bicarbonate ions (HCO_3^-). Blood acidity is minimally affected by the released hydrogen ions because blood proteins, especially hemoglobin, are effective buffering agents. (A buffer solution resists change in acidity by combining with added hydrogen ions and, essentially, inactivating them.) The natural conversion of carbon dioxide to carbonic acid is a relatively slow process; however, carbonic anhydrase, a protein enzyme present inside the red blood cell, catalyzes this reaction with sufficient rapidity that it is accomplished in only a fraction of a second. Because the enzyme is present only inside the red blood cell, bicarbonate accumulates to a much greater extent within the red cell than in the plasma. The capacity of blood to carry carbon dioxide as bicarbonate is enhanced by an ion transport system inside the red blood cell membrane that simultaneously moves a bicarbonate ion out of the cell and into the plasma in exchange for a chloride ion. The simultaneous exchange of these two ions, known as the chloride shift, permits the plasma to be used as a storage site for bicarbonate without changing the electrical charge of either the plasma or the red blood cell. Only 26 percent of the total carbon dioxide content of blood exists as bicarbonate inside the red blood cell, while 62 percent exists as bicarbonate in plasma; however, the bulk of bicarbonate ions is first produced inside the cell, then transported to the plasma. A reverse sequence of reactions occurs when blood reaches the lung, where the partial pressure of carbon dioxide is lower than in the blood.

Buffering effect of hemoglobin

Chloride shift

Hemoglobin acts in another way to facilitate the transport of carbon dioxide. Amino groups of the hemoglobin molecule react reversibly with carbon dioxide in solution to yield carbamates. A few amino sites on hemoglobin are oxylabile, that is, their ability to bind carbon dioxide depends on the state of oxygenation of the hemoglobin molecule. The change in molecular configuration of hemoglobin that accompanies the release of oxygen leads to increased binding of carbon dioxide to oxylabile amino groups. Thus, release of oxygen in body tissues enhances binding of carbon dioxide as carbamate. Oxygenation of hemoglobin in the lungs has the reverse effect and leads to carbon dioxide elimination.

Only 5 percent of carbon dioxide in the blood is transported free in physical solution without chemical change or binding, yet this pool is important, because only free carbon dioxide easily crosses biologic membranes. Virtually every molecule of carbon dioxide produced by metabolism must exist in the free form as it enters blood in the tissues and leaves capillaries in the lung. Between these two events, most carbon dioxide is transported as bicarbonate or carbamate.

Gas exchange in the lung. The introduction of air into the alveoli allows the removal of carbon dioxide and the addition of oxygen to venous blood. Because ventilation is a cyclic phenomenon that occurs through a system of conducting airways, not all inspired air participates in gas exchange. A portion of the inspired breath remains in the conducting airways and does not reach the alveoli where gas exchange occurs. This portion is approximately one-third of each breath at rest but decreases to as little as 10 percent during exercise, due to the increased size of inspired breaths.

Cyclic nature of ventilation

In contrast to the cyclic nature of ventilation, blood flow through the lung is continuous, and almost all blood entering the lungs participates in gas exchange. The efficiency of gas exchange is critically dependent on the uniform distribution of blood flow and inspired air throughout the lungs. In health, ventilation and blood flow are extremely well matched in each exchange unit throughout the lungs. The lower parts of the lung receive slightly more blood flow than ventilation because gravity has a greater effect on the distribution of blood than on the distribution of inspired air. Under ideal circumstances, partial pressures of oxygen and carbon dioxide in alveolar gas and arterial blood are identical. Normally there is a small difference between oxygen tensions in alveolar gas and arterial blood because of the effect of gravity on matching and the addition of a small amount of venous drainage to the bloodstream after it has left the lungs. These events have no measurable effect on carbon dioxide partial pressures because the difference between arterial and venous blood is so small.

Abnormal gas exchange. Lung disease can lead to severe abnormalities in blood gas composition. Because of the differences in oxygen and carbon dioxide transport, impaired oxygen exchange is far more common than impaired carbon dioxide exchange. Mechanisms of abnormal gas exchange are grouped into four categories—hypoventilation, shunting, ventilation–blood flow imbalance, and limitations of diffusion.

If the quantity of inspired air entering the lungs is less than is needed to maintain normal exchange—a condition known as hypoventilation—the alveolar partial pressure of carbon dioxide rises and the partial pressure of oxygen falls almost reciprocally. Similar changes occur in arterial blood partial pressures because the composition of alveolar gas determines gas partial pressures in blood perfusing the lungs. This abnormality leads to parallel changes in both gas and blood and is the only abnormality in gas exchange that does not cause an increase in the normally small difference between arterial and alveolar partial pressures of oxygen.

Shunting

In shunting, venous blood enters the bloodstream without passing through functioning lung tissue. Shunting of blood may result from abnormal vascular (blood vessel) communications or from blood flowing through unventilated portions of the lung (*e.g.*, alveoli filled with fluid or inflammatory material). A reduction in arterial blood oxygenation is seen with shunting, but the level of carbon dioxide in arterial blood is not elevated even though the shunted blood contains more carbon dioxide than arterial blood.

The differing effects of shunting on oxygen and carbon dioxide partial pressures are the result of the different configurations of the blood-dissociation curves of the two gases. As noted above, the oxygen-dissociation curve is S-shaped and plateaus near the normal alveolar oxygen partial pressure, but the carbon dioxide-dissociation curve is steeper and does not plateau as the partial pressure of carbon dioxide increases. In the example in Figure 16, blood perfusing the collapsed, unventilated area of the lung, labeled A, leaves the lung without exchanging oxygen or carbon dioxide. The content of carbon dioxide, indicated by the letter A next to the carbon dioxide-dissociation curve, is greater than the normal carbon dioxide content, indicated by the square. The remaining healthy portion of the lung, labeled B in the figure, receives both its usual ventilation and the ventilation that normally would be directed to the abnormal lung. This lowers the partial pressure of carbon dioxide in the alveoli of the normal area of the lung. As a result, blood leaving the healthy portion of the lung, indicated by the letter B next to the carbon dioxide-dissociation curve, has a lower carbon dioxide content than normal. The lower carbon dioxide content in this blood counteracts the addition of blood with a higher carbon dioxide content from the abnormal area, and the composite arterial blood carbon dioxide content remains normal, as indicated by the arrow next to the carbon dioxide-dissociation curve. This compensatory mechanism is less efficient than normal carbon dioxide exchange and requires a modest increase in overall ventilation, which is usually achieved without difficulty. Because the carbon dioxide-dissociation curve is steep and relatively linear, compensation for decreased carbon dioxide exchange in one portion of the lung can be counterbalanced by increased excretion of carbon dioxide in another area of the lung.

In contrast, shunting of venous blood has a substantial effect on arterial blood oxygen content and partial pressure. Blood leaving an unventilated area of the lung has an oxygen content (indicated by the letter A next to the oxygen-dissociation curve) that is less than the normal content (indicated by the square). In the healthy area of the lung, labeled B in the figure, the increase in ventilation above normal raises the partial pressure of oxygen in the alveolar gas and, therefore, in the arterial blood. The oxygen-dissociation curve, however, reaches a plateau at the normal alveolar partial pressure, and an increase in blood partial pressure results in a negligible increase in oxygen content. Mixture of blood from this healthy portion of the lung (with normal oxygen content) and blood from the abnormal area of the lung (with decreased oxygen content) produces a composite arterial oxygen content that is less than the normal level, indicated by the arrow next to the oxygen-dissociation curve. Thus, an area of healthy lung cannot counterbalance the effect of an abnormal portion of the lung on blood oxygenation because the oxygen-dissociation curve reaches a plateau at a normal alveolar partial pressure of oxygen. This effect on blood oxygenation is seen not only in shunting but in any abnormality that results in a localized reduction in blood oxygen content.

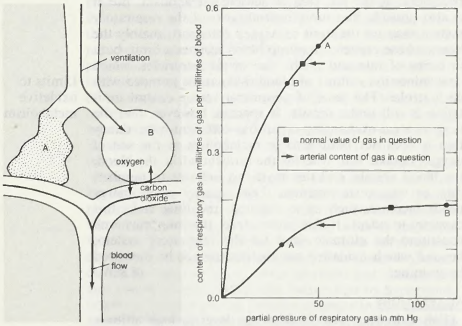


Figure 16: *The effect of shunting of blood through nonfunctioning lung tissue on oxygen and carbon dioxide exchange.*
(Left) Distribution of ventilation and blood flow. (Right) The oxygen- (lower curve) and carbon dioxide- (upper curve) dissociation curves of blood showing the relationships between blood contents and partial pressures of the respiratory gases (see text).

Ventilation–blood flow imbalance

Mismatching of ventilation and blood flow is by far the most common cause of a decrease in partial pressure of oxygen in blood. There are minimal changes in blood carbon dioxide content unless the degree of mismatch is extremely severe. Inspired air and blood flow normally are distributed uniformly, and each alveolus receives approximately equal quantities of both. As matching of inspired air and blood flow deviates from the normal ratio of 1 to 1, alveoli become either overventilated or underventilated in relation to their blood flow. In alveoli that are overventilated, the amount of carbon dioxide eliminated is increased, which counteracts the fact that there is less carbon dioxide eliminated in the alveoli that are relatively underventilated. Overventilated alveoli, however, cannot compensate in terms of greater oxygenation for underventilated alveoli because, as is shown in the oxygen-dissociation curve, a plateau is reached at the alveolar partial pressure of oxygen, and increased ventilation will not increase blood oxygen content. In healthy lungs there is a narrow distribution of the ratio of ventilation to blood flow throughout the lung that is centred around a ratio of 1 to 1. In disease, this distribution can broaden substantially so that individual alveoli can have ratios that markedly deviate from the ratio of 1 to 1. Any deviation from the usual clustering around the ratio of 1 to 1 leads to decreased blood oxygenation—the more disparate the deviation, the greater the reduction in blood oxygenation. Carbon dioxide exchange, on the other hand, is not affected by an abnormal ratio of ventilation and blood flow as long as the increase in ventilation that is required to maintain carbon dioxide excretion in overventilated alveoli can be achieved.

A fourth category of abnormal gas exchange involves limitation of diffusion of gases across the thin membrane separating the alveoli from the pulmonary capillaries. A variety of processes can interfere with this orderly exchange; for oxygen, these include increased thickness of the alveolar–capillary membrane, loss of surface area available for diffusion of oxygen, a reduction in the alveolar partial pressure of oxygen required for diffusion, and decreased time available for exchange due to increased velocity of flow. These factors are usually grouped under the broad description of "diffusion limitation," and any can cause incomplete transfer of oxygen with a resultant reduction in blood oxygen content. There is no diffusion limitation of the exchange of carbon dioxide because this gas is more soluble than oxygen in the alveolar–capillary membrane, which facilitates carbon dioxide exchange. The complex reactions involved in carbon dioxide transport proceed with sufficient rapidity to avoid being a significant limiting factor in exchange. (R.A.Kl.)

INTERPLAY OF RESPIRATION, CIRCULATION, AND METABOLISM

The interplay of respiration, circulation, and metabolism is the key to the functioning of the respiratory system as a whole. Cells set the demand for oxygen uptake and carbon dioxide discharge, that is, for gas exchange in the lungs. The circulation of the blood links the sites of oxygen utilization and uptake. The proper functioning of the respiratory system depends on both the ability of the system to make functional adjustments to varying needs and the design features of the sequence of structures involved, which set the limit for respiration.

Role of respiration

The main purpose of respiration is to provide oxygen to the cells at a rate adequate to satisfy their metabolic needs. This involves transport of oxygen from the lung to the tissues by means of the circulation of blood. In antiquity and the medieval period, the heart was regarded as a furnace where the "fire of life" kept the blood boiling. Modern

cell biology has unveiled the truth behind the metaphor. Each cell maintains a set of furnaces, the mitochondria, where, through the oxidation of foodstuffs such as glucose, the energetic needs of the cells are supplied. The precise object of respiration therefore is the supply of oxygen to the mitochondria.

Cell metabolism depends on energy derived from high-energy phosphates such as adenosine triphosphate (ATP), whose third phosphate bond can release a quantum of energy to fuel many cell processes, such as the contraction of muscle fibre proteins or the synthesis of protein molecules. In the process, ATP is degraded to adenosine diphosphate (ADP), a molecule with only two phosphate bonds. To recharge the molecule by adding the third phosphate group requires energy derived from the breakdown of foodstuffs, or substrates. Two pathways are available: (1) anaerobic glycolysis, or fermentation, which operates in the absence of oxygen; and (2) aerobic metabolism, which requires oxygen and involves the mitochondria. The anaerobic pathway leads to acid waste products and is wasteful of resources: The breakdown of one molecule of glucose generates only two molecules of ATP. In contrast, aerobic metabolism has a higher yield (36 molecules of ATP per molecule of glucose) and results in "clean wastes"—water and carbon dioxide, which are easily eliminated from the body and are recycled by plants in the process of photosynthesis. For any sustained high-level cell activity, the aerobic metabolic pathway is therefore preferable. Since oxidative phosphorylation occurs only in mitochondria, and since each cell must produce its own ATP (it cannot be imported), the number of mitochondria in a cell reflects its capacity for aerobic metabolism, or its need for oxygen.

The supply of oxygen to the mitochondria at an adequate rate is a critical function of the respiratory system, because the cells maintain only a limited store of high-energy phosphates and of oxygen, whereas they usually have a reasonable supply of substrates in stock. If oxygen supply is interrupted for a few minutes, many cells, or even the organism, will die.

Route of oxygen

Oxygen is collected from environmental air, transferred to blood in the lungs, and transported by blood flow to the periphery of the cells where it is discharged to reach the mitochondria by diffusion. The transfer of oxygen to the mitochondria involves several structures and different modes of transports. It begins with ventilation of the lung, which is achieved by convection or mass flow of air through an ingeniously branched system of airways; in the most peripheral airways ventilation of alveoli is completed by diffusion of oxygen through the air to the alveolar surface. The transfer of oxygen from alveolar air into the capillary blood occurs by diffusion across the tissue barrier; it is driven by the oxygen partial pressure difference between alveolar air and capillary blood and depends on the thickness (about 0.5 micrometre) and the surface area of the barrier (about 130 square metres in humans). Convective transport by the blood depends on the blood flow rate (cardiac output) and on the oxygen capacity of the blood, which is determined by its content of hemoglobin in the red blood cells. The last step is the diffusive discharge of oxygen from the capillaries into the tissue and cells, which is driven by the oxygen partial pressure difference and depends on the quantity of capillary blood in the tissue. In this process the blood plays a central role and affects all transport steps: oxygen uptake in the lung, transport by blood flow, and discharge to the cells. Blood also serves as carrier for both respiratory gases: oxygen, which is bound to hemoglobin in the red blood cells, and carbon dioxide, which is carried by both plasma and red blood cells and which also serves as a buffer for acid-base balance in blood and tissues.

Metabolism, or, more accurately the metabolic rate of the cells, sets the demand for oxygen. At rest, a human consumes about 250 millilitres of oxygen each minute. With exercise this rate can be increased more than 10-fold in a normal healthy individual, but a highly trained athlete may achieve a more than 20-fold increase. As more and more muscle cells become engaged in doing work, the demand for ATP and oxygen increases linearly with work rate. This is accompanied by an increased cardiac output, essentially due to a higher heart rate, and by increased ventilation of the lungs; as a consequence, the oxygen partial pressure difference across the air–blood barrier increases and oxygen transfer by diffusion is augmented. These dynamic adjustments to the muscles' needs occur up to a limit that is twice as high in the athlete as in the untrained individual. This range of possible oxidative metabolism from rest to maximal exercise is called the aerobic scope. The upper limit to oxygen consumption is not conferred by the ability of muscles to do work, but rather by the limited ability of the respiratory system to provide or utilize oxygen at a higher rate. Muscle can do more work, but beyond the aerobic scope they must revert to anaerobic metabolism, with the result that waste products, mainly lactic acid, accumulate and limit the duration of work.

Aerobic scope

The limit to oxidative metabolism is therefore set by some features of the respiratory system, from the lung to the mitochondria. Knowing precisely what sets the limit is important for understanding respiration as a key vital process, but it is not straightforward, because of the complexity of the system. Much has been learned from comparative physiology and morphology, based on observations that oxygen consumption rates differ significantly among species. For example, the athletic species in nature, such as dogs or horses, have an aerobic scope more than twofold greater than that of other animals of the same size; this is called adaptive variation. Then, oxygen consumption per unit body mass increases as animals become smaller, so that a mouse consumes six times as much oxygen per gram of body mass as a cow, a feature called allometric variation. Furthermore, the aerobic scope can be increased by training in an individual, but this induced variation achieves at best a 50 percent difference between the untrained and the trained state, well below interspecies differences.

Within the aerobic scope the adjustments are due to functional variation. For example, cardiac output is augmented by increasing heart rate. Mounting evidence indicates that the limit to oxidative metabolism is related to structural design features of the system. The total amount of mitochondria in skeletal muscle is strictly proportional to maximal oxygen consumption, in all types of variation. In training, the mitochondria increase in proportion to the augmented aerobic scope. Mitochondria set the demand for oxygen, and they seem able to consume up to five millilitres of oxygen per minute and gram of mitochondria. If energy (ATP) needs to be produced at a higher rate, the muscle cells make more mitochondria. It is thus possible that oxygen consumption is limited at the periphery, at the last step of aerobic metabolism. But it is also possible that more central parts of the respiratory system may set the limit to oxygen transport, mainly the heart, whose capacity to pump blood reaches a limit, both in terms of rate and of the size of the ventricles, which determines the volume of blood that can be pumped with each stroke. The issue of peripheral versus central limitation is still under debate. It appears, however, that the lung as a gas-exchanging organ has sufficient redundancy that it does not limit aerobic metabolism at the site of oxygen uptake. But, whereas the mitochondria, the blood, the blood vessels, and the heart can increase in number, rate, or volume to augment their capacity when energy needs increase, such as in training, the lung lacks this capacity to adapt. If this proves true, the lung may well constitute the ultimate limit for the respiratory system, beyond which oxidative metabolism cannot be increased by training. (E.R.W.)

Limits to oxidative metabolism

ADAPTATIONS

High altitudes. Ascent from sea level to high altitude has well-known effects upon respiration. The progressive fall in barometric pressure is accompanied by a fall in the partial pressure of oxygen, both in the ambient air and in the alveolar spaces of the lung; and it is this fall that poses the major respiratory challenge to humans at high altitude. Humans and some mammalian species like cattle adjust to the fall in oxygen pressure through the reversible and

non-inheritable process of acclimatization, which, whether undertaken deliberately or not, commences from the time of exposure to high altitudes. Indigenous mountain species like the llama, on the other hand, exhibit an adaptation that is heritable and has a genetic basis.

Respiratory acclimatization

Respiratory acclimatization in humans is achieved through mechanisms that heighten the partial pressure of oxygen at all stages, from the alveolar spaces in the lung to the mitochondria in the cells, where oxygen is needed for the ultimate biochemical expression of respiration. The decline in the ambient partial pressure of oxygen is offset to some extent by greater ventilation, which takes the form of deeper breathing rather than a faster rate at rest. Diffusion of oxygen across the alveolar walls into the blood is facilitated, and in some experimental animal studies the alveolar walls are thinner at altitude than at sea level. The scarcity of oxygen at high altitudes stimulates increased production of hemoglobin and red blood cells, which increases the amount of oxygen transported to the tissues. The extra oxygen is released by increased levels of inorganic phosphates in the red blood cells, such as 2,3-diphosphoglycerate. With a prolonged stay at altitude, the tissues develop more blood vessels, and, as capillary density is increased, the length of the diffusion path along which gases must pass is decreased—a factor augmenting gas exchange. In addition, the size of muscle fibres decreases, which also shortens the diffusion path of oxygen.

The initial response of respiration to the fall of oxygen partial pressure in the blood on ascent to high altitude occurs in two small nodules, the carotid bodies, attached to the division of the carotid arteries on either side of the neck. As the oxygen deprivation persists, the carotid bodies enlarge but become less sensitive to the lack of oxygen. The low oxygen partial pressure in the lung is associated with thickening of the small blood vessels in pulmonary alveolar walls and a slight increase in pulmonary blood pressure, thought to enhance oxygen perfusion of the lung apices.

Indigenous mountain animals like the llama, alpaca, and vicuña in the Andes or the yak in the Himalayas are adapted rather than acclimatized to the low oxygen partial pressures of high altitude. Their hemoglobin has a high oxygen affinity, so that full saturation of the blood with oxygen occurs at a lower partial pressure of oxygen. In contrast to acclimatized humans, these indigenous, adapted mountain species do not have increased levels of hemoglobin or of organic phosphates in the red cells; they do not develop small muscular blood vessels or an increased blood pressure in the lung; and their carotid bodies remain small.

Native human highlanders are acclimatized rather than genetically adapted to the reduced oxygen pressure. After living many years at high altitude, some highlanders lose this acclimatization and develop chronic mountain sickness, sometimes called Monge's disease, after the Peruvian physician who first described it. This disease is characterized by greater levels of hemoglobin. In Tibet some infants of Han origin never achieve satisfactory acclimatization on ascent to high altitude. A chemodectoma, or benign tumour, of the carotid bodies may develop in native highlanders in response to chronic exposure to low levels of oxygen. (Do.A.H.)

Swimming and diving. Fluid is not a natural medium for sustaining human life after the fetal stage; human respiration requires ventilation with air. Nevertheless, all vertebrates, including humans, exhibit a set of responses that may be called a "diving reflex," which involves cardiovascular and metabolic adaptations to conserve oxygen during diving into water. Other physiological changes are also observed, either artificially induced (as by hyperventilation) or resulting from pressure changes in the environment at the same time that a diver is breathing from an independent gas supply.

Hyperventilation, a form of overbreathing that increases the amount of air entering the pulmonary alveoli, may be used intentionally by swimmers to prolong the time they are able to hold their breath under water. Hyperventilation can be dangerous, and this danger is greatly increased if the swimmer descends to depth, as sometimes happens in snorkeling. The increased ventilation prolongs the duration of the breath-hold by reducing the carbon dioxide pressure in the blood, but it cannot provide an equivalent increase in oxygen. Thus the carbon dioxide that accumulates with exercise takes longer to reach the threshold at which the swimmer is forced to take another breath, but concurrently the oxygen content of the blood falls to unusually low levels. The increased environmental pressure of the water around the breath-holding diver increases the partial pressures of the pulmonary gases. This allows an adequate oxygen partial pressure to be maintained in the setting of reduced oxygen content, and consciousness remains unimpaired. When the accumulated carbon dioxide at last forces the swimmer to return to the surface, however, the progressively diminishing pressure of the water on his ascent reduces the partial pressure of the remaining oxygen. Unconsciousness may then occur in or under the water.

Hazards of breath holding

Divers who breathe from an apparatus that delivers gas at the same pressure as that of the surrounding water need not return to the surface to breathe and can remain at depth for prolonged periods. But this apparent advantage introduces additional hazards, many of them unique in human physiology. Most of the hazards result from the environmental pressure of water. Two factors are involved. At the depth of a diver, the absolute pressure, which is approximately one additional atmosphere for each 10-metre increment of depth, is one factor. The other factor, acting at any depth, is the vertical hydrostatic pressure gradient across the body. The effects of pressure are seen in many processes at the molecular and cellular level and include the physiological effects of the increased partial pressures of the respiratory gases, the increased density of the respiratory gases, the effect of changes of pressure upon the volumes of the gas-containing spaces in the body, and the consequences of the uptake of respiratory gases into, and their subsequent elimination from, the blood and tissues of the diver, often with the formation of bubbles. The multiple effects of submersion upon respiration are not easily separated from one another or clearly distinguishable from related effects of pressure upon other bodily systems.

Hazards due to pressure

The increased work of breathing, rather than cardiac or muscular performance, is the limiting factor for hard physical work underwater. Although the increased work of breathing may be largely due to the effects of increased respiratory gas density upon pulmonary function, the use of underwater breathing apparatus adds significant external breathing resistance to the diver's respiratory burden.

Arterial carbon dioxide pressure should remain unchanged during changes of ambient pressure, but the impaired alveolar ventilation at depth leads to some carbon dioxide retention (hypercapnia). This may be compounded by an increased inspiratory content of carbon dioxide, especially if the diver uses closed-circuit and semiclosed-circuit rebreathing equipment or wears an inadequately ventilated helmet. Alveolar oxygen levels can also be disturbed in diving. Hypoxia may result from failure of the gas supply and may occur without warning. More commonly, the levels of inspired oxygen are increased. Oxygen in excess can be a poison; at a partial pressure greater than 1.5 bar ("surface equivalent value" = 150 percent), it may cause the rapid onset of convulsions, and after prolonged exposures at somewhat lower partial pressures it may cause pulmonary oxygen toxicity with reduced vital capacity and later pulmonary edema. In mixed-gas diving, inspired oxygen is therefore maintained at a partial pressure somewhere between 0.2 and 0.5 bar, but at great depths the inhomogeneity of alveolar ventilation and the limitations of gas diffusion appear to require oxygen provision at greater than normal levels.

Effects on pulmonary function

The maximum breathing capacity and the maximum voluntary ventilation of a diver breathing compressed air diminish rapidly with depth, approximately in proportion to the reciprocal of the square root of the increasing gas density. Thus the practice of using an inert gas such as helium as the oxygen diluent at depths where nitrogen becomes narcotic, like an anesthetic, has the additional advantage of providing a breathing gas of lesser density. The use of hydrogen, which in a mixture with less than

4 percent oxygen is noncombustible, provides a greater respiratory advantage for deep diving.

At the extreme depths now attainable by humans—some 500 metres in the sea and more than 680 metres in the laboratory—direct effects of pressure upon the respiratory centre may be part of the "high-pressure neurological syndrome" and may account for some of the anomalies of breathlessness (dyspnea) and respiratory control that occur with exercise at depth.

The term carbon dioxide retainer is commonly applied to a diver who fails to eliminate carbon dioxide in the normal manner. An ability to tolerate carbon dioxide may increase the work capacity of a diver at depth but also may predispose him to other consequences that are less desirable. High values of end-tidal carbon dioxide with only moderate exertion may be associated with a diminished tolerance to oxygen neurotoxicity, a condition that, if it occurs underwater, places the diver at great risk. Nitrogen narcosis is enhanced by the presence of excess carbon dioxide, and the physical properties of carbon dioxide facilitate the nucleation and growth of bubbles on decompression.

Independent of the depth of the dive are the effects of the local hydrostatic pressure gradient upon respiration. The supporting effect of the surrounding water pressure upon the soft tissues promotes venous return from vessels no longer solely influenced by gravity; and, whatever the orientation of the diver in the water, this approximates the effects of recumbency upon the cardiovascular and respiratory systems. Also, the uniform distribution of gas pressure within the thorax contrasts with the hydrostatic pressure gradient that exists outside the chest. Intrathoracic pressure may be effectively lower than the pressure of the surrounding water, in which case more blood will be shifted into the thorax, or it may be effectively greater, resulting in less intrathoracic blood volume. The concept of a hydrostatic balance point within the chest, which represents the net effect of the external pressures and the effects of chest buoyancy, has proved useful in designing underwater breathing apparatuses.

Decompression illness

Intrapulmonary gas expands exponentially during the steady return of a diver toward the surface. Unless vented, the expanding gas may rupture alveolar septa and escape into interstitial spaces. The extra-alveolar gas may cause a "burst lung" (pneumothorax) or the tracking of gas into the tissues of the chest (mediastinal emphysema), possibly extending into the pericardium or into the neck. More seriously, the escaped alveolar gas may be carried by the blood circulation to the brain (arterial gas embolism). This is a major cause of death among divers. Failure to exhale during ascent causes such accidents and is likely to occur if the diver makes a rapid emergency ascent, even from depths as shallow as two metres. Other possible causes of pulmonary barotrauma include retention of gas by a diseased portion of lung and gas trapping due to dynamic airway collapse during forced expiration at low lung volumes.

Decompression sickness may be defined as the illness, following a reduction of pressure, that is caused by the formation of bubbles from gases that were dissolved in the tissues while the diver was at an increased environmental pressure. The causes are related to the inadequacy of the diver's decompression, perhaps failure to follow a correct decompression protocol, or occasionally a diver's idiosyncratic response to an apparently safe decompression procedure. The pathogenesis begins both with the mechanical effects of bubbles and their expansion in the tissues and blood vessels and with the surface effects of the bubbles upon the various components of the blood at the blood–gas interface. The lung plays a significant role in the pathogenesis and natural history of this illness and may contribute to the clinical picture. Shallow, rapid respiration, often associated with a sharp retrosternal pain on deep inspiration, signals the onset of pulmonary decompression sickness, the "chokes." Whether occurring alone or as part of a more complex case of decompression sickness, this respiratory pattern constitutes an acute emergency. It usually responds rapidly to treatment by recompression in a compression chamber. (D.H.E.)

Diseases and disorders of respiration

Diseases of the respiratory system may affect any of the structures and organs that have to do with breathing—the nasal cavities, the throat (pharynx), the larynx, the windpipe (trachea), the airways (bronchi), and the lung tissue. In addition, respiration is dependent on normal functioning of the muscles in the chest wall and the diaphragm, which may also be affected by disease. The respiratory tract is the site of an exceptionally large range of disorders for three main reasons: first, it is exposed to the environment and therefore may be affected by dust or gases in the air; second, it possesses a large network of capillaries through which the entire output of the heart has to pass, which means that diseases that affect the small blood vessels are likely to affect the lung; and third, it may be the site of "sensitivity" or allergic phenomena that may profoundly affect function.

SIGNS AND SYMPTOMS

By contrast, the symptoms of lung disease are relatively few. Cough is a particularly important sign of all diseases that affect any part of the bronchial tree. A cough productive of sputum is the most important manifestation of disease of the major airways, of which bronchitis is a common example. In severe bronchitis the mucous glands lining the bronchi enlarge greatly, and up to a cupful of sputum may be produced in a few hours; more commonly, 30 to 60 millilitres of sputum are produced in a 24-hour period, particularly in the first two hours after awakening in the morning. An irritative cough without sputum may be caused by extension of malignant disease to the bronchial tree from nearby organs. The presence of blood in the sputum (hemoptysis) is an important sign that should never be disregarded. Although it may result simply from an exacerbation of an existing infection, it may also indicate the presence of inflammation, capillary damage, or tumour. It is also a classic sign of tuberculosis of the lungs.

Blood-stained sputum

The second most important symptom of lung disease is dyspnea, or shortness of breath. This sensation, of complex origin, may arise acutely, as when a foreign body is inhaled into the trachea, or with the onset of a severe attack of asthma. More usually, it is insidious in onset and slowly progressive. What is noted is a slowly progressive difficulty in completing some task, such as walking up a flight of stairs, playing golf, or walking uphill. The shortness of breath may vary in severity, but in diseases such as emphysema (described below), in which there is irreversible lung damage, it is constantly present. It may become so severe as to immobilize the victim, and tasks such as dressing cannot be performed without difficulty. Severe fibrosis of the lung, resulting from occupational lung disease or arising from no identifiable antecedent condition, may also cause severe and unremitting dyspnea. Dyspnea is also an early symptom of congestion of the lung as a result of impaired function of the left ventricle of the heart. When this occurs, if the right ventricle that pumps blood through the lungs is functioning normally, the lung capillaries become engorged, and fluid may accumulate in small airways. It is commonly dyspnea that first causes a patient to seek medical advice, but absence of the symptom does not mean that serious lung disease is not present, since, for example, a small lung cancer that is not obstructing an airway will not produce shortness of breath.

Dyspnea

Chest pain may be an early symptom of lung disease, but it is most often associated with an attack of pneumonia, in which case it is due to an inflammation of the pleura that follows the onset of the pneumonic process. This pain is characteristically felt when a deep breath is taken, and it disappears when fluid accumulates in the pleural space, a condition known as a pleural effusion. Acute pleurisy with pain may signal a blockage in a pulmonary vessel, which leads to acute congestion of the affected part, sometimes with a pleurisy over it. Severe chest pain may by occasioned by the spread of malignant disease to involve the pleura, or by a tumour, such as a mesothelioma, arising from the pleura itself. Severe, intractable pain caused by such conditions may require surgery to cut the nerves

Chest pain

supplying the affected segment to give relief. Fortunately, pain of this severity is rare.

To these major symptoms of lung disease—coughing, dyspnea, and chest pain—may be added several of less importance. A wheeziness in the chest may be heard. This is caused by airway obstruction, such as occurs in asthma. Some diseases of the lung are associated with the swelling of the fingertips (and, rarely, of the toes) called "clubbing." Clubbing may be a feature of bronchiectasis (chronic inflammation and dilation of the major airways), diffuse fibrosis of the lung from any cause, and lung cancer. In the case of lung cancer, this unusual sign may disappear after surgical removal of the tumour. In some lung diseases, the first symptom may be a swelling of the lymph nodes that drain the affected area, particularly the small nodes above the collarbone in the neck; enlargement of the lymph nodes in these regions should always lead to a suspicion of intrathoracic disease. Not infrequently, the presenting symptom of a lung cancer is caused by spread of the tumour to other organs. Thus, a hip fracture from bone metastases, cerebral signs from intracranial metastases, or jaundice from liver involvement may all be the first evidence of a primary lung cancer, as may sensory changes in the legs, since a peripheral neuropathy may also be the presenting evidence of these tumours.

General effects of lung disease

The generally debilitating effect of many lung diseases is well recognized. A person with primary lung tuberculosis or with lung cancer, for example, may be conscious only of a general feeling of malaise, unusual fatigue, or seemingly minor symptoms as the first indication of disease. Loss of appetite and loss of weight, a disinclination for physical activity, general psychological depression, and some symptoms apparently unrelated to the lung such as mild indigestion or headaches, may be diverse indicators of lung disease. Not infrequently, the patient may feel as one does when convalescent after an attack of influenza. Because the symptoms of lung disease, especially in the early stage, are variable and nonspecific, physical and radiographic examination of the chest are an essential part of the evaluation of persons with these complaints.

THE DEFENSES OF THE LUNG

Exposed as it is to the outside environment, the respiratory tract possesses a complicated but comprehensive series of defenses against inhaled material. The cells that form the first line of defense in the smaller radicles of the airway and in the alveoli of the lung are the alveolar macrophage cells. These cells can ingest and destroy bacteria and viruses and can remove small particles. They also secrete chemicals that attract white blood cells to the site, and hence they can initiate an inflammatory response in the lung. Particles picked up by macrophages are removed by them into the lymphatic system of the lung and stored in adjacent lymph glands. Soluble particles are removed into the bloodstream, to be finally excreted by the kidney. This is the route followed by the small lead particles emitted in automobile exhaust, which are inhaled, ingested by macrophages, cleared into the bloodstream, and finally excreted. The half-life of these particles in the lung is about 12 hours.

The macrophages

METHODS OF INVESTIGATION

Physical examination of the chest remains important, as it may reveal the presence of an area of inflammation, a pleural effusion, or an airway obstruction; the first two of these conditions will be visible on chest radiographs, but the third is usually audible only. Examination of the sputum for bacteria allows the identification of many infectious organisms and the institution of specific treatment; sputum examination for malignant cells is occasionally helpful. The conventional radiological examination of the chest has been greatly enhanced by the technique of computerized tomography, which shows small lesions and permits their precise location to be defined. The introduction of the flexible bronchoscope (in place of a rigid instrument) has greatly lessened the discomfort and danger associated with visual inspection of the bronchi and allows small tissue samples to be taken for histological study at the same time.

A number of tests are available to determine the functional status of the lung and the effects of disease on pulmonary function. A simple ventilatory test, the measurement of the velocity of a forced maximal expiration after a full inspiration, allows the degree of airway obstruction to be quantified. Airflow obstruction occurs in asthma and emphysema. A related test, of ventilatory capability, measures the volume that can be forcibly expired in one second after a full inspiration. Ventilatory capability can be simply measured with a spirometer or flow meter, and these devices are widely used in field studies. More complex laboratory equipment is necessary to measure the volumes of gas in the lung; the distribution of ventilation within the lung; airflow resistance; the stiffness of the lung, or the pressure required to inflate it; and the rate of gas transfer across the lung, which is commonly measured by recording the rate of absorption of carbon monoxide—used for this purpose because of the high affinity of hemoglobin for it. Gas transfer is impaired in diseases that destroy the lung or cause severe generalized thickening. Arterial blood gases and pH values indicate the adequacy of oxygenation and ventilation and are routinely measured in intensive care unit patients. Tests of exercise capability, in which work load, total ventilation, and gas exchange are compared before and in response to exercise, are useful in assessing disability.

Tests of lung function

MORPHOLOGICAL CLASSIFICATION OF RESPIRATORY DISEASE

It is helpful to recall the main divisions of the respiratory system as a basis for the morphological description of respiratory system diseases. The upper airway consists of the nose, nasopharynx, and larynx. Below these structures lies the trachea, or windpipe. Thereafter the airway divides into two major airways, right and left, then into progressively smaller tubes, until finally the terminal bronchioles, which are about one millimetre in diameter, are reached. On average, 16 generations of division occur between the trachea and the terminal bronchioles. Although there is only one airway at the beginning—the trachea—there are about 650,000 terminal bronchioles. The cross-sectional area of the bronchial tree increases with increasing subdivision. The end of each terminal bronchiole opens into an acinus, so called because the structure resembles a cluster of grapes, and from this point onward the gas-exchanging portion of the lung is reached. The alveoli or air sacs, which are divided into groups or lobules by fibrous partitions, or septa, are small hexagonal structures forming a blind end to the acinus. The wall of the acinus consists of blood capillaries, and the remaining structures are extremely thin, only providing supporting tissue for the rich capillary bed that constitutes the parenchyma, or the essential tissue of the lung itself. The parenchyma is the gas-exchanging tissue of the lung and has a surface area roughly comparable to that of a tennis court. Blood is distributed to the lung through the branching pulmonary artery, which subdivides with the bronchial tree and accompanies the smaller bronchioles into the region of the acinus to supply the capillaries of the alveolar wall. Oxygenated blood from the acini is collected into pulmonary veins, which run at some little distance from the bronchioles. An interstitial space exists around the alveoli and around the bronchioles and blood vessels, and this connects the lymph nodes (the small masses of lymphatic tissue that occur along the course of the lymph vessels) situated in the midline of the thoracic cavity and extending in a chain up into the neck and down into the abdomen.

Overview of the respiratory system

The lung is covered by a protective membrane, the pleura, and the inner lining of the chest wall consists of a similar membrane. The space between these two fibrous coverings, called the intrapleural space, normally contains no air, and only a few millilitres of fluid for lubrication purposes, as during breathing one layer must slide on another. The pleurae may become involved by inflammation or neoplastic disease, in which case an effusion of fluid may occur between the two layers.

From this general description, diseases of the respiratory system may be grouped into the following categories.

Upper airway disease. The nasal sinuses are frequently

the site of both acute and chronic infections. In common with the palate and the nasopharynx, they are also the site of malignant neoplastic changes. Cancer of the larynx is much more common in smokers than in nonsmokers; and cancers of other parts of the upper respiratory tract, some of which may be caused by exposure to dusts and metals, are also more likely to develop in smokers.

The occurrence of upper airway obstruction (particularly common in people who snore) has been documented in sleep laboratory studies. In the sleep apnea syndrome, episodes of obstruction are accompanied by cessation of breathing for up to half a minute and a marked fall in blood oxygen levels, terminating in arousal from sleep.

Sleep apnea syndrome

The sleep apnea syndrome is not uncommon. It is not confined to the very obese, although it forms part of the syndrome of severe obesity in which sleep disturbance is common; and it is associated with the daytime somnolence known as the pickwickian syndrome, after Charles Dickens' description of the fat boy in *The Pickwick Papers.* Sleep apnea is caused by relaxation of muscles around the pharynx and obstruction of the airway by the palate. It is related to narrow anatomical dimensions in this area but is also more likely to occur if alcohol is ingested shortly before sleep. Sleep apnea is more than a medical curiosity: it may cause a rise in systemic blood pressure, and the daytime somnolence may occur while the affected person is at work.

Diseases of major bronchi. The major bronchi can become the seat of chronic inflammation, as in chronic bronchitis or bronchiectasis. The latter disease is not uncommonly caused by the familial disease of cystic fibrosis. The major bronchi may also be the site of development of malignant disease.

Asthma

Diseases of smaller bronchi and bronchioles. It is in the smaller bronchi that major obstruction commonly occurs in asthma, since these bronchioles contain smooth muscle in their walls and the muscle may contract, causing airway obstruction. The small radicles of the bronchial tree, the bronchioles, are commonly involved in infective processes such as viral infections; they are also the primary site of deposition of inhaled dust and particles. Because of the large cross-sectional area of this part of the airway, considerable disease may be present in the bronchioles without affecting the expiratory flow rate. The bronchioles are occasionally the site of a primary noninfective bronchiolitis in persons with rheumatoid arthritis.

The alveolar ducts and alveoli. These structures are the site of primary involvement in many infections, including pneumonia, and it is on the parenchyma of the lung that the main effects of blockage of a pulmonary artery (pulmonary embolus) occur. The capillary bed surrounding the alveoli is subject to damage, and fluid may leak through the alveolar capillaries to accumulate in the lungs (pulmonary edema). The capillary bed is also extensively damaged in the condition known generally as acute respiratory distress syndrome; the exact mechanism of the damage is not yet fully understood. The alveolar walls themselves may undergo diffuse interstitial thickening, a characteristic of diseases grouped under the heading of "diffuse interstitial fibrosis"; interstitial thickening may also occur as a manifestation of collagen diseases such as scleroderma. One of the common forms of emphysema, in which alveolar destruction occurs, entails early loss of tissue at the point where the bronchiole ends in the acinus, resulting in a punched-out lesion in the centriacinar region. It is believed that this form of emphysema is the one that most commonly develops after years of cigarette smoking.

The lung parenchyma is the site of the discrete aggregations of cells, usually giant cells, that form the granulomas characteristic of the generalized disease known as sarcoidosis, and it is in the lung parenchyma that nodules caused by the inhalation of silica particles are found.

The lymphatic system. The system of channels draining the lung may be involved in primary disease of the lungs, but more commonly it is involved in metastatic invasion by distant tumours arising in the breast, stomach, or pancreas.

The pleura. The pleura may be involved in inflammatory or neoplastic processes, either of which may lead to fluid accumulation (pleural effusion) between the two layers.

Although these divisions provide a general outline of the ways in which diseases may affect the lung, they are by no means rigid. It is common for more than one part of the system to be involved in any particular disease process, and disease in one region not infrequently leads to involvement of other parts.

Interdependence of respiratory sites

VIRAL INFECTIONS

A wide variety of viruses are responsible for acute respiratory disease. The common cold—frequently of viral origin—can cause inflammation of the trachea and laryngitis, and these processes may extend to involve the lower bronchial tree. After such episodes the ciliary lining of the bronchial tree may be damaged, but the repair process is usually rapid.

Infections with rhinoviruses and adenoviruses are especially important in children, in whom they cause a febrile illness, occasionally with severe bronchiolar involvement. Although recovery is usually rapid, in some infections with respiratory syncytial virus an extensive bronchiolitis develops that may be severe enough to threaten life. In epidemics of these diseases, occasional cases occur in which the course is complicated by inflammation of the pericardium—the membrane enclosing the heart—or by a pleural effusion.

Influenza viruses

Influenza and parainfluenza viruses are capable of causing severe illness. The influenza virus attacks many systems of the body simultaneously, but the primary site of viral replication appears to be the alveolar cells of the lung. There the virus multiplies many times over within a 24-hour period, and the pulmonary involvement may begin in the parenchyma and cause considerable consolidation and inflammation of lung tissue. It is common for there to be severe tracheitis, bronchitis, and bronchiolitis at the same time. Another form of the disorder is that described as viral pneumonia, in which a distinguishing feature is the presence of patchy areas of atelectasis or partial collapse of lung tissue, without extensive involvement of the bronchial tree. All of these conditions are more dangerous in small children and in the elderly, and the lung that is the seat of a severe attack of influenza may quickly become secondarily infected.

It was secondary bacterial infection that accounted for the high mortality in the influenza epidemic of 1918 and 1919, one of the worst human catastrophes on record. It has been estimated that more than 20,000,000 people around the world died during the epidemic, and of the 20,000,000 people who suffered from the illness in the United States, approximately 850,000 died. It was a characteristic of this epidemic that young people were commonly severely affected. The high mortality resulted from the lack of antibiotics for treating the secondary bacterial infection; widespread malnutrition probably also contributed to the death rate.

Three types of influenza virus

There are three immunologically distinct types of influenza virus, designated A, B, and C; parainfluenza viruses are designated by the letter D. Types A, B, and D cause epidemic disease. Within type A there are now known to be at least four distinct strains. The "Asian" strain of type A was responsible for the 1957 influenza epidemic. Epidemic influenza tends to occur in two- or three-year cycles; careful study has allowed predictions to be made of their future occurrence. Although infected individuals develop lasting immunity to a particular strain following an attack of influenza, the immunity is highly specific as to type, and no protection is afforded against even closely related strains. Artificial immunization with high potency vaccines is of value in protecting against previous strains, and the vaccines have been shown to ameliorate the infection in the general population. Their use is particularly indicated in elderly people whose cardiac or lung function is already compromised.

Mycoplasma, identified in 1944 as responsible for a group of pneumonias previously thought to be of viral origin, is a member of a group of organisms known as the pleuropneumonia-like organisms and has also been termed the

"Eaton agent" after the scientist who first described it. *Mycoplasma pneumoniae* is the single most common cause of pneumonia in school-age children and young adults. The infection produces soft patchy shadows on the chest radiograph and relatively few signs on physical examination. A nonproductive cough and fever occur for a few days. Familial spread is common, and disease occurs in epidemic form in young healthy people brought together in clusters, as in military recruit camps and colleges, where a number of outbreaks have been documented. It is not usually a life-threatening disease, but in rare cases it may progress to cause acute respiratory distress syndrome.

Psittacosis and ornithosis, primarily infections of birds, and particularly common among parakeets and parrots, are transmitted to human beings by inhalation of dust particles from the droppings of infected birds. The onset of psittacosis may be quite severe, with headache, insomnia, and even delirium. Gastrointestinal symptoms such as vomiting and pain are frequent, and a cough productive of clear sputum usually develops after a few days. Mild attacks are often unrecognized and dismissed as due to influenza. Recovery is usually complete, but convalescence may be slow. A pandemic of this disease in 1929 was caused by the shipment of 5,000 parrots into Argentina from Brazil for auction. Many of the birds died, and there was considerable human mortality.

Q fever is an infection with *Coxiella burnetii.* The disease was first described in Queensland, Australia; areas in which Q fever is now known to be endemic include Australia, the western United States, Africa, England, and the Mediterranean countries. Animal infection is widespread and involves a large variety of domestic farm animals, particularly cattle and sheep, and some wild animals. Transmission is believed to occur between mammals through ticks and mice. Human disease, which is uncommon, is probably acquired through inhalation of infected material. Laboratory workers and employees in slaughterhouses are particularly at risk. Q fever is usually a mild and self-limited disease, requiring only symptomatic treatment.

The disease chickenpox (varicella), particularly when it occurs in adults, may affect the lung. Acute lesions may occur in the lung parenchyma, leading to a transient but significant fall in arterial oxygen tension (hypoxemia), occasionally necessitating oxygen therapy. Recovery may be slow but is usually complete, although shadows may remain on the radiograph as a result of it.

Whooping cough

Whooping cough occurs in epidemic form among children and appears to be linked to the later development of the chronic infective process known as bronchiectasis, which occurs as a result of bronchial damage. In western countries, both whooping cough and measles (which causes an acute bronchiolitis) have been controlled by effective vaccines. In developing countries, where these vaccines are not administered, these diseases are still a major cause of mortality in children. Resistance to acute respiratory diseases is reduced when there is malnutrition.

Slow process of repair

Lung repair after viral infection. The reparative processes in the lung after any viral attack may be quite slow. Apparent clinical recovery may occur relatively quickly and the radiograph may show no remaining shadows, yet repair and restitution of the alveolar wall may take several additional weeks. There is speculation that the occurrence of a severe viral infection in childhood may impair subsequent development of the lung or even set the stage for chronic respiratory disease in later life, but this has not been proved.

BACTERIAL PNEUMONIA

Before effective antibiotics became available, pneumonia was the respiratory disease responsible for the greatest mortality and consequently was one of the most feared diseases. Because it frequently led to the death of severely disabled, elderly people, it was also known as the "old man's friend." The most common form of the disease is caused by a streptococcus, *Diplococcus pneumoniae.* Infection is followed by an acute illness of sudden onset with high fever, involvement of one or more lobes of the lung with resultant consolidation of the lung tissue, followed by complications such as a lung abscess, pleurisy, or heart failure; but terminating naturally after about seven days with a high peak of fever and a sudden crisis, followed by a sharp fall in temperature and slow resolution. The classic form of the disease is now rarely seen, since prompt antibiotic therapy controls the acute process within 24 hours. Streptococci cause a diffuse type of bronchopneumonia, and it is this condition that is most likely to occur as a consequence of forced recumbency in elderly people. The infection probably develops in those parts of the lung where airway closure has occurred and where some extravasation of fluid is present. Streptococcal pneumonia may also occur as a complication of an acute attack of influenza, and the much lowered mortality of influenza is to be explained by effective antibiotics against streptococci. Staphylococcal pneumonia occurs as an acute illness in small children and may lead to rapid destruction of lung tissue with abscess formation; however, if the acute state is survived, as it usually is with chemotherapy, the lung recovers fully. This type of pneumonia may also occur as a complication of preexisting lung disease of any kind and may follow aspiration of stomach contents into the lung. The development of antibiotic-resistant staphylococci has meant that this form of pneumonia may be a problem in the hospital environment, complicating other lung disease or occurring postoperatively. Pneumonia due to infection with *Klebsiella pneumoniae* may be difficult to treat and characteristically may occur as a repetitive series of episodes of pneumonia, each running a rather long course with slow resolution. The organism *Hemophilus influenzae* is commonly isolated from the sputum of patients with chronic bronchitis during acute exacerbations of infection, but its exact role as a cause of disease is not well understood.

In all these bacterial pneumonias, the diagnosis may be made from the characteristic radiographic pattern, together with isolation from the sputum of the bacterium primarily responsible.

Legionnaires' disease of 1976

Legionnaires' disease. In July 1976, an outbreak of severe pneumonia occurred among U.S. veterans attending a convention of the American Legion in Philadelphia. Of the 147 persons admitted to hospitals, 29 died. Identification of the organism responsible (subsequently named *Legionella pneumophila*) constituted a classic medical detective story. The bacterium had evaded detection before because it does not stain with the usual stains used in sputum examination. It is now known that this bacterium may grow in air-conditioning systems or on shower heads, and it has been shown to be responsible for sporadic but severe outbreaks of pneumonia, particularly but not exclusively in older people. Fortunately, the bacterium is sensitive to erythromycin.

Pneumonia of any cause can lead to the development of the serious state known as adult respiratory distress syndrome, which is discussed in detail in a later section.

Pneumonia in immunocompromised persons. For some years prior to 1980, it had been known that if the immune system was compromised by immunosuppressive drugs (given, for example, before organ transplantation to reduce the rate of rejection), the patient was at risk for developing pneumonia from organisms or viruses not normally pathogenic. Patients with AIDS may develop pneumonia from cytomegalovirus or *Pneumocystis* infections, capable of causing invasive pneumonic lesions in the setting of reduced immunity. Such infections are a major cause of illness in these patients, are difficult to treat, and may prove fatal. Infections with fungi such as *Candida* also occur. The diagnosis and management of these cases has become a challenging and time-consuming responsibility for respiratory specialists in cities with large numbers of AIDS cases.

Pulmonary tuberculosis

Of all the lung diseases caused by bacteria, pulmonary tuberculosis is historically by far the most important. Particular features of this dreaded condition, to which many writers of the last hundred years have turned for dramatic material, include the severe general debilitation and weakness that it may cause; the insidious nature of the onset of its initial symptoms, which may not be pulmonary in nature; the familial tendency; the long-drawn-out course of the disease and the distressing nature of many of

its manifestations, particularly severe hemorrhage from the lung and from tuberculous involvement of the brain (meningitis), or involvement of the adrenal gland leading to adrenal insufficiency (Addison's disease); and, above all, the general inefficacy of medical treatment before effective antibiotic therapy became available. Antibiotics have greatly reduced the mortality from pulmonary tuberculosis in all developed countries, but the decline in mortality began well before their introduction, and it is clear that improved diet and housing were responsible for this. With antibiotic therapy, however, the bacilli quickly disappear from the sputum and the spread of infection is quickly controlled.

In its classic form, tuberculosis first causes pulmonary inflammation at the apices (upper portions) of the lungs, and it may progress slowly to form a chronic cavity in this region. Secondary infection of the cavity may occur and may be difficult to eradicate. When still active, pulmonary tuberculosis is a constant threat to the patient, because blood-borne spread may occur at any time. Diffuse spread of tuberculosis in the lung (known as miliary tuberculosis) may occur at the onset of the disease. The chest radiograph reveals many small and diffuse shadows. The exact sequence of events that leads to this disseminated form of disease is not understood, but prompt treatment is required to prevent spread to the brain and other organs. Pulmonary tuberculosis remains an important disease.

Treatment of tuberculosis

Streptomycin was the first clinically successful antituberculous drug, and it is still occasionally used. More commonly, rifampin together with isonicotinic hydrazide (isoniazid, or INH) is used first, and ethambutol may be used in addition. INH is never given alone as the tubercle bacilli acquire resistance to it; hence it is often combined with another drug, para-aminosalicylic acid (PAS). It is important to stress that there is no one "best" regimen preferable to all others in every case. Tubercle bacilli can acquire resistance to most of the antituberculous drugs, and skillful treatment consists in combining antibiotics so that resistant bacilli are less likely to be produced. The development of resistance to antibiotics can be delayed by the concomitant use of two or more drugs, by continuous treatment without significant interruption until all bacterial growth has ceased, and by the use of bed rest and resectional surgery in a few selected cases. Surgery may be indicated when a chronic cavity has developed. In most newly diagnosed cases, a year of antibiotic therapy is recommended, although recent observations have suggested that with modern treatment regimens, effective control may be achieved in a shorter period.

The major problem in treating pulmonary tuberculosis is ensuring continued medication and supervision. This may be very difficult in developing countries and in indigent and alcoholic populations of large cities in developed countries. Detection and treatment also pose problems in native communities living in isolated regions, such as the Canadian Eskimos. Although the death rate from respiratory tuberculosis in the Western world has fallen greatly since 1900, it remains a serious and difficult problem in some underprivileged communities, in many tropical countries, and in any population with inadequate medical care and poor diet and hygiene. In addition, pulmonary tuberculosis has reappeared in the West in persons with AIDS, in whom treatment is complicated by diminished immunity.

ALLERGIC LUNG DISEASES

There are at least three reasons why the lungs are particularly liable to be involved in allergic reponses to proteins. First, the lungs are exposed to the outside environment, and hence particles of foreign protein—*e.g.,* in pollen—may be deposited directly in the lung; second, the walls of the bronchial tree contain smooth muscle that is very likely to be stimulated to contract if histamine is released by cells affected by the allergic reaction; and third, the lung contains a very large vascular bed, which may be involved in any general inflammatory response. It is therefore not surprising to find that sensitivity phenomena are common and represent an important aspect of pulmonary disease as a whole. The most common and most important of these is asthma. This word is loosely applied to all kinds of conditions in which there is airflow obstruction, but it is better reserved for those conditions in which an allergic component of the bronchial obstruction is likely to be present.

Asthma. Spasmodic asthma is characterized by contraction of the smooth muscle of the airways and, in severe attacks, by airway obstruction from mucus that has accumulated in the bronchial tree. This results in a greater or lesser degree of difficulty in breathing. One approach to classifying asthma differentiates cases that occur with an identifiable antigen, in which antigens affect tissue cells sensitized by a specific antibody, and cases that occur without an identifiable antigen or specific antibody. The former condition is known as "extrinsic" asthma and the latter as "intrinsic" asthma. Extrinsic asthma commonly manifests in childhood because the subject inherits an "atopic" characteristic: the serum contains specific antigens to pollens, mold spores, animal proteins of different kinds, and proteins from a variety of insects, particularly cockroaches and mites that occur in house dust. Exacerbation of extrinsic asthma is precipitated by contact with any of the proteins to which sensitization has occurred; airway obstruction is often worse in the early hours of the morning, for reasons not yet entirely elucidated. The other form of asthma, intrinsic, may develop at any age, and there may be no evidence of specific antigens. Persons with intrinsic asthma experience attacks of airway obstruction unrelated to seasonal changes, although it seems likely that the airway obstruction may be triggered by infections, which are assumed to be viral in many cases.

Extrinsic and intrinsic forms of asthma

Asthma acquired as the result of occupational exposure (a special form of intrinsic asthma) is now recognized to be more common than previously suspected. Exposure to solder resin used in the electronics industry, to toluene diisocyanate (used in many processes as a solvent), to the dust of the western red cedar (in which plicatic acid is the responsible agent), and to many other substances can initiate an asthmatic state, with profound airflow obstruction developing when the subject is challenged by the agent.

It is a characteristic of all types of asthma that those with the condition may exhibit airflow obstruction when given aerosols of histamine or acetylcholine (both normally occurring smooth muscle constrictors) at much lower concentrations than provoke airflow obstruction in normal people; affected individuals may also develop airflow obstruction while breathing cold air or during exercise. These characteristics are used in the laboratory setting to study the airway status of patients. As a result of much recent work, it is thought that the diagnosis of asthma of any kind is difficult to sustain in the absence of a general increase in airway reactivity.

The acute asthmatic attack is alarming both for the sufferer and for the onlooker. There is acute difficulty in breathing, and the chest assumes a more and more inspiratory position. Despite the severe respiratory difficulty, the patient remains fully conscious. The most dangerous form of the condition is known as status asthmaticus. The bronchial spasm worsens over several hours or a day or so, the bronchi become plugged with thick mucus, and airflow is progressively more obstructed. The affected person becomes fatigued; the arterial oxygen tension falls still further, carbon dioxide accumulates in the blood (leading to drowsiness), and the acidity of the arterial blood increases to dangerous levels and may lead to cardiac arrest. Prompt treatment with intravenous corticosteroids and bronchodilators is usually sufficient to relieve the attack, but in occasional cases ventilatory assistance is required. In a few cases, death from asthma is remarkably rapid—too rapid for this complete sequence of events to have occurred, although at autopsy the lungs are overinflated. The exact mechanism of death in these cases is not completely understood.

Status asthmaticus

Although the state of the airway is influenced by psychogenic factors, asthma is not correctly regarded as a disease commonly caused by psychological factors. It may interrupt normal activities and schooling to such an extent that it casts a shadow over the development of the personality. More commonly, it tends to diminish in severity

with age, and people who had quite severe asthma in childhood may lead normal lives after the age of 20. It is now known that asthma attacks may be precipitated by food—in small children, possibly by milk; and some adults are extremely sensitive to sulfite compounds in food or wine. A subgroup of asthmatics are so sensitive to aspirin (acetylsalicylic acid) that ingestion of this chemical may lead to a life-threatening attack.

Changes in mortality from asthma in different countries have been closely studied, but the causes are obscure. It is clear, however, that there has been a considerable increase in the rate of hospital admissions for asthma in children and in adults up to the age of 60. Because there is now more effective treatment for asthma than was available previously, it is not clear why this should be occurring. Unless the asthma is complicated by infection (of which that by the fungus *Aspergillus* is common in damp climates), the chest radiograph remains normal. Asthma does not lead to the destructive lesions of emphysema (described below), although the physical appearance of the patient and the sounds of airflow obstruction in the lung may be similar in the two conditions.

Hay fever. Hay fever is a common seasonal condition caused by allergy to grasses and pollens. It is frequently familial, and the sensitivity is often to ragweed pollen. Conjunctival infection and edema of the nasal mucosa lead to attacks of sneezing. Allergic inflammation and the development of polyps in the nasal passages represent a severer form of hay fever that is often associated with asthma.

Hypersensitivity pneumonitis. This is an important group of conditions in which the lung is sensitized by contact with a variety of agents and in which the response consists of an acute pneumonitis, with inflammation of the smaller bronchioles, alveolar wall edema, and a greater or lesser degree of airflow obstruction due to smooth muscle contraction. In more chronic forms of the condition, granulomas, or aggregations of giant cells, may be found in the lung. One of these illnesses is the so-called farmer's lung, caused by the inhalation of spores from moldy hay (thermophilic *Actinomyces*). This causes an acute febrile illness with a characteristically fine opacification in the basal regions of the lung on the chest radiograph. Airflow obstruction in small airways is present, and there may be measurable interference with diffusion of gases across the alveolar wall. If untreated, the condition may become chronic, with shortness of breath persisting after the radiographic changes have disappeared. Farmer's lung is common in Wisconsin, on the eastern seaboard of Canada, in the west of England, and in France. Education of farmers and their families and the wearing of a simple mask can completely prevent the condition.

Farmer's lung

A similar group of diseases occurs in those with close contact with birds. Variously known as pigeon breeder's lung or bird fancier's lung, these represent different kinds of allergic responses to proteins from birds, particularly proteins contained in the excreta of pigeons, budgerigars (parakeets), and canaries. An acute hypersensitivity pneumonitis may also occur in those cultivating mushrooms (particularly where this is done below ground), after exposure to redwood sawdust, or in response to a variety of other agents. An influenza-like illness resulting from exposure to molds growing in humidifier systems in office buildings ("humidifier fever") has been well-documented. It is occasionally attributable to *Aspergillus*, but sometimes the precise agent cannot be identified. The disease may present as an atypical nonbacterial pneumonia and may be labeled a viral pneumonia if careful inquiry about possible contacts with known agents is not made.

Bird fancier's lung

BRONCHITIS AND BRONCHIOLITIS

Acute bronchitis. Acute bronchitis most commonly occurs as a consequence of viral infection. It may also be precipitated by acute exposure to irritant gases, such as ammonia, chlorine, or sulfur dioxide. In people with chronic bronchitis—a common condition in cigarette smokers—exacerbations of infection are common. The bronchial tree in acute bronchitis is reddened and congested, and minor blood streaking of the sputum may occur. Most cases of acute bronchitis resolve over a few days, and the mucosa repairs itself.

Bronchiolitis refers to inflammation of the small airways. Bronchiolitis probably occurs to some extent in acute viral disorders, particularly in children between the ages of one and two years, and particularly in infections with respiratory syncytial virus. In severe cases the inflammation may be severe enough to threaten life, but it normally clears spontaneously, with complete healing in all but a very small percentage of cases. In adults, acute bronchiolitis of this kind is not a well-recognized clinical syndrome, though there is little doubt that in most patients with chronic bronchitis, acute exacerbations of infection are associated with further damage to small airways. In isolated cases, an acute bronchiolitis is followed by a chronic obliterative condition, or this may develop slowly over time. This pattern of occurrence has only recently been recognized. In addition to patients acutely exposed to gases, in whom such a syndrome may follow the acute exposure, patients with rheumatoid arthritis may develop a slowly progressive obliterative bronchiolitis that may prove fatal. An obliterative bronchiolitis may appear after bone marrow replacement for leukemia and may cause shortness of breath and disability.

Exposure to oxides of nitrogen, which may occur from inhaling gas in silos, when welding in enclosed spaces such as boilers, after blasting underground, or in fires involving plastic materials, is characteristically not followed by acute symptoms. These develop some hours later, when the victim develops a short cough and progressive shortness of breath. A chest radiograph shows patchy inflammatory change, and the lesion is an acute bronchiolitis. Symptomatic recovery may mask incomplete resolution of the inflammation.

An inflammation around the small airways, known as a respiratory bronchiolitis, is believed to be the earliest change that occurs in the lung in cigarette smokers, although it does not lead to symptoms of disease at that stage. The inflammation is probably reversible if smoking is discontinued. It is not known whether those who develop this change (after possibly only a few years of smoking) are or are not at special risk of developing the long-term changes of chronic bronchitis and emphysema.

Early changes due to smoking

Chronic bronchitis. The chronic cough and sputum production of chronic bronchitis were once dismissed as nothing more than "smoker's cough," without serious implications. But the striking increase in mortality from chronic bronchitis and emphysema that occurred after World War II in all Western countries indicated that the long-term consequences of chronic bronchitis could be serious. This common condition is characteristically produced by cigarette smoking. After about 15 years of smoking, a blob of mucus is coughed up in the morning, owing to an increase in size and number of mucous glands lining the large airways. The increase in mucous cells and the development of chronic bronchitis may be enhanced by breathing polluted air (particularly in areas of uncontrolled coal burning) and by a damp climate. The changes are not confined to large airways, though these produce the dominant symptom of chronic sputum production. Changes in smaller bronchioles lead to obliteration and inflammation around their walls. All of these changes together, if severe enough, can lead to disturbances in the distribution of ventilation and perfusion in the lung, causing a fall in arterial oxygen tension and a rise in carbon dioxide tension. By the time this occurs, the ventilatory ability of the patient, as measured by the velocity of a single forced expiration, is severely compromised; in a cigarette smoker, ventilatory ability has usually been declining rapidly for some years. It is not clear what determines the severity of these changes, since many people can smoke for decades without evidence of significant airway changes, while others may experience severe respiratory compromise after 15 years or less of exposure.

Pulmonary emphysema. This irreversible disease consists of destruction of alveolar walls. It occurs in two forms, centrilobular emphysema, in which the destruction begins at the centre of the lobule, and panlobular (or panacinar) emphysema, in which alveolar destruction

Types of emphysema

occurs in all alveoli within the lobule simultaneously. In advanced cases of either type, this distinction can be difficult to make. Centrilobular emphysema is the form most commonly seen in cigarette smokers, and some observers believe it is confined to smokers. It is more common in the upper lobes of the lung (for unknown reasons) and probably causes abnormalities in blood gases out of proportion to the area of the lung involved by it. By the time the disease has developed, some impairment of ventilatory ability has probably occurred. Panacinar emphysema may also occur in smokers, but it is the type of emphysema characteristically found in the lower lobes of patients with a deficiency in the antiproteolytic enzyme known as alpha$_1$-antitrypsin. Like centrilobular emphysema, panacinar emphysema causes ventilatory limitation and eventually blood gas changes. Other types of emphysema, of less importance than the two major varieties, may develop along the dividing walls of the lung (septal emphysema) or in association with scars from other lesions.

A major step forward in understanding the development of emphysema followed the identification, in Sweden, of families with an inherited deficiency of alpha$_1$-antitrypsin, an enzyme essential for lung integrity. Members of affected families commonly developed panacinar emphysema in the lower lobes, unassociated with chronic bronchitis but leading to ventilatory impairment and disability. Intense investigation of this major clue led to the "protease-antiprotease" theory of emphysema. It is postulated that cigarette smoking either increases the concentration of protease enzymes released in the lung (probably from white blood cells), or impairs the lung's defenses against these enzymes, or both. Although many details of the essential biochemical steps at the cellular level remain to be clarified, this represents a major step forward in understanding a disease whose genesis was once ascribed to overinflation of the lung (like overdistending a bicycle tire).

Theories concerning smoking and emphysema

Chronic bronchitis and emphysema are distinct processes. Both may follow cigarette smoking, however, and they commonly occur together, so determination of the extent of each during life is not easy. In general, significant emphysema is more likely if ventilatory impairment is constant, gas transfer in the lung (usually measured with carbon monoxide) is reduced, and the lung volumes are abnormal. The radiological technique of computerized tomography may improve the accuracy of detection of emphysema. Many people with emphysema suffer severe incapacity before the age of 60; thus, emphysema is not a disease of the elderly only. A reasonably accurate diagnosis can be made from pulmonary function tests, careful radiological examination, and a detailed history. The physical examination of the chest reveals evidence of airflow obstruction and overinflation of the lung, but the extent of lung destruction cannot be reliably gauged from these signs, and therefore laboratory tests are required.

Signs and symptoms

The prime symptom of emphysema, which is always accompanied by a loss of elasticity of the lung, is shortness of breath, initially on exercise only, and associated with loss of normal ventilatory ability. The severity of this loss is a predictor of survival in this condition. But once ventilatory ability is reduced to less than half the normal value, what determines outcome is the severity of the changes in blood gases, chiefly the lowering of arterial blood oxygen tension. The chronic hypoxemia (lowered oxygen tension) is believed to lead to the development of increased blood pressure in the pulmonary circulation, which in turn leads to failure of the right ventricle of the heart. The symptom (subjective evidence perceived by the patient) of right ventricular failure is swelling of the ankles; the signs (objective evidence discovered by the examining physician) are engorgement of the neck veins and enlargement of the liver. These are portents of advanced lung disease in this condition. The hypoxemia may also lead to an increase in total hemoglobin content and in the number of circulating red blood cells, as well as to psychological depression, irritability, loss of appetite, and loss of weight. Thus the advanced syndrome of chronic obstructive lung disease may cause not only such shortness of breath that the afflicted person is unable to dress without assistance, but also numerous other symptoms.

The slight fall in ventilation that normally accompanies sleep may exacerbate the failure of lung function in chronic obstructive lung disease, leading to a further fall in arterial oxygen tension and an increase in pulmonary arterial pressure.

Unusual forms of emphysema also occur. In one form the disease appears to be unilateral, involving one lung only and causing few symptoms. Unilateral emphysema is believed to result from a severe bronchiolitis in childhood that prevented normal maturation of the lung on that side. "Congenital lobar emphysema" of infants is usually a misnomer, since there is no alveolar destruction. It is most commonly caused by overinflation of a lung lobe due to developmental malformation of cartilage in the wall of the major bronchus. Such lobes may have to be surgically removed to relieve the condition.

Bronchiectasis. Bronchiectasis is believed usually to begin in childhood, possibly after a severe attack of whooping cough or pneumonia. It consists of a dilatation of major bronchi. The bronchi become chronically infected, and excess sputum production and episodes of chest infection are common. The disease may develop as a consequence of airway obstruction or of undetected (and therefore untreated) aspiration into the airway of small foreign bodies such as plastic toys.

Bronchiectasis may also develop as a consequence of inherited conditions, of which the most important is the familial disease of cystic fibrosis. The essential defect in this condition probably has to do with the transport of sodium and other ions across the wall of membranes, but it is not yet completely understood. The most important consequence of cystic fibrosis, apart from the malnutrition it causes, is the development of chronic pulmonary changes, with repetitive infections and bronchiectasis as characteristic features. This condition does not progress to pulmonary emphysema but rather causes obliteration and fibrosis of small airways and dilation and infection of the larger bronchi. Thick, viscid secretions in the bronchial tree are difficult to expectorate. However, the modern management of the condition, with the control of pulmonary infections, has markedly improved survival in affected persons, many of whom, who would formerly have died in childhood, now reach adult life.

Cystic fibrosis

OCCUPATIONAL LUNG DISEASE

Silica dust produces a distinctive reaction in the lung that eventually leads to the development of masses of fibrous tissue and distinctive nodules of dense fibrosis, which, by contracting, distort and damage the lung. Silicosis is a hazard in any occupation in which workers are exposed to silica dust, particularly rock drilling above or below ground, quarrying, or grinding with a wheel containing silica. Cases have also been reported in dental technicians, who use the material ground into a fine powder. Silicosis is usually fairly easy to detect on radiographs, and in its later stages it causes considerable shortness of breath and reduction of the vital capacity (a maximal breath). Sandblasting without respiratory protection is exceedingly dangerous, and fatal cases of acute silicosis caused by unprotected sandblasting have been reported. The dangers of silica are generally well recognized, and better protection has reduced the incidence of this condition. The disease may advance, with increasing disability, for years after the person has stopped inhaling the dust.

Silicosis

Coal dust alone, even if its silica content is very low, causes a distinctive pattern of change in the lung known as coalworker's pneumoconiosis (also called black lung). Initially the dust is deposited in the terminal bronchioles, where it causes a fibrotic reaction. At this stage there is little disability, but later the disease may progress to a more generalized form, and in some instances large masses of fibrotic tissue form in the lung. This condition, known as progressive massive fibrosis, is usually associated with severe disability and the risk of a secondary heart failure. It is not clear whether this stage is more likely to develop if pulmonary tuberculosis is superimposed on the respiratory damage caused by coal dust inhalation.

The widespread use of asbestos as an insulating material during World War II, and later in flooring, ceiling tiles,

brake linings, and as a fire protectant sprayed inside buildings, led to a virtual epidemic of asbestos-related disease 20 years later. At first only the form of disease known as asbestosis, with radiographic changes and impaired function at an early stage, was recognized. Then it became apparent that exposure to much less asbestos than was needed to cause asbestosis led to thickening of the pleura, and, when both cigarette smoking and asbestos exposure occurred, there was a major increase in the risk for lung cancer. It is currently believed that the risks from smoking and from significant asbestos exposure are multiplicative in the case of lung cancer. Finally, a malignant tumour of the pleura known as mesothelioma was found to be caused almost exclusively by inhaled asbestos. Often a period of 20 years or more elapsed between exposure to asbestos and the development of the tumour.

Asbestosis

As far as is known, all the respiratory changes associated with asbestos exposure are irreversible. Malignant mesothelioma is rare and unrelated to cigarette smoking, but survival after diagnosis is less than two years. Usually the pleural thickenings are not associated with disturbance of function or symptoms, although in occasional cases the pleuritis is more aggressive, in which case both may occur. It is not yet understood why asbestos causes such devastating changes. Furthermore, not all types of asbestos are equally dangerous; the risk of mesothelioma in particular appears to be much higher if crocidolite, a blue asbestos that comes from South Africa, is inhaled than if chrysotile is inhaled. But exposure to any type of asbestos is believed to increase the risk of lung cancer when associated with cigarette smoking. There has been much discussion of the advisability of removing asbestos from all buildings. It has been argued on the one hand that this is the only responsible policy, and on the other that more exposure might follow the careless removal of the material than would occur if it were left in place. All industrialized countries have imposed strict regulations for handling asbestos, and the work force is generally aware of the material's dangers.

The increasing use of man-made mineral fibres (as in fibreglass and rock wool) has led to concern that these may also be dangerous when inhaled; present evidence suggests that they do increase the risk of lung cancer in persons occupationally exposed to them. Standards for maximal exposure have been proposed.

The toxicity of beryllium was first discovered when it was widely used in the manufacture of fluorescent light tubes shortly after World War II. Beryllium causes the formation of granulomas in the lung and alveolar wall thickening, often with considerable disability as a result. Although beryllium is no longer used in the fluorescent light industry, it is still important in the manufacture of special steels and ceramics, and new cases of beryllium poisoning are occasionally reported.

It is not only inorganic minerals and dusts that may affect the lung. The dust produced in the processing of raw cotton may cause chronic obstructive lung disease. This does not have a characteristic pathology, however, and it does not give rise to emphysema. It is unclear whether cotton dust alone or the combination of cigarette smoke and cotton dust is particularly dangerous. The disease that results is known as byssinosis, or "brown lung." Workers in cotton plants in England used to complain of "Monday morning fever" and were found to suffer an easily measurable decrement in ventilatory function when they returned to work after spending a weekend away from the plant. The active particle or contaminant in the cotton dust that is responsible for the syndrome has not yet been identified.

Byssinosis

The dust from western red cedar may cause occupational asthma, and dust from the redwood and other trees may cause an acute hypersensitivity pneumonitis. Workers in the sugarcane industry may be affected by a similar syndrome, known as bagassosis; sisal workers also develop airflow obstruction.

Toluene diisocyanate, used in the manufacture of polyurethane foam, may cause occupational asthma at very low concentrations; in higher concentrations, such as may occur with accidental spillage, it causes a transient flulike illness associated with airflow obstruction. Prompt recognition of this syndrome has led to modifications in the industrial process involved.

Although the acute effects of exposure to many of these gases and vapours are well-documented, there is less certainty about the long-term effects of repeated low-level exposures over a long period of time. This is particularly the case when the question of whether work in a generally dusty environment has contributed to the development of chronic bronchitis or later emphysema—in other words, whether such nonspecific exposures increase the risk of these diseases in cigarette smokers. There is little unanimity on this question, but it is generally recognized that the differentiation is difficult.

Many chemicals can damage the lung in high concentration: these include oxides of nitrogen, ammonia, chlorine, oxides of sulfur, ozone, gasoline vapour, and benzene. In industrial accidents, such as occurred in 1985 in Bhopāl, India, and in 1976 in Seveso, near Milan, people in the neighbourhood of chemical plants were acutely exposed to lethal concentrations of these or other chemicals. The custom of transporting dangerous chemicals by rail or road has led to the occasional exposure of bystanders to toxic concentrations of gases and fumes. Although in many cases recovery may be complete, it seems clear that long-term damage may occur.

The assessment of disability and the writing of opinions on attributability have become important tasks for many respiratory specialists. Disability consequent upon a specific lung disease can be assessed by pulmonary function testing and in some cases by tests of exercise capability; these measures provide a good indication of the impact of the disease on the physical ability of the patient. It is much more difficult to decide how much of the disability is attributable to occupational exposure. If the exposure is historically known to cause a specific lesion in a significant percentage of exposed persons, such as mesothelioma in workers exposed to asbestos, attribution may be fairly straightforward. In many cases, however, the exposure may cause only generalized pulmonary changes leading to airflow obstruction or may cause lung lesions of multifactorial etiology, the precise cause of which cannot be determined by histological examination of the tissue. The question of attributability in these instances, already diffuse, may be complicated by a history of cigarette smoking, which may be mild or moderate, or of short or long duration. Physicians asked to give opinions on attributability in multifactorial disease processes before a legal body frequently must rely on the application of probability statistics to the individual case, a not wholly satisfactory procedure to those who must assign compensation and disability benefits.

Assessment of disability

LUNG CANCER

Up to the time of World War II, cancer of the lung was a relatively rare condition. The increase in its incidence in Europe after World War II was at first ascribed to better diagnostic methods, but by 1956 it had become clear that the rate of increase was too great to be accounted for in this way. At that time the first epidemiological studies began to indicate that a long history of cigarette smoking was associated with a great increase in risk of death from lung cancer. By 1965 cancer of the lung and bronchus accounted for 43 percent of all cancers in the United States in men, an incidence nearly three times greater than that of the second most common cancer (of the prostate gland) in men, which accounted for 16.7 percent of cancers. The 1964 *Report of the Advisory Committee to the Surgeon General of the Public Health Service* (United States) concluded categorically that cigarette smoking was causally related to lung cancer in men. Since then, many further studies in diverse countries have confirmed this conclusion.

Cigarette smoking and lung cancer

The incidence of lung cancer in women began to rise in 1960 and continued rising through the mid-1980s. This is believed to be explained by the later development of heavy cigarette smoking in women compared with men, who greatly increased their cigarette consumption during World War II. By 1988 there was evidence suggesting that the peak incidence of lung cancer due to cigarette smok-

ing in men may have been passed. The incidence of lung cancer mortality in women, however, is increasing.

The reason for the carcinogenicity of tobacco smoke is not known. Tobacco smoke contains many carcinogenic materials, and although it is assumed that the "tars" in tobacco smoke probably contain a substantial fraction of the cancer-causing condensate, it is not yet established which of these is responsible. In addition to its single-agent effects, cigarette smoking greatly potentiates the cancer-causing proclivity of asbestos fibres, increases the risk of lung cancer due to inhalation of radon daughters (products of the radioactive decay of radon gas), and possibly also increases the risk of lung cancer due to arsenic exposure. Cigarette smoke may be a promoter rather than an initiator of lung cancer, but this question cannot be resolved until the process of cancer formation is better understood. Recent data suggest that those who do not smoke but who live or work with smokers and who therefore are exposed to environmental tobacco smoke may be at increased risk for lung cancer, eloquent testimony to the power of cigarettes to induce or promote the disease.

Varied symptoms of lung cancer

Because lung cancer is caused by different types of tumour, because it may be located in different parts of the lung, and because it may spread beyond the lungs at an early stage, the first symptoms noted by the patient vary from blood staining of the sputum, to a pneumonia that does not resolve fully with antibiotics, to shortness of breath due to a pleural effusion; the physician may discover distant metastases to the skeleton, or in the brain that cause symptoms unrelated to the lung. Lymph nodes may be involved early, and enlargement of the lymph nodes in the neck may lead to a chest examination and the discovery of a tumour. In some cases a small tumour metastasis in the skin may be the first sign of the disease. Lung cancer may develop in an individual who already has chronic bronchitis and who therefore has had a cough for many years. The diagnosis depends on securing tissue for histological examination, although in some cases this entails removal of the entire neoplasm before a definitive diagnosis can be made.

Survival from lung cancer has improved very little in the past 40 years. Early detection with routine chest radiographs has been attempted, and large-scale trials of routine sputum examination for the detection of malignant cells have been conducted, but neither screening method appears to have a major impact on mortality. Therefore, attention has been turned to prevention by every means possible. Foremost among them are efforts to inform the public of the risk and to limit the advertising of cigarettes. Steps have been taken to reduce asbestos exposure, both in the workplace and in public and private buildings, and to control air pollution. The contribution of air pollution to the incidence of lung cancer is not known with certainty, though there is clearly an "urban" factor involved.

Exposure to radon

Persons exposed to radon daughters are at risk for lung cancer. The hazard from exposure was formerly thought to be confined to uranium miners, who, by virtue of their work underground, encounter high levels of these radioactive materials. However, significant levels of radon daughters have been detected in houses built over natural sources, and with increasingly efficient insulation of houses, radon daughters may reach concentrations high enough to place the occupants at risk for lung cancer. A recent survey of houses in the United States indicated that about 2 percent of all houses had a level of radon daughters that posed some risk to the occupants. Major regional variations in the natural distribution of radon occur, and it is not yet possible to quantify precisely the actual magnitude of the risk. In some regions of the world (such as the Salzburg region of Austria) levels are high enough that radon daughters are believed to account for the majority of cases of lung cancer in nonsmokers.

Workers exposed to arsenic in metal smelting operations, and the community around the factories from which arsenic is emitted, have an increased risk for lung cancer. Arsenic is widely used in the electronics industry in the manufacture of microchips, and careful surveillance of this industry may be needed to prevent future disease.

Some types of lung cancer are unrelated to cigarette smoking. Alveolar cell cancer is a slowly spreading condition that affects men and women in equal proportion and is not related to cigarette smoking. Pulmonary adenocarcinoma of the lung also has a more equal sex incidence than other types, and although its incidence is increased in smokers, it may also be caused by other factors.

It is common to feel intuitively that one should be able to apportion cases of lung cancer among discrete causes, on a percentage basis. But in multifactorial disease, this is not possible. Although the incidence of lung cancer would probably be far lower without cigarette smoking, the contribution of neither this factor nor any of the other factors mentioned can be precisely quantified.

MISCELLANEOUS PULMONARY CONDITIONS

Nonoccupational diffuse interstitial fibrosis. This condition has many names, including cryptogenic fibrosing alveolitis. It is a diffuse disease characterized by thickening of the alveolar walls, which may become so severe as to interfere with gas transport, and hypoxemia (reduced oxygen supply to tissues) results. Hypoxemia particularly occurs with exercise, since exercise reduces the time the blood spends in the alveolar capillary; and when the wall is thickened, there is insufficient time for oxygen to pass into the blood and achieve full saturation. Thus an early indication of abnormality in this condition is a significant fall in the pressure of oxygen in the arterial blood during exercise. The pulmonary diffusing capacity is usually reduced when this occurs. The cause of diffuse interstitial fibrosis is not known, although it sometimes appears after a viral infection; it also occurs in some diseases of the collagen system. It may run a relatively short course, but often it is slowly progressive over several years and leads eventually to secondary heart failure. The radiographic appearances are characteristic, although in some instances a lung biopsy may be necessary to establish the diagnosis. In a few patients the course of the disease seems to be improved by steroid treatment, but in many cases the downhill course of the disease is uninfluenced by it.

Early indications

Sarcoidosis. Sarcoidosis is a disease characterized by the development of small aggregations of cells, or granulomas, in different organs; the lung is commonly involved. Other common changes are enlargement of the lymph glands at the root of the lung, skin changes, inflammation in the eye, and liver dysfunction; occasionally there is inflammation of nerve sheaths, leading to signs of involvement in the affected area. The kidney is not commonly involved, but some changes in blood calcium levels occur in a small percentage of cases. The disease seems to affect the Scandinavian population more often than other Caucasians; and in the North American population it is more severe in blacks than in whites. In most cases the disease is first detected on chest radiographs. Evidence of granulomas in the lung may be visible, but often there is little interference with lung function. The disease usually remits without treatment within the next year or so, but in a small proportion of cases it progresses, leading finally to lung fibrosis and retraction.

Eosinophilic granuloma. This disease causes granulomas associated with masses of eosinophil cells, a subgroup of the white blood cells. It also causes lesions in bone. Eosinophilic granuloma is a lung condition that may spontaneously "burn out," leaving the lung with some permanent cystic changes. Its cause is not known.

Alveolar proteinosis. Alveolar proteinosis is characterized by accumulation of protein-rich material in the alveolar spaces of the lung. It is not usually associated with irreversible changes in the lung but runs a remittent course and finally may resolve completely. Its cause is unknown. It is often treated by lavage (washing out) of the lung with saline during bronchoscopy. The disease produces a characteristic radiographic picture and is not associated with changes in any other organ.

Immunologic conditions. The lung is often affected by generalized diseases of the blood vessels. Periarteritis nodosa, an acute inflammatory disease of the blood vessels believed to be of immunologic origin, is an important cause of pulmonary blood vessel inflammation. Acute hemorrhagic pneumonitis occurring in the lung in associa-

tion with changes in the kidney is known as Goodpasture's syndrome. The condition has been successfully treated by exchange blood transfusion, but its cause is not fully understood. Pulmonary hemorrhage also occurs as part of a condition known as pulmonary hemosiderosis, which results in the accumulation of the iron-containing substance hemosiderin in the lung tissues. The lung may also be involved in a variety of ways by the disease known as systemic lupus erythematosus, which is also believed to have an immunologic basis. Pleural effusions may occur, and the lung parenchyma may be involved. These conditions have only recently been recognized and differentiated; accurate diagnosis has been much improved by refinements in radiological methods, by the use of pulmonary function tests, and especially by improvement in thoracic surgical techniques and anesthesia that have made lung biopsy much less dangerous than it formerly was.

The common condition of rheumatoid arthritis may be associated with scattered zones of interstitial fibrosis in the lung or with solitary isolated fibrotic lesions. More rarely, a slowly obliterative disease of small airways (bronchiolitis) occurs, leading finally to respiratory failure.

Radiation damage. The lung may be damaged by irradiation of the chest wall in the treatment of cancer of the breast and other conditions. About three weeks or so after the end of the treatment, a pneumonitis may develop in the underlying lung, signaled by an irritant and unproductive cough. The condition may resolve, but in a few cases the lung becomes fibrotic and contracts to a small fraction of its normal volume. There is considerable individual variation in the response to the same dose of radiation.

Circulatory disorders. The lung is commonly involved in disorders of the circulation. The most important and common of these is blockage of a branch of the pulmonary artery by blood clot, which has usually formed in the veins of the legs or the pelvis. The resulting pulmonary embolus leads to changes in the lung supplied by the affected artery. These changes are known as a pulmonary infarction. The consequences of embolism range from sudden death, when the infarction is massive, to an increased respiratory rate, slight fever, and occasionally some pleuritic pain over the site of the infarction. People are at increased risk for pulmonary embolism whenever they are immobilized in bed and the circulation is sluggish, and particularly during the postoperative period. Early mobilization after surgery or childbirth is considered an important preventive measure. Repetitive small emboli may lead to pulmonary thromboembolic disease, in which the pressure in the main pulmonary artery is permanently increased. This disease is believed to be more common in women who have regularly used birth-control pills, but it is not confined to them.

Pulmonary embolism

In primary pulmonary hypertension, a condition of unknown origin, a marked increase in pulmonary arterial pressure occurs, with changes in the small radicles of the pulmonary artery that may not be clearly embolic in nature. Both repetitive thromboembolic disease and primary pulmonary hypertension eventually lead to failure of the right ventricle of the heart, usually after increasing disability with severe shortness of breath.

Congestion of the lungs and the development of fluid in the pleural cavity, with consequent shortness of breath, follows left ventricular failure, usually as a consequence of coronary arterial disease. When the valve between the left atrium of the heart and the left ventricle is thickened and deformed by rheumatic fever (mitral stenosis), chronic changes develop in the lung as a result of the increased pressure in the pulmonary circulation. These changes contribute to the shortness of breath in that condition and account for the not-infrequent blood staining of the sputum.

Diseases of the pleura. The pleura lining the lung may become perforated and spontaneously rupture, usually over a small collection of congenital blebs or cysts at the apex of the lung. This causes spontaneous pneumothorax, a partial or occasionally complete collapse of the lung. In the majority of cases a pneumothorax resolves slowly of its own accord, although pleural suction may be needed to expedite recovery. If repetitive attacks occur, the blebs may be removed surgically, and the pleura lining the lung may be sealed to the pleura lining the inner wall of the thorax to prevent a recurrence.

Spontaneous pneumothorax

The most common disease of the pleura is inflammatory. A pleurisy with an effusion may be the presenting symptom of pulmonary tuberculosis, and pleurisy may accompany any kind of pneumonia, though it is rare in viral infections. When a pleural effusion in a person with bacterial pneumonia becomes infected, pus accumulates in the pleural cavity (empyema). This complication, dreaded before the antibiotic era, required drainage of the pleural space. Such episodes are now rarely seen as a result of acute infections, but draining sinuses may still occur in pulmonary tuberculosis or fungal infections. Infection of the lung and later the pleural cavity by the moldlike bacteria *Actinomyces* and *Nocardia* is particularly likely to lead to this complication.

Diseases of the mediastinum and diaphragm. The mediastinum comprises the fibrous membrane in the centre of the thoracic cavity, together with the many important structures situated within it. Enlargement of lymph glands in this region is common, particularly in the presence of lung tumours or as part of a generalized enlargement of lymphatic tissue in disease. Primary tumours of mediastinal structures may arise from the thymus gland or the lower part of the thyroid gland; noninvasive cysts of different kinds are also found in the mediastinum.

The diaphragm may be incompletely formed, leading to herniation of abdominal viscera through it. In adult life the important disease involving the diaphragm is bilateral diaphragmatic paralysis. This leads to a severe reduction in the vital capacity when the subject is recumbent, although exercise capability may be relatively well preserved. In many cases the cause of the paralysis cannot be determined. The function of the diaphragm may be compromised when the lung is highly overinflated, as occurs in emphysema; diaphragmatic fatigue may limit the exercise capability of affected persons.

Acute respiratory distress syndrome of adults. Bacterial or viral pneumonia, exposure of the lung to gases, aspiration of material into the lung (including water in near-drowning episodes), or any generalized septicemia (blood poisoning) or severe lung injury may lead to capillary leakage throughout the lung, a syndrome known as the acute respiratory distress syndrome of adults. It was first recognized when cases occurred following septicemia induced by intrauterine birth control devices; however, this is only one pathway to a generalized lung injury that has many causes. Acute respiratory distress syndrome carries about a 50 percent mortality, later infection by certain types of bacteria being particularly serious. Life-support treatment with assisted ventilation rescues some patients, but there has been little improvement in survival during the 1970s and '80s. Recovery and repair of the lung may take months after clinical recovery from the acute event.

Effects of air pollution on health. The disastrous fog and attendant high levels of sulfur dioxide and particulate pollution (and probably also sulfuric acid) that occurred in London in the second week of December 1952 led to the deaths of more than 4,000 people during that week and the subsequent three weeks. Many, but not all, of the victims already had chronic heart or lung disease. Prize cattle at an agricultural show also died in the same period as a result of the air pollution. This episode spurred renewed attention to this problem, which had been intermittently considered since the 14th century in England, and finally the passage of legislation banning open coal burning, the factor most responsible for the pollution. This form of pollution, common in many cities using coal as heating fuel, was associated with excess mortality and increased prevalences of chronic bronchitis, respiratory tract infections in the young and old, and possibly lung cancer. Many industrial cities in Western countries now have legislation restricting the use of specific fuels and mandating emission-control systems in factories.

The 1952 London epidemic

In 1952 a different kind of air pollution was characterized for the first time in Los Angeles. The large number of automobiles in that city, together with the bright sunlight and frequently stagnant air, leads to the formation

Smog

of photochemical smog. This begins with the emission of nitrogen oxide during the morning commuting hour, followed by the formation of nitrogen dioxide by oxygenation, and finally, through a complex series of reactions in the presence of hydrocarbons and sunlight, to the formation of ozone and peroxyacetyl nitrite and other irritant compounds. Eye irritation, chest irritation with cough, and possibly the exacerbation of asthma occur as a result. It is now recognized that ozone is formed in many large cities of the world. Modern air pollution consists of some combination of the reducing form consequent upon sulfur dioxide emissions, and the oxidant form, which begins as emissions of nitrogen oxides. Ozone is the most irritant gas known. In controlled exposure studies it reduces the ventilatory capability of healthy people in concentrations as low as 0.12 parts per million. These levels are commonly exceeded in many places, including Mexico City, Bangkok, and São Paulo, where there is a high automobile density and the meteorologic conditions favour the formation of photochemical oxidants. Although acute episodes of communal air exposure leading to demonstrable mortality are unlikely, there is much concern over the possible long-term consequences of brief but repetitive exposures to oxidants and acidic aerosols; such exposures are now common in the lives of millions of people. Their impact has not yet been precisely defined.

The indoor environment can be important in the genesis of respiratory disease. In developing countries, disease may be caused by inhalation of fungi from roof thatch materials or by the inhalation of smoke when the home contains no chimney. In developed countries, exposure to oxides of nitrogen from space heaters or gas ovens may promote respiratory tract infections in children. Inhalation of tobacco smoke in the indoor environment by nonsmokers impairs respiration and may cause lung cancer. A tightly sealed house may act as a reservoir for radon seeping in from natural sources.

Acute carbon monoxide poisoning. Acute carbon monoxide poisoning is a common and dangerous hazard. The British physiologist J.S. Haldane pioneered the study of the effects of carbon monoxide at the end of the 19th century, as part of his detailed analysis of atmospheres in underground mines. Carbon monoxide is produced by incomplete combustion, including combustion of gas in automobile engines, and for a long period it was a major constituent of domestic gas made from coal (its concentration in natural gas is much lower). When the carbon

Mechanism of action

monoxide concentration in the blood reaches 40 percent (that is, when the hemoglobin is 40 percent saturated with carbon monoxide, leaving only 60 percent available to bind to oxygen), the subject feels dizzy and is unable to perform simple tasks; judgment is also impaired. Hemoglobin's affinity for carbon monoxide is 200 times greater than for oxygen, and in a mixture of these gases hemoglobin will preferentially bind to carbon monoxide; for this reason, carbon monoxide concentrations of less than 1 percent in inspired air seriously impair oxygen–hemoglobin binding capacity. The partial pressure of oxygen in the tissues in carbon monoxide poisoning is much lower than when the oxygen-carrying capacity of the blood has been reduced an equivalent amount by anemia, a condition in which hemoglobin is deficient. The immediate treatment for acute carbon monoxide poisoning is assisted ventilation with 100 percent oxygen.

The carbon monoxide inhaled by smokers who smoke more than two packs of cigarettes a day may cause up to 10 percent hemoglobin saturation with carbon monoxide. A 4 percent increase in the blood carbon monoxide level in patients with coronary artery disease is believed to shorten the exercise that may be taken before chest pain is felt, in those who have that symptom on exercise.

LUNG TRANSPLANTATION

Early attempts at transplanting a single lung in patients with severe bilateral lung disease were not successful, but from the late 1970s bilateral lung transplantation had some striking results. Persons severely disabled by cystic fibrosis, emphysema, sarcoidosis, pulmonary fibrosis, or severe primary pulmonary hypertension reportedly achieved nearly normal lung function several months after the procedure. Combined heart–lung transplantation has been attempted, with some good long-term results. One study of seven cases showed that ventilatory capacity after bilateral lung transplantation was about 75 percent of the normal value for that individual. Because transplantation offers the only hope for persons with severe lung disease, who may be relatively young, the techniques are being pursued aggressively in specialized centres. (D.V.B.)

BIBLIOGRAPHY

General features of the respiratory process: AUGUST KROGH, *The Comparative Physiology of Respiratory Mechanisms* (1941, reissued 1968), is classic in its field. JULIUS H. COMROE, JR., *Physiology of Respiration: An Introductory Text*, 2nd ed. (1974) covers the basic aspects of respiration in mammals. More recent texts include JOHN WIDDICOMBE and ANDREW DAVIES, *Respiratory Physiology* (1983), a good introduction; PETER SEBEL *et al., Respiration: The Breath of Life* (1985), an overview of respiration, the respiratory system, and its diseases; N. BALFOUR SLONIM and LYLE H. HAMILTON, *Respiratory Physiology*, 5th ed. (1987); and ALLAN H. MINES, *Respiratory Physiology*, 2nd ed. (1986). F. HAROLD MCCUTCHEON, "Organ Systems in Adaptation: The Respiratory System," in D.B. DILL (ed.), *Handbook of Physiology*, sect. 4, *Adaptation to the Environment* (1964), pp. 167–191, discusses respiration in relation to the environment, including chemical regulation, gas transport, and evolutionary patterns. STEPHEN C. WOOD (ed.), *Evolution of Respiratory Processes: A Comparative Approach* (1979), compares respiratory processes in modern animals to gain insights into evolutionary changes. DAVID J. RANDALL *et al., The Evolution of Air Breathing in Vertebrates* (1981), begins with the aquatic ancestral form.

Respiration in animals: Introductions to the field are provided by G.M. HUGHES, *Comparative Physiology of Vertebrate Respiration*, 2nd ed. (1974); RUFUS M.G. WELLS, *Invertebrate Respiration* (1980), a short but useful study; F. REED HAINSWORTH, *Animal Physiology: Adaptations in Function* (1981), which includes chapters on respiration, circulation, temperature, and energetics and their interplay; WILLIAM S. HOAR, *General and Comparative Physiology*, 3rd ed. (1983), in which phylogeny in animal functions is used as a framework for depicting animal physiology; MARTIN E. FEDER and WARREN W. BURGGREN, "Skin Breathing in Vertebrates," *Scientific American*, 253(5):126–142 (Nov. 1985); KNUT SCHMIDT-NIELSEN, *Animal Physiology: Adaptation and Environment*, 3rd ed. (1983), which explains systematically how animals cope with their environments; and a supplement to it, C. RICHARD TAYLOR, KJELL JOHANSEN, and LIANA BOLIS (eds.), *A Companion to "Animal Physiology"* (1982), which probes certain topics, including respiratory physiology. See also V.B. WIGGLESWORTH, *The Principles of Insect Physiology*, 7th ed. (1972, reprinted 1982), an excellent introduction to the form and function of insect respiration. C. LADD PROSSER, "Oxygen: Respiration and Metabolism," ch. 5 in C. LADD PROSSER (ed.), *Comparative Animal Physiology*, 3rd ed. (1973), pp. 165–211, is a comprehensive chapter on oxygen and its role. CHARLOTTE P. MANGUM, "Oxygen Transport in Invertebrates," *The American Journal of Physiology: Regulatory, Integrative and Comparative Physiology*, 248(5):R505–R514 (May 1985), provides a succinct overview of oxygen-carrying proteins.

(A.P.F./F.N.W.)

Respiration in humans: The design of the human respiratory system is covered by PETER H. BURRI, JOAN GIL, and EWALD R. WEIBEL, "Ultrastructure and Morphometry of the Human Lung," in THOMAS W. SHIELDS (ed.), *General Thoracic Surgery*, 2nd ed. (1983), pp. 18–42; and two essays in *Handbook of Physiology*, sect. 3, *The Respiratory System*, vol. 1, *Circulation and Nonrespiratory Functions*, ed. by ALFRED P. FISHMAN and ARON B. FISHER (1985); PETER H. BURRI, "Development and Growth of the Human Lung," pp. 1–46; and EWALD R. WEIBEL, "Lung Cell Biology," pp. 47–91.

(P.H.B.)

Control of breathing is described in *Handbook of Physiology*, sect. 3, *The Respiratory System*, vol. 2, *Control of Breathing*, 2 vol., ed. by NEIL S. CHERNIACK and JOHN G. WIDDICOMBE (1986); JACK L. FELDMAN, "Neurophysiology of Breathing in Mammals," in *Handbook of Physiology*, sect. 1, *The Nervous System*, vol. 4, *Intrinsic Regulatory Systems of the Brain*, ed. by FLOYD E. BLOOM (1986), pp. 463–524, an overview of the control of breathing; and the following journal articles: EUGENE N. BRUCE and NEIL S. CHERNIACK, "Central Chemoreceptors," *Journal of Applied Physiology*, 62(2):389–402 (Feb. 1987), a short review of advances in understanding the physiological mechanisms that account for the effect of carbon dioxide on breathing; NEIL S. CHERNIACK and MURRAY D. ALTOSE, "Mechanisms of Dyspnea," *Clinics in Chest Medicine*,

8(2):207–214 (June 1987), a brief description of the physiological basis of shortness of breath; HUGO LAGERCRANTZ and THEODORE A. SLOTKIN, "The 'Stress' of Being Born," *Scientific American,* 254(4):100–107 (April 1986), an article on breathing in the infant; MICHAEL E. LONG, "What Is This Thing Called Sleep?" *National Geographic,* 172(6):787–821 (Dec. 1987), an update on sleep physiology and apnea and on the kinds of research being conducted; and KINGMAN P. STROHL, NEIL S. CHERNIACK, and BARBARA GOTHE, "Physiologic Basis of Therapy for Sleep Apnea," *American Review of Respiratory Disease,* 134(6):791–802 (June 1986), on abnormal breathing during sleep and how it can be treated. (N.S.C.)

The principles of gas exchange in animals and humans are discussed in MALCOLM S. GORDON, *Animal Physiology: Principles and Adaptation,* 4th ed. (1982), a consideration of the mechanisms of gas exchange among animals; "Gas Exchange and Circulation," in R. MCNEILL ALEXANDER (ed.), *The Encyclopedia of Animal Biology* (1987), pp. 50–65; and *Handbook of Physiology,* sect. 3, *The Respiratory System,* vol. 4, *Gas Exchange,* ed. by LEON E. FARHI and S. MARSH TENNEY (1987), a critical, comprehensive presentation of physiological knowledge and concepts. (A.P.F./R.A.Kl.)

The interplay between respiration, circulation, and metabolism is outlined by EWALD R. WEIBEL, *The Pathway for Oxygen: Structure and Function in the Mammalian Respiratory System* (1984); R. GILLES (ed.), *Circulation, Respiration, and Metabolism: Current Comparative Approaches* (1985), essays on oxygen transport and utilization in animals; and C.R. TAYLOR *et al.,* "Adaptive Variation in the Mammalian Respiratory System in Relation to Energetic Demand," *Respiration Physiology,* 69(1):1–127 (July 1987), an entire issue devoted to the subject. (E.R.W.)

Adaptations of the respiratory system to high altitude are described in a comprehensive but readable manner in DONALD HEATH and DAVID REID WILLIAMS, *High-Altitude Medicine and Pathology* (1989). (Do.A.H.)

The effects of swimming and diving on respiration are detailed in PETER B. BENNETT and DAVID H. ELLIOTT (eds.), *The Physiology and Medicine of Diving,* 3rd ed. (1982); and N.R. ANTHONISEN, "Respiration," in CHARLES W. SHILLING, CATHERINE B. CARLSTON, and ROSEMARY A. MATHIAS (eds.), *The Physician's Guide to Diving Medicine* (1984), pp. 71–85, part of a chapter on the physiology of diving. (D.H.E.)

Diseases and disorders of respiration: The respiratory system is described in DAVID V. BATES, PETER T. MACKLEM, and RONALD V. CHRISTIE, *Respiratory Function in Disease: An Introduction to the Integrated Study of the Lung,* 2nd ed. (1971), a detailed text on impairment of lung function caused by disease; and ROBERT G. FRASER *et al., Diagnosis of Diseases of the Chest,* 2nd ed., 4 vol. (1977–79), with vol. 1 also available in a 3rd ed. (1988). H. CORWIN HINSHAW and JOHN F. MURRAY, *Diseases of the Chest,* 4th ed. (1980), is a general textbook covering diagnosis and treatment of chest diseases; see also J.G. SCADDING and GORDON CUMMING (eds.), *Scientific Foundations of Respiratory Medicine* (1981). STEVEN E. WEINBERGER, *Principles of Pulmonary Medicine* (1986), is an introductory text in which respiratory pathophysiology is considered from the clinical vantage. Comprehensive texts include GORDON CUMMING and STEPHEN J. SEMPLE, *Disorders of the Respiratory System,* 2nd ed. (1980); JOHN CROFTON and ANDREW DOUGLAS, *Respiratory Diseases,* 3rd ed. (1981); and IAN R. CAMERON and NIGEL T. BATEMAN, *Respiratory Disorders* (1983). ALFRED P. FISHMAN (ed.), *Pulmonary Diseases and Disorders,* 2nd ed., 3 vol. (1988), provides a comprehensive overview of pathophysiology as related to clinical syndromes. See also JOHN F. MURRAY and JAY A. NADEL (eds.), *Textbook of Respiratory Medicine* (1988); and WILLIAM M. THURLBECK (ed.), *Pathology of the Lung* (1988). (D.V.B./A.P.F.)

Rhetoric

The term rhetoric has traditionally applied to the principles of training communicators—those seeking to persuade or inform; in the 20th century it has undergone a shift of emphasis from the speaker or writer to the auditor or reader. This article deals with rhetoric in both its traditional and its modern forms. For information on applications of rhetoric, see the articles BROADCASTING, COMMUNICATION, and PROPAGANDA. For coverage of related topics in the *Macropædia* and the *Micropædia*, see the *Propædia*, sections 621 and 10/12, and the *Index*.

The article is divided into the following sections:

Rhetoric in literature 758
The nature and scope of rhetoric
Rhetorical traditions
Toward a new rhetoric
The rhetoric of non-Western cultures
Rhetoric in philosophy: the new rhetoric 762
Nature of the new rhetoric
Systematic presentation of the new rhetoric
Significance of the new rhetoric
Bibliography 764

Rhetoric in literature

THE NATURE AND SCOPE OF RHETORIC

Traditional and modern rhetoric. The traditional rhetoric is limited to the insights and terms developed by rhetors, or rhetoricians, in the Classical period of ancient Greece, about the 5th century BC, to teach the art of public speaking to their fellow citizens in the Greek republics and, later, to the children of the wealthy under the Roman Empire. Public performance was regarded as the highest reach of education proper, and rhetoric was at the centre of the educational process in western Europe for some 2,000 years. *Institutio oratoria* (before AD 96; "The Training of an Orator"), by the Roman rhetorician Quintilian, perhaps the most influential textbook on education ever written, was in fact a book about rhetoric. Inevitably, there were minor shifts of emphasis in so long a tradition, and for a long time even letter writing fell within the purview of rhetoric; but it has consistently maintained its emphasis upon creation, upon instructing those wishing to initiate communication with other people.

Function of rhetoric

Modern rhetoric has shifted its focus to the auditor or reader. Literary criticism always borrowed from rhetoric—stylistic terms such as antithesis and metaphor were invented by Classical rhetoricians. When language became a subject of sustained scholarly concern, it was inevitable that scholars would turn back to Classical theories of rhetoric for help. But modern rhetoric is far more than a collection of terms. The perspective from which it views a text is different from that of other disciplines. History, philosophy, literary criticism, and the social sciences are apt to view a text as though it were a kind of map of the author's mind on a particular subject. Rhetoricians, accustomed by their traditional discipline to look at communication from the communicator's point of view, regard the text as the embodiment of an intention, not as a map. They know that that intention in its formulation is affected by its audience. They know also that the structure of a piece of discourse is a result of its intention. A concern for audience, for intention, and for structure is, then, the mark of modern rhetoric. It is as involved with the process of interpretation, or analysis, as it is with the process of creation, or genesis.

The situations of a message

Rhetorical analysis is actually an analogue of traditional rhetorical genesis: both view a message through the situation of the auditor or reader as well as the situation of the speaker or writer. Both view the message as compounded of elements of time and place, motivation and response. An emphasis on the context automatically makes a rhetorician of the literary critic or interpreter and distinguishes that approach from the other kinds of verbal analysis. Critics who have insisted upon isolating, or abstracting, the literary text from the mind of its creator and from the milieu of its creation have found themselves unable to abstract it from the situation of its reader. Certain modern critics have joined with rhetoricians in denouncing the folly of all such attempts at abstraction. In interpreting any text—say a speech by Elizabeth I of England at Tilbury, Essex, or a play by the great Hindu poet of the 5th century, Kālidāsa—the rhetorician must imaginatively re-create the original situation of that text as well as endeavour to understand those factors that condition a present understanding.

All discourse now falls within the rhetorician's purview. Modern rhetoricians identify rhetoric more with critical perspective than with artistic product. They justify expanding their concerns into other literary provinces on the basis of a change in thinking about the nature of human reason. Modern philosophers of the Existentialist and Phenomenologist schools have strongly challenged the assumptions whereby such dualities as knowledge and opinion, persuasion and conviction, reason and emotion, rhetoric and poetry, and even rhetoric and philosophy have in the past been distinguished. The old line between the demonstrable and the probable has become blurred. According to these modern philosophers, a person's basic method of judgment is argumentation, whether in dialogue with others or with a text, and the results are necessarily relative and temporal. Such modern philosophers use legal battles in a courtroom as basic models of the process every person goes through in acquiring knowledge or opinion. For some, philosophy and rhetoric have become conflated, with rhetoric itself being a further conflation of the subject matter Aristotle discusses not only in his *Rhetoric* but also in his *Topics*, which he had designed for dialectics, for disputation among experts. According to this view, philosophers engage in a rhetorical transaction that seeks to persuade through a dialogic process first themselves and then, by means of their utterances, others. It is in this "argumentative" light that a rhetorically trained reader or auditor interprets all texts and justifies their inclusion within the province of rhetoric.

The courtroom model

Rhetoric has come to be understood less as a body of theory or as certain types of artificial techniques and more as an integral component of all human discourse. As a body of discursive theory, rhetoric has traditionally offered rules that are merely articulations of contemporary attitudes toward certain kinds of prose and has tended to be identified with orations in which the specific intent to persuade is most obvious. But modern rhetoric is limited neither to the offering of rules nor to studying topical and transient products of controversy. Rather, having linked its traditional focus upon creation with a focus upon interpretation, modern rhetoric offers a perspective for discovering the suffusion of text and content inhering within any discourse. And for its twin tasks, analysis and genesis, it offers a methodology as well: the uncovering of those strategies whereby the interest, values, or emotions of an audience are engaged by any speaker or writer through his discourse. The perspective has been denoted with the term situation; the methodology, after the manner of certain modern philosophers, may be denoted by the term argumentation. It should be noted at the outset that one may study not only the intent, audience, and structure of a discursive act but also the shaping effects of the medium itself on both the communicator and the communicant. Those rhetorical instruments that potentially work upon

an audience in a certain way, it must be assumed, produce somewhat analogous effects within the writer or speaker as well, directing and shaping his discourse.

Figures of speech

Elements of rhetoric. For the tasks imposed by the rhetorical approach some of the most important tools inherited from antiquity are the figures of speech: for example, the metaphor, or comparison between two ostensibly dissimilar phenomena, as in the famous comparison by the 17th-century English poet John Donne of his soul and his mistress's to the legs on a geometer's compass in his "A Valediction: Forbidding Mourning"; another is the allegory, the extended metaphor, as in John Bunyan's classic of English prose *Pilgrim's Progress* (1678, 1684), wherein man's method of earning Christian salvation is compared to a road on which he journeys, and the comparison is maintained to such an extent that it becomes the central structural principle of the entire work. Such figures may be said to pertain either to the texture of the discourse, the local colour or details, or to the structure, the shape of the total argument. Ancient rhetoricians made a functional distinction between trope (like metaphor, a textural effect) and scheme (like allegory, a structural principle). To the former category belong such figures as metaphor, simile (a comparison announced by "like" or "as"), personification (attributing human qualities to a nonhuman being or object), irony (a discrepancy between a speaker's literal statement and his attitude or intent), hyperbole (overstatement or exaggeration) or understatement, and metonymy (substituting one word for another which it suggests or to which it is in some way related—as part to whole, sometimes known as synecdoche). To the latter category belonged such figures as allegory, parallelism (constructing sentences or phrases that resemble one another syntactically), antithesis (combining opposites into one statement—"To be or not to be, that is the question"), congeries (an accumulation of statements or phrases that say essentially the same thing), apostrophe (a turning from one's immediate audience to address another, who may be present only in the imagination), enthymeme (a loosely syllogistic form of reasoning in which the speaker assumes that any missing premises will be supplied by the audience), *interrogatio* (the "rhetorical" question, which is posed for argumentative effect and requires no answer), and *gradatio* (a progressive advance from one statement to another until a climax is achieved). However, a certain slippage in the categories trope and scheme became inevitable, not simply because rhetoricians were inconsistent in their use of terms but because well-constructed discourse reflects a fusion of structure and texture. One is virtually indistinguishable from the other. Donne's compass comparison, for example, creates a texture that is not isolable from other effects in the poem; rather, it is consonant with a structural principle that makes the comparison both appropriate and coherent. Above all, a modern rhetorician would insist that the figures, like all elements of rhetoric, reflect and determine not only the conceptualizing processes of the speaker's mind but also an audience's potential response. For all these reasons figures of speech are crucial means of examining the transactional nature of discourse.

Rhetoric of or in a discourse. In making a rhetorical approach to various discursive acts, one may speak of the rhetoric *of* a discourse—say, Robert Browning's poem "My Last Duchess" (1842)—and mean by that the strategies whereby the poet communicated with his contemporaries, in this case the Victorians, or with modern man, his present readers; or one may speak of the rhetoric *in* a discourse and mean by that the strategies whereby the persona, the Duke of Ferrara who speaks Browning's poem in dramatic-monologue fashion, communicates with his audience in the poem, in this case an emissary from the father of Ferrara's next duchess. The two kinds of rhetoric are not necessarily discrete: in oratory or in lyric poetry, for example, the creator and his persona are assumed to be identical. To a degree Aristotle's distinction between the three voices of discourse still holds. A poet, according to Aristotle, speaks in his own voice in lyric poetry, in his own voice and through the voices of his characters in epic (or narrative), and only through the voices of his characters in drama. Thus, the speaker of oratory or of most nonfictional prose is similar to the lyric speaker, with less freedom than the latter either to universalize or to create imaginatively his own audience.

The three voices of discourse

RHETORICAL TRADITIONS

Although knowledge of rhetorical traditions is essential to the modern student's work, it must be borne in mind that he is nonetheless divorced from those traditions in two important ways. First, there is an almost exclusive emphasis upon the speaker or writer in traditional rhetoric; and, second, there is an implicit belief that the truth can be detached from the forms of discourse and can be divided into the demonstrable and the probable. In both of these respects, modern rhetorical practice differs.

Ancient Greece and Rome. Since the time of Plato it has been conventional to posit a correlative if not causal relationship between rhetoric and democracy. Plato located the wellsprings of rhetoric in the founding of democracy at Syracuse in the 5th century BC. Exiles returning to Syracuse entered into litigation for the return of their lands from which they had been dispossessed by the overthrown despotic government. In the absence of written records, claims were settled in a newly founded democratic legal system. To help litigants improve their persuasiveness, certain teachers began to offer something like systematic instruction in rhetoric.

In this experience at Syracuse, certain identifiable characteristics become prototypal: the rhetor, or speaker, is a pleader; his discourse is argumentative; and members of his audience are participants in and judges of a controversy. Later, in Athens, these characteristics began to aggregate to themselves some serious intellectual issues.

Role of the Sophists in Athens

In Athens early teachers of rhetoric were known as Sophists. These men did not simply teach methods of argumentation; rather, they offered rhetoric as a central educational discipline and, like modern rhetoricians, insisted upon its usefulness in both analysis and genesis. With the growth of Athenian democracy and higher systematized education, the Sophists became very powerful and influential. Today the word sophistic refers to a shabby display of learning or to specious reasoning; it refers, consequently, to an image of the Sophists that resulted from the attacks upon them led by such reformers as Plato. The ideal rhetoric proposed by Socrates in Plato's dialogue the *Phaedrus,* however, is itself not unlike the ideal sought by the Sophists in general, Isocrates in particular. Though the Platonic-Socratic ideal is more specialized in its focus on creating discourse, nonetheless, like the Sophistic ideal, it sought a union of verbal skills with learning and wisdom. Specifically, Platonic-Socratic rhetoric became a means of putting into practice the wisdom one acquires in philosophy. In this way Plato and Socrates resolved one of the most serious intellectual issues surrounding the subject: the relationship between truth and rhetorical effectiveness. The resolution, of course, presupposes and maintains a bifurcation between the two.

Aristotle's contributions to rhetoric

Aristotle, too, presupposed and maintained the same division between truth, which was knowable to varying degrees of certainty, and verbal skills, which for Aristotle were primarily useful in assisting truth to prevail in a controversy. But Aristotle lived in a world different from Plato's, one that was closer to the present in the premium it placed upon literacy and upon those patterns of thought that literacy encourages. The literate function of Aristotle's brilliance at recording and categorizing is well captured in Donne's phrase, "Nature's Secretary." Aristotle's *Rhetoric* both recorded contemporary practice and sought its reform through fitting it into its proper category among the arts. One of the masterstrokes of Aristotle's thought on the subject is his teaching that rhetoric itself is not a productive art of making but is an art of doing, embodying a power which is employed in certain kinds of speaking. Further evidence of his brilliance on the subject is his division of speaking into the forensic, the deliberative, and the epideictic and of persuasive appeals into the ethical, the emotional, and the logical. His division of speaking into three kinds reflects his efforts to distinguish rhetoric and its counterpart, dialectics, from philosophy

and science. Rhetoric and dialectics, he felt, are concerned with probable matters, in which there are several roads to truth; philosophy and science, on the other hand, are concerned with demonstrable matters, in which the roads are fewer but the truth more certain. In dividing persuasive appeals into three kinds, Aristotle indicated an unmistakable preference for the logical. This preference has been interpreted variously as a result of Aristotle's naïve assumption about the rationality of most audiences and as an attempt to reform the emotionally charged rhetoric of his contemporaries. In discussing elements of style, Aristotle treated metaphor, perhaps the major figure of speech, in a way that was to plague rhetoricians and poets for centuries. He describes it not as an instrument of thought but as an ornamentation, an adornment that at best serves the functions of clarity and vividness. The effect is further reflection of the principle noted earlier: for Aristotle the truth with which rhetoric is concerned is not demonstrable. It is, moreover, detachable from the forms of argument, and it can be tested by such analytical means as dialectics, which is the counterpart of rhetoric but which does not have what Aristotle viewed as rhetoric's cloying concerns with that beast of many heads, the heterogeneous audience composed of experts and laymen alike.

The Sophistic doctrine that rhetoric should be the central discipline in the educational scheme had a long history, rising to its fullest statement in the writings of Quintilian in Rome of the 1st century AD. By the age of Quintilian three intellectual issues had become firmly fixed within the orbit of rhetoric. Two of these were consciously faced: (1) the relationship between truth and verbal expression and (2) the difficulties of achieving intellectual or artistic integrity while communicating with a heterogeneous audience. In a sense, both of these issues were not faced at all but dodged, as they had been in the past, with the implicit assumption that wisdom and eloquence were not necessarily synonymous and that truth and integrity were ultimately dependent upon the character of the speaker. The orator, according to Cato the Elder, must be a good man skilled in speaking. Through the writings of Cicero, the ancient Roman orator of the 1st century BC whom later ages were to adulate both for his statesmanship and for his prose style, Cato's doctrine was spread in the Western world for centuries. Quintilian's tediously prescriptive *Institutio oratoria* is built on Cato's thesis: it offers an educational program for producing generations of Ciceronian statesmen. But for all its importance and influence, the work never found its time so far as being used as a text for political leaders to follow. Quintilian's program was impossible to achieve in the age of tyranny in which he lived, and it was impracticable in the Renaissance. Nevertheless, it was in the Renaissance that the *Institutio oratoria* began to be revered as the greatest educational treatise ever written.

Rhetorical issues in the age of Quintilian

A third issue arose in part as a consequence of literacy and in part as a consequence of social change: rhetoric became a productive art, but one whose role and status were unclear. The audience was no longer quite the full partner in the creative event that it had been in older days of freer public discussion; subsequently, from the classical period through the Middle Ages rhetoricians began to conceive of their art as a kind of methodical, solitary progress toward literary creation. Rhetoric was thought of less in terms of a power and more in terms of certain products of that power—orations; elaborate rules were given for distinguishing the kinds of orations and for arranging the material in them. Accompanying this shift, the entire creative process taught by the rhetoricians became linear and sequential in concept, with some activities located at further and further removes from the serious operations of the mind. A certain linearity, or step-by-step procedure, is evident in Aristotle's *Rhetoric,* but the attendant dangers of compartmentalization and fragmentation into increasingly trivial matters did not make themselves felt for centuries. By the time of Cicero, rhetoric was considered to be a discipline that encompassed five "offices": invention, analyzing the speech topic and collecting the materials for it; disposition, arranging the material into an oration; elocution, fitting words to the topic, the speaker, the audience, and the occasion; pronunciation or action, delivering the speech orally; and memory, lodging ideas within the mind's storehouse. Not only orations but also poems, plays, and almost every kind of linguistic product except those belonging peculiarly to logic (or dialectics) fell within the rhetoricians' creative art. Thus, the function of rhetoric appeared to be the systematic production of certain kinds of discourse, but the significance of this now clearly productive art became increasingly dubious in ages when governments did not allow public deliberation on social or political issues or when the most significant speaking was done by church authorities whose training had been capped by logic and theology.

The five "offices" of rhetoric

The Middle Ages. The early Church Father St. Augustine made one of the earliest efforts to write a rhetoric for the Christian orator. Book IV of *On Christian Doctrine* is usually considered the first rhetorical theory specifically designed for the minister. Of course, the kind of truth to which Augustine sought to give verbal effectiveness was the "revealed" truth as contained in the Scriptures. The first three books of *On Christian Doctrine,* which describe procedures for a proper interpretation of the Bible, actually set forth the invention part of Augustine's rhetoric. There is no basis here for replacing either logic or theology with rhetoric as the capstone of professional training. The work does represent, however, one of the first theoretical efforts to bring together interpretation—that is, interpreting a text, as opposed to interpreting the facts of a case—and rhetoric.

Late in the 13th century, two students of the German philosopher Albertus Magnus produced a great impact upon the thought—particularly the educational thought—of succeeding generations. Thomas Aquinas, who became in effect the preceptor of the theological curriculum, and Peter of Spain (later Pope John XXI), the preceptor of the general or "arts" curriculum, gave articulate force to the current educational practice of making logic the specialty toward which the professional student advanced beyond rhetoric. Thomas wrote on the logic of abstract, symbolic thought, and Peter wrote on the logic of dialectics, disputation among experts.

The Renaissance and after. In the 16th century, at a time marked by a tremendous growth of interest in creating vernacular rhetorics to satisfy a new self-consciousness in the use of native tongues, the French philosopher Petrus Ramus and his followers merely completed the incipient fragmentation of rhetorical theory by affirming the offices as discrete specialties. Invention and disposition were assigned to dialectics, by now largely a silent art of disputation which in the Ramist system placed a premium upon self-evident, axiomatic statements. Memory was considered not a matter of creating sound effects to enhance the memorization of the orator's ideas but a matter of effective disposition, so that separate attention to memory disappeared. Elocution and pronunciation were considered the only two offices proper to rhetoric, and these fell under peculiar opprobrium.

The Ramist system

Elocution, or style, became the centre of rhetorical theory, and in Ramist hands it was almost solely concerned with figures of speech. Actually, a strong emphasis upon the figures of speech had been evolving since the late Middle Ages. When responsibly taught, as linguistic postures, stances, gestures of the mind in confrontation with external reality, the figures served a useful purpose; and in Renaissance education they were widely employed, as in the modern manner, in the interpretation or analysis of discourse. Less responsibly taught, the figures became merely an ornamentation, like the metaphor in Aristotle. In the Ramistic system, the figures ranged between serving as arguments and serving as extrinsic decorations. The figures of speech fell into greater disrepute in the new culture of the Renaissance, which was marked not only by an enthusiasm for printed vernacular discourse in a "plain" style but also by an increasing perplexity over doctrines of the passions. For centuries rhetoricians had taught figures of speech as means of "amplifying" ideas so that they would appeal to the passions in an audience. With Ramus, rhetoric discarded its principles of amplification, leaving the passions to be discussed primarily by "moral

philosophers," who battled heatedly over which were ordinate and which were inordinate passions. Ultimately, the passions themselves became subjects, or objects, of the new scientists, who divorced them from moral or religious dogma. It was the end of the 18th century before doctrines of the passions fell once more within the rhetorician's purview; however, at that time the figures were regarded less as appeals to an audience's passions and more as manifestations of the author's or speaker's psychology—or, to use the metaphor employed earlier, as places on the map of his mind.

The other part of the fragmented Ramist rhetoric, pronunciation or action, was rarely mentioned in the Renaissance; it hath not yet been perfected, was the excuse the Ramists gave. The first real impetus for a scientizing of English oral delivery came at the beginning of the 17th century from Francis Bacon, who, in touching on rhetoric in his writings, called for a scientific approach to the study of gesture. The Ramists had created a context within which Bacon's call would have peculiar force and meaning. John Bulwer's *Chirologia* (1644) was the first work to respond, and in its wake came a host of studies of the physical, nonverbal expression of ideas and passions, including works by Charles Darwin and Alexander Melville Bell in the 19th century and modern writings on "silent language" by the American linguist Edward T. Hall.

The elocutionary movement

But, so far as rhetorical theory is concerned, even more significant attempts to specialize in the study of pronunciation or action came in the elocutionary movement of the 18th century, which was the first large-scale, systematic effort to teach reading aloud (oral interpretation). The elocutionists named their study for the third office of rhetoric partly because "pronunciation" was coming to refer solely to correct English phonation and partly because "elocution" had traditionally referred to the decorous expression of previously composed material. The most important elocutionists were actors or lexicographers, such as Thomas Sheridan and John Walker, both of whom acted in London and went on to write dictionaries in the late 18th century. At first glance, their efforts to describe or prescribe the oral delivery of written or printed discourse (poems, plays, as well as speeches) appear to operate on extremely inadequate theory: exactly how one discovered the meaning on the page seems mysterious, almost divinatory. Some of their efforts produced such absurdities as statelike posing or a contempt for the verbal later associated in America with the 19th-century French teacher of dramatic and musical expression François Delsarte. Yet, their efforts may also be seen as attempts to restore the voice to that entire language process which the page abstracted—as attempts to bridge the gap left in concepts of "natural" meaning by the decay of the oral traditions. Moreover, it is most significant that of all theorists within the history of rhetoric, the elocutionists were the first to place an exclusive concern upon interpreting discourse. Indeed, it was through the elocutionary emphasis upon interpretation that something like a meaningful restoration of pronunciation occurred within the rhetorical tradition.

Sheridan had found within the teachings of the 17th-century English philosopher John Locke a foundation on which the study of elocution could be built: words are the signs of ideas, tones the signs of passions. A new, virtually irrevocable split had apparently occurred between spoken language and printed or written discourse. But the split did not produce in other rhetoricians quite the anxiety it produced in the elocutionists. Other rhetoricians began to discover faculty psychology (*i.e.,* the obsolete notion that supposed faculties of the mind such as will and reason account for all human behaviour) and associationism (*i.e.,* the philosophy expostulated by the 18th-century Scot David Hume and others that most mental activity is based on the association of ideas). In these concepts they found a fragmented, compartmentalized means whereby a fragmented, compartmentalized rhetorical theory could recover part of its earlier vast province, as, for example, doctrines of the passions. Pathetic appeals could simply become, as in Hugh Blair's *Lectures on Rhetoric and Belles Lettres* (1783), something like the sixth office of rhetoric. Besides Blair's, the most important rhetorical treatises of the period were George Campbell's *Philosophy of Rhetoric* (1776) and Richard Whately's *Elements of Rhetoric* (1828). All three books were written by Protestant clerics, and all reveal the pervasive assumptions of the Age of Reason. Though rhetoric may involve the whole man—indeed, that is the very reason Campbell believed rhetoric properly seen is naturally allied with a science of the mind—nonetheless, man was viewed as an animal with higher and lower faculties, whose intellect was susceptible to being disordered by his passions and whose noble achievement was the creation of rational, preferably written, discourse.

Dissolution of rhetoric

Theories of rhetorical invention of the 18th and 19th centuries seldom treated the exigencies of oral composition before live audiences or even involved an imaginative projection of oneself into a public situation. Rather, they posited an inventive process that was silent, solitary, meditative—a process of conducting solitary, or inward, dialogues. Imagination, that faculty by which man may potentially synthesize what faculty psychology termed his rational and sensory experiences, was not vindicated philosophically until the Romantic movement of the 19th century (and perhaps never effectively). By that time, rhetoric had fallen into discredit. Printed matter had proliferated to such an extent that traditional principles of invention had become antiquated. Eventually all traditional techniques of style and all organized rhetorical study were devalued by interest in experiments; in Switzerland, cultural historian Jacob Burckhardt described antiquity's interest in rhetoric as a "monstrous aberration." In America, the Delsarteans, who stressed gesture rather than words, spread an antirhetorical approach to imagination, the passions, sensory experience, and delivery. Thus, well into the 20th century, "elocution" in popular speech meant florid delivery and "rhetoric," because of its principal concern with oratory, meant purple prose. In academic circles, "rhetoric" referred largely to principles of "belles lettres" until "belletristic" became a pejorative; then "rhetoric" in a host of college "composition" courses referred to less philosophically troublesome principles of paragraph development and thematic arrangement. More than the medieval logicians, more than Ramus, more than all Rationalist philosophers, and more than even the new philosophies of science, it was probably the very momentum of the revolution begun by Johannes Gutenberg's invention of the printing press that caused traditional rhetoric, both as an educational principle and as a theory, to go under.

TOWARD A NEW RHETORIC

These extremely negative views toward rhetoric prevailed until the 1930s, when attention to the importance of studying how language is used was stimulated by Logical Positivism, the philosophical movement that insists that all statements be verifiable by observation or experiment, and that movement had ironically been stimulated in turn by the very scientism that had earlier disparaged rhetoric. Substantial attempts were made, particularly in the United States, to develop an art of discourse suitable for teaching in schools and universities.

In the opening decades of the 20th century, an attempt was made in American universities to restore rhetoric to the serious study of communication (that is, of creating discourse). Teachers of public speaking were the first to turn to rhetorical traditions for help, followed by teachers of writing. (The teaching of speaking had been divorced from the teaching of writing in America since the third quarter of the 19th century—a divorce that has been recognized by modern universities but challenged by the temper of modern life.) Appropriately, considering the impetus of Logical Positivism, the restored rhetoric was largely Aristotelian, an Aristotelianism that was filtered through centuries of faculty psychology, that was becoming part of a doctrinaire stance against the Romantics and the elocutionists, and that was interpreted in terms of lingering presuppositions of a typographical age. Nonetheless, the rhetoric offered through the tenets of a restored Aristotelianism was potentially more comprehensive—more inclusive of all the offices of rhetoric—than any in Western education since the Renaissance. The political facts of

Restored Aristotelianism

modern life, however, made the Rationalist proclivity of this rhetoric appear naïve. The new media—films, radio, and television—and the new orality of modern life was felt by those interested in rhetoric as a challenge to older linguistic notions, not simply those of the print-oriented teachers of written or spoken composition but those of the Aristotelian Positivists as well.

Moreover, the restoration of traditional rhetoric was at first—within speech departments and then later within English departments—an attempt to serve as an emphasis upon training students in how to communicate. When modern rhetoricians shifted their emphasis to interpretation and shifted their concerns from the speaker or writer to the auditor or reader, traditional rhetoric was seen in a new perspective and the subject itself was given its strongest modern impetus and relevance. As noted earlier, the latter effect was the combined result of the work of modern philosophers and literary critics as well as educators.

The 20th century witnessed the publication of some highly provocative works on rhetoric, which potentially carry the subject beyond its Aristotelian confines and give it new relevance to an age dissatisfied with older epistemologies (or theories of knowledge) and their curious, divisive assumptions about truth and verbal expression or about oral and written discourse. Particular attention must be called to the work of the American critic and philosopher Kenneth Burke. A controversial writer, partly because of his extension of rhetoric into the study of nonverbal transactions and sensations, he has perhaps done more than anyone else to create a theoretical basis for the use of rhetoric in interpretation.

As noted at the opening of this article, modern literary critics have helped to free rhetoric from its traditional emphasis by proving its instrumentality in literary analysis—"practical criticism," as the English critic I.A. Richards called his 1929 book on the subject. But in turn the practical critic has helped preserve traditional rhetoric for the analyses of traditional literature, and through his work on modern literature, he has stimulated the demand for a new rhetoric.

THE RHETORIC OF NON-WESTERN CULTURES

Freed, too, of the parochialism engendered by its Western traditions, rhetoric could undertake a variety of analytical endeavour, even "cross-cultural" studies—for example, the mingling of Malaysian and Western cultures in the political oratory of the Philippines, structure and intention in the oral literatures of Africa, or the communicative strategy of the Japanese verse form haiku.

The search for the rhetoric of non-Western cultures

Indeed, the search for the rhetoric of non-Western cultures has become a crucial scholarly and political endeavour, as people seek bases for understanding the politics as well as the poetry of other lands—and, hopefully, bases for dialogue across tribal and national boundaries. The avenues this search has taken thus far reveal a significant fact both about rhetoric and about the nature of its Western tradition: the true rhetoric of any age and of any people is to be found deep within what might be called attitudinizing conventions, precepts that condition one's stance toward experience, knowledge, tradition, language, and other people. Searching for those precepts, the scholar realizes the extent to which Western culture has become secularized and compartmentalized. In Western culture one may seek out a body of writing under such special rubrics as "rhetoric," "religion," "ethics." But in some Oriental or Middle Eastern cultures, the search may begin and end with religious thought and practices. The Talmudic rabbis, with their disputatious hermeneutics and their attitudes toward Oral Law, gave centuries of Jews a pattern of reasoning and communication. No less so did the *Tao-te Ching*—the basic text of the Chinese religious system of Taoism—shape a mentality that is as inherent in certain Chinese poetry as in the oratory, dance, painting, architecture, and government of that ancient culture. And for all the Western studies one might encourage into the haiku, surely only one thoroughly grounded in the mysterious doctrines of Zen Buddhism can fully understand how that imagistic poetry itself "works." Moreover, as rhetorical doctrine, the form and function of the "sayings" of a modern, secular Oriental revolutionary may not be so far distant from the form and function of the ancient analects of the sage Confucius. Though rhetoric is to be found in every use of language, only Westerners have attempted to divide its precepts discretely from the great body of ethical, moral, or religious precepts that condition the very nature of a culture.

In sum, the basic rhetorical perspective is simply this: all utterance, except perhaps the mathematical formula, is aimed at influencing a particular audience at a particular time and place, even if the only audience is the speaker or writer himself; any utterance may be interpreted rhetorically by being studied in terms of its situation—within its original milieu or even within its relationship to any reader or hearer—as if it were an argument. (T.O.S.)

Rhetoric in philosophy: the new rhetoric

There is nothing of philosophical interest in a rhetoric that is understood as an art of expression, whether literary or verbal. Rhetoric, for the proponents of the new rhetoric, is a practical discipline that aims not at producing a work of art but at exerting through speech a persuasive action on an audience.

NATURE OF THE NEW RHETORIC

Theory of argumentation

The new rhetoric is defined as a theory of argumentation that has as its object the study of discursive techniques that aim to provoke or to increase the adherence of men's minds to the theses that are presented for their assent. It also examines the conditions that allow argumentation to begin and to be developed, as well as the effects produced by this development.

This definition indicates in what way the new rhetoric continues classical rhetoric and in what way it differs from it. The new rhetoric continues the rhetoric of Aristotle insofar as it is aimed at all types of hearers. It embraces what the ancients termed dialectics (the technique of discussion and debate by means of questions and answers, dealing especially with matters of opinion), which Aristotle analyzed in his *Topics;* it includes the reasoning that Aristotle qualified as dialectical, which he distinguished from the analytical reasoning of formal logic. This theory of argumentation is termed new rhetoric because Aristotle, although he recognized the relationship between rhetoric and dialectic, developed only the former in terms of the hearers.

It should be noted, moreover, that the new rhetoric is opposed to the tradition of modern, purely literary rhetoric, better called stylistic, which reduces rhetoric to a study of figures of style, because it is not concerned with the forms of discourse for their ornamental or aesthetic value but solely insofar as they are means of persuasion and, more especially, means of creating "presence" (*i.e.,* bringing to the mind of the hearer things that are not immediately present) through the techniques of presentation.

The elaboration of a rhetoric thus conceived has an undeniable philosophical interest because it constitutes a response to the challenge of Logical Empiricism. The Logical Empiricists proclaim the irrationality of all judgments of value—*i.e.,* those judgments that relate to the ends of men's actions—because such judgments can be grounded neither in experience nor in calculation, neither in deduction nor in induction. But it is not clearly necessary, after discarding the recourse to intuition as an insufficient basis for a judgment of value, to declare all such judgments equally arbitrary. This amounts to considering as futile the hopes of philosophers to elaborate a wisdom that would guide men in their public as well as their private lives. The alternative offered by the new rhetoric would furnish a complementary tool to traditional logic, which is limited to the technique of demonstration, or necessary proof according to the rules of deduction and induction; it would add the technique of argumentation. This would allow men not only to verify and to prove their beliefs but also to justify their decisions and their choices. Thus, the new rhetoric, elaborating a logic for judgments of value, is indispensable for the analysis of practical reasoning.

SYSTEMATIC PRESENTATION OF THE NEW RHETORIC

Adaptation to the audience

Personal relations with the audience. Argumentation, whether it be called rhetorical or dialectical, always aims at persuading or convincing the audience to whom it is addressed of the value of the theses for which it seeks assent. Because the purpose of all argumentation is to gain or reinforce the adherence of an audience, it must be prepared with this audience in mind. Unlike demonstration, it cannot be conceived in an impersonal manner. On the contrary, it is essential that it be adapted to the audience if it is to have any effectiveness. Consequently, the orator—the person who presents an argument either by speech or in writing to an audience of listeners or readers—must seek to build his argumentative discourse on theses already accepted by his audience. The principal fallacy in argumentation is the *petitio principii* ("begging of the question"), in which the speaker presupposes that the audience accepts a thesis that actually is contested by them, even implicitly.

Taken in a broad sense, the new rhetoric can treat the most varied questions and be addressed to the most diverse audiences. The audience may involve only the individual deliberating within himself or it may involve another person in a dialogue. The discourse may be addressed to various particular audiences or to the whole of mankind—to what may be called the universal audience—in which case the orator appeals directly to reason.

Classical rhetoric was traditionally addressed to an audience made up of a crowd of generally incompetent hearers gathered in a public place; argumentation, however, can be addressed to highly qualified audiences, such as the members of an academy or some learned society. As a result, effectiveness is not the only means of testing the value of an argument, for this value also depends on the quality and competence of the minds whose adherence is sought. An argument may persuade an audience of less informed persons and remain without effect on a more critical audience. For Plato, the argumentation worthy of a philosopher should convince the gods themselves.

Basis of agreement and types of argumentation. The orator, in order to succeed in his undertaking, must start from theses accepted by his audience and eventually reinforce this adherence by techniques of presentation that render the facts and values on which his argument rests present to the listener. Thus, the orator can have recourse to literary devices, using figures of rhetoric and other techniques of style and composition that are well known to writers.

If the discourse is addressed to a nonspecialized audience, its appeal will be to common sense and common principles, common values, and common *loci*, or "places." Agreement about common values is general, but their object is vague and ill-defined. Thus, the appeal to universal values, such as the good and the beautiful, truth and justice, reason and experience, liberty and humanity, will leave no one indifferent, but the consequences to be drawn from these notions will vary with the meaning attached to them by the different individuals. Therefore, an agreement about common values must be accompanied by an attempt to interpret and define them, so that the orator can direct the agreement to make it tally with his purposes. If the discourse is addressed to a specialized group—such as a group of philosophers or jurists or theologians—the basis of agreement will be more specific.

Types of arguments

To pass from the premises accepted by the audience to the conclusions he wishes to establish, the orator can use arguments of various types of association and dissociation. A detailed analysis of such arguments would require a whole treatise; the best known, however, are arguments by example, by analogy, by the consequences, *a pari* (arguing from similar propositions), *a fortiori* (arguing from an accepted conclusion to an even more evident one), *a contrario* (arguing from an accepted conclusion to the rejection of its contrary), and the argument of authority. The traditional figures of rhetoric are usually only abridged arguments, as, for instance, a metaphor is an abbreviated analogy.

Associative arguments transfer the adherence from the premises to the conclusion; for example, the act–person association enables one to pass from the fact that an act is courageous to the consequence that the agent is a courageous person. Argumentation leads to the dissociation of concepts if appearance is opposed to reality. Normally, reality is perceived through appearances that are taken as signs referring to it. When, however, appearances are incompatible—an oar in water looks broken but feels straight to the touch—it must be admitted, if one is to have a coherent picture of reality, that some appearances are illusory and may lead to error regarding the real. Because the status of appearance is equivocal, one is forced to distinguish between those appearances that correspond with reality and those that are only illusory. The distinction will depend on a conception of reality that can serve as a criterion for judging appearances. Whatever is conformable to this conception of the real will be given value; whatever is opposed to it will be denied value.

Every concept can be subjected to a similar dissociation of appearance and reality. Real justice, democracy, and happiness can be opposed to apparent justice, democracy, and happiness. The former, being in conformity with the criteria of what justice, democracy, and happiness really are, will keep the value normally attached to these notions. The apparent—what is taken for real by common sense or unenlightened opinion—will be depreciated because it does not correspond to what actually deserves the name of justice, democracy, or happiness. By means of this technique of dissociating concepts, philosophers can direct men's actions toward what they hold to be true values and can reject those values that are only apparent. Every ontology, or theory about the nature of being, makes use of this philosophical process that gives value to certain aspects of reality and denies it to others according to dissociations that it justifies by developing a particular conception of reality.

Scope and organization of argumentation. A discourse that seeks to persuade or convince is not made up of an accumulation of disorderly arguments, indefinite in number; on the contrary, it requires an organization of selected arguments presented in the order that will give them the greatest force. After its analysis of the various types of arguments, the new rhetoric naturally deals with the study of the problems raised by the scope of the argumentation, the choice of the arguments, and their order in the discourse.

Although formal demonstrative proof is most admired when it is simple and brief, it would seem theoretically that there would be no limit to the number of arguments that could be usefully accumulated; in fact, because argumentation is concerned not with the transfer from the truth of premises to a conclusion but with the reinforcement of the adherence to a thesis, it would appear to be effective to add more and more arguments and to enlarge the audience. Because the argumentation that has persuaded some may fail to have any effect on others, it would appear to be necessary to continue the search for arguments better adapted to the enlarged audience or to the fraction of the audience that has been hitherto ignored.

Limits to argumentation

In practice, however, three different reasons point to the need to set bounds to the scope of an argumentation. First, there are limits to the capacity and the will of an audience to pay attention. It is not enough for an orator to speak or write; he must be listened to or read. Few people are prepared to listen to a 10-hour speech or read a book of 1,000 pages. Either the subject must be worth the trouble or the hearer must feel some obligation to the subject or orator. Normally, when a custom or an obligation exists, it binds not the hearer but the orator, setting limits to the space or time allotted to the presentation of a thesis. Second, it is considered impolite for an orator to draw out a speech beyond the normally allotted time. Third, by the mere fact that he occupies the platform, an orator prevents other people from expressing their point of view. Consequently, in almost all circumstances in which argumentation can be developed, there are limits that are not to be overstepped.

It thus becomes necessary to make a choice between the available arguments, taking into account the following considerations: first, arguments do not have equal strength

nor do they act in the same manner on an audience. They must be considered relevant for the thesis the speaker upholds and must provide valuable support for it. It is essential that they do not—instead of reinforcing adhesion—call the thesis into question again by raising doubts that would not have occurred to the audience had they not been mentioned. Thus, proofs of the existence of God have shaken believers who would never have thought of questioning their faith had such proofs not been submitted to them. Second, there is constant interaction between the orator and his discourse; thus, the speaker's prestige intensifies the effect of his discourse, but, inversely, if his arguments are weak, the audience's opinion of his intelligence, competence, or sincerity is influenced. Therefore, it is best to avoid using weak arguments; they may induce the belief that the speaker has no better arguments to support his thesis. Third, certain arguments, especially in the case of a mixed audience whose beliefs and aspirations are greatly varied, may be persuasive for only one part of an audience. Therefore, arguments should be chosen that will not be opposed to the beliefs and aspirations of some part of the audience. Thus, by stressing the revolutionary effect of a particular measure, for example, one stiffens the opposition to that measure on the part of those who wish to prevent the revolution, but one draws to the measure the favour of those who wait for the revolution to break out. For this reason arguments that have value for all men are superior to those that have more limited appeal; they are capable of convincing all the members of what could be called the universal audience, which is composed of all normally reasonable and competent men. An argumentation that aims at convincing a universal audience is considered philosophically superior to one that aims only at persuading a particular audience without bothering about the effect it might have on another audience in some other context or circumstances.

The need for order in argumentation

Further, for a discourse to be persuasive, the arguments presented must be organized in a particular order. If they are not, they lose their effectiveness, because an argument is neither strong nor weak in an absolute sense and for every audience but only in relation to a particular audience that is prepared to accept it or not. In the first place, the orator must have a certain amount of prestige, and the problem in question must raise some interest. Should the orator be a small child, a man of ill-repute, or one supposed to be hostile to the audience or should the question be devoid of interest for the audience, there is little chance that the orator will be allowed to speak or that he will be listened to. Thus, an orator is normally introduced by someone who has the public ear, and the orator then uses the exordium, or beginning portion of his discourse, not to speak about his subject but to gain the audience's sympathy.

Effective arguments can modify the opinions or the dispositions of an audience. An argument that is weak because it is ill-adapted to the audience can become strong and effective when the audience has been modified by a previous argument. Similarly, an argument that is ineffective because it is not understood can become relevant once the audience is better informed. Research into the effectiveness of discourse can determine the order in which arguments should be presented. The best order, however, will often be whatever is expected, whether it be a chronological order, a conventional order, or the order followed by an opponent whose argumentation has to be refuted point by point.

In all these considerations—concerning the techniques of presentation and argumentation and the arrangement of a discourse—form is subordinated to content, to the action on the mind, to the effort to persuade and to convince. Consequently, the new rhetoric is not part of literature; it is concerned with the effective use of informal reasoning in all fields.

It has been seen that common principles and notions and common *loci* play a part in all nonspecialized discourses. When the matter that is debated belongs to a specialized field, the discussion will normally be limited to the initiated—*i.e.,* those who, because of their more or less extensive training, have become familiar with the theses and methods that are currently accepted and regarded as valid in the field in question. In such instances, the basis of the argumentation will not be limited to common *loci* but to specific *loci.* The introduction in some field of a new thesis or new methods is always accompanied by criticism of the theses or methods that are being replaced; thus, criticism must be convincing to the specialists if the new thesis or method is to be accepted. Similarly, the rejection of a precedent in law has to be justified by argumentation giving sufficient reasons for not applying the precedent to the case in question.

SIGNIFICANCE OF THE NEW RHETORIC

The new rhetoric introduces a fundamental change in the philosophical outlook. Insofar as it aims at directing and guiding human action in all of the fields in which value judgments occur, philosophy is no longer conceived as the search for self-evident, necessary, universally and eternally valid principles but, rather, as the structuring of common principles, values, and *loci,* accepted by what the philosopher sees as the universal audience. The way the philosopher sees this universal audience, which is the incarnation of his idea of reason, depends on his situation in his cultural environment. The facts a philosopher recognizes, the values he accepts, and the problems he attends to are not self-evident; they cannot be determined *a priori.* The dialectical interaction between an orator and his audience is imposed also on the philosopher who wishes to influence his audience. Therefore, each philosophy reflects its own time and the social and cultural conditions in which it is developed. This is the fundamental truth in the thought of G.W.F. Hegel, a German Idealist: the history of philosophy is not regarded as an abstract and timeless dialectic that proceeds in a predetermined direction but as an argumentation that aims at universality at a concrete moment in history.

To the extent that the new rhetoric views all informal discourse and all philosophical discourse from the viewpoint of its action on the minds of the hearers, it integrates into the analysis of thought valuable elements from both Pragmatism and Existentialism. In stressing the effects of discourse it allows Analytical philosophy to be given the dynamic dimension that some scholars believe it has heretofore lacked. The new rhetoric can thus contribute to the development of a theory of knowledge and to a better understanding of the history of philosophy. (C.Pe.)

BIBLIOGRAPHY. The following works may be regarded as fundamental to the points made in the preceding article: EDWIN BLACK, *Rhetorical Criticism* (1965); WAYNE C. BOOTH, *The Rhetoric of Fiction* (1960); WILLIAM J. BRANDT, *The Rhetoric of Argumentation* (1970); KENNETH BURKE, *The Philosophy of Literary Form* (1941), *A Grammar of Motives* (1945), and *A Rhetoric of Motives* (1950); CHAIM PERELMAN and LUCIE OLBRECHTS-TYTECA, *La Nouvelle Rhetorique: traité de l'argumentation,* 2 vol. (1958; Eng. trans., *The New Rhetoric: A Treatise on Argumentation,* 1969); JOHN CROWE RANSOM, *The New Criticism* (1941); and STEPHEN E. TOULMIN, *The Uses of Argument* (1958). See also CHAIM PERELMAN, "The New Rhetoric: A Theory of Practical Reasoning," in *The Great Ideas Today* (1970).

In addition, the following is helpful in understanding the modern critique of rhetorical traditions: LLOYD F. BITZER and EDWIN BLACK (eds.), *The Prospect of Rhetoric: Report of the National Developmental Project* (1971), RAYMOND F. HOWES (ed.), *Historical Studies of Rhetoric and Rhetoricians* (1961); and R.S. CRANE (ed.), *Critics and Criticism, Ancient and Modern* (1952), are particularly useful in understanding respectively the critics and rhetoricians of Cornell and Chicago, the universities at which modern rhetoric received especially strong impetus. Other works useful in a study of the history of rhetoric include WILBUR SAMUEL HOWELL, *Logic and Rhetoric in England, 1500–1700* (1956); GEORGE KENNEDY, *The Art of Persuasion in Greece* (1963); and WALTER J. ONG, *Ramus: Method, and the Decay of Dialogue* (1958). In addition to Ransom's book, I.A. RICHARDS, *The Philosophy of Rhetoric* (1936), helped illuminate the early stages of the modern relationship between rhetoric and literary criticism. A book-length treatment of non-Western rhetoric is ROBERT T. OLIVER, *Communication and Culture in Ancient India and China* (1971).

Rio de Janeiro

Rio de Janeiro, a major port city of Brazil and the capital of Rio de Janeiro state, is located on the Atlantic Ocean, in the southeastern part of the tropical zone in South America. The name was given to the city's original site by Portuguese navigators who arrived at the port on Jan. 1, 1502, and mistook the entrance of the bay for the mouth of a river (*rio* is the Portuguese word for "river" and *janeiro*, the word for "January"). When the foundations of the future town were laid in 1565, it was named Cidade de São Sebastião do Rio de Janeiro ("City of Saint Sebastian of Rio de Janeiro"), for both St. Sebastian and Dom Sebastian, king of Portugal. It is officially and commonly called Rio de Janeiro but is often referred to, in a shortened form, as Rio.

Rio was the capital of Brazil from 1822 until 1960, when the national capital was moved to Brasília and the territory belonging to the former Federal District was converted into Guanabara state, which formed an enclave in Rio de Janeiro state. In March 1975 the two states were fused as the state of Rio de Janeiro; the former Guanabara state, including the city of Rio de Janeiro, became one of the 14 municipalities of the Metropolitan Region of Rio de Janeiro, or Greater Rio. The city of Rio de Janeiro then became the capital of the reorganized state of Rio de Janeiro.

This article is divided into the following sections:

Physical and human geography 765
Character of the city 765
The landscape 765
The city site
Climate
The city layout
The people 767
The economy 767
Industry
Management, finance, and trade
Transportation
Administration and social conditions 767
Government
Services
Education
Cultural life 768
History 768
The colonial period 768
The city after independence 768
The republican period 769
Bibliography 769

Physical and human geography

CHARACTER OF THE CITY

Rio de Janeiro is well known for the beauty of its beaches and of its peaks, ridges, and hills—all partly covered by tropical forests. The city is a centre of leisure for national and foreign tourists, and people wearing bathing suits can be seen walking in the streets and along the beaches or traveling on the city's buses. Perhaps at no time is the city's festive reputation better displayed than during the annual pre-Lenten Carnival, which enlivens the city night and day with music, singing, parties, balls, and street parades of brilliantly-costumed dancers performing to samba rhythms. Rio is also an important economic centre, however, with activities ranging from industry and national and international trade to administration, banking, education, culture, and research.

The city's economic and social prominence grew in the 18th century after it became the main trade centre for the gold- and diamond-mining areas of Minas Gerais. Later, its status as a national capital and as the royal residence of the Portuguese monarch influenced Rio's continued growth and helped it acquire a cosmopolitan atmosphere and a national character, free of regional conflict. After the city reverted to being a state capital in the mid-20th century, however, a new regional consciousness began to develop.

THE LANDSCAPE

The city site. Rio de Janeiro lies on a strip of Brazil's Atlantic coast, close to the Tropic of Capricorn, where the shoreline is oriented east–west, the city facing south. It was founded on an inlet of this stretch of the coast, Guanabara Bay (Baía de Guanabara), the entrance to which is marked by a point of land called Sugar Loaf (Pão de Açúcar), a "calling card" of the city.

The core, or Centre, of Rio lies on the plains of the western shore of Guanabara Bay. The greater portion of the city—commonly referred to as the North Zone—extends to the northwest on plains composed of marine and continental sediments and on hills and several rocky mountains. The South Zone of the city, reaching the beaches fringing the open sea, is cut off from the Centre and from the North Zone by coastal mountains. These mountains and hills are offshoots of the Serra do Mar, an ancient gneiss-granite mountain chain that forms the slopes of the Brazilian Highlands.

Climate. Although the region's climate is generally tropical, hot, and humid, the climate of Greater Rio is strongly affected by its topography, its proximity to the ocean, and the shape of the southern cone of South America. Along the coast, the breeze, blowing alternately onshore and offshore, modifies the temperature. Because of its geographic situation, the city is reached often, especially during autumn and winter, by cold fronts advancing from the Antarctic, which cause frequent weather changes. But it is mostly in summer that strong showers may provoke catastrophic floods and landslides. The mountainous areas register greater rainfall since they constitute a barrier to the humid wind that comes from the Atlantic. The highest rainfall rate is found in the urban district of Jardim Botânico (more than 63 inches [1,600 millimetres]), where nearby coastal mountains trap humid winds from the Atlantic.

Coastal weather effects

The temperature varies according to altitude, distance from the coast, and type of vegetation. Winter (from June 21 to September 23) is particularly pleasant, both because of its mild temperatures and because it is, in general, less rainy than the summer (December 21 to March 21), which is hotter as well. The annual average temperature at Rio is about 73° F (23° C).

The city layout. The core of the city of Rio de Janeiro (452 square miles [1,171 square kilometres]) and of its large metropolitan area are the sectors called the Centre and the South Zone, respectively.

The Centre corresponds approximately to the old city and is referred to as Cidade (Portuguese: "City," "Downtown"). There are few colonial buildings or monuments because of a series of remodeling and modernizing efforts. Included in these changes were the demolition of old buildings and their substitution with larger and higher constructions; the destruction of hills and the filling of lagoons, swamps, and stretches of the sea; the enlarging of streets and avenues for automobile traffic; and the construction of new infrastructure, such as the port, rebuilt in 1907.

The Centre also contains a number of public buildings with styles that reflect these historical remodeling phases. The Municipal Theatre, still the main national theatre, was built at the beginning of the 20th century and is almost a replica of the Paris Opera House. The Ministry of Education (1936), conceived by Le Corbusier and Brazilian architects, represents the modernismo movement in

Architectural diversity of the Centre



Rio de Janeiro and Sugar Loaf mountain on Guanabara Bay.
Four By Five Inc.

Brazilian architecture of the 1930s, while the headquarters of the Bank of Brazil is an example of a modern skyscraper. Two- or three-storied houses, built at the turn of the century and resembling those of some areas of Lisbon, compete for space with historical monuments, 8- to 12-storied buildings constructed before the 1940s, 20- to 30-storied buildings of the post-World War II era, and 40- to 60-storied skyscrapers of the 1970s. The Avenida Rio Branco is the main street of the Centre, which is also noted as a shopping and entertainment district.

The South Zone

The South Zone grew primarily as a residential area for Rio's wealthy population, which moved from the Centre and North Zone, attracted by the amenities of the shore and the social status that became associated with the area. Settlement along the sea and in valleys of the coastal mountains was made possible, from the 20th century's first decades, by the construction of tunnels and the establishment of a streetcar system. In the shore areas, generally the first street parallel to the shore avenue became the main commercial street. In Copacabana, for example, Avenida Atlántica lies along the shore, and the main commercial artery, Avenida N.S. Copacabana, is the first parallel street inland. Each neighbourhood was named for a bay—such as Flamengo, Botafogo, Copacabana, Ipanema-Leblon—or a valley—such as Laranjeiras or Gávea.

The South Zone was once dominated by individual houses but today is mainly an area of apartment buildings. Despite high rents, the middle class has increasingly moved to the apartments of the South Zone, while the very poor occupy *favelas* (slums) on the slopes of the mountains.

West of the historical South Zone, along the sandy coast, is another urban ring, the development of which began in the 1970s and was based on the growth of automotive transport. Streetcars had disappeared from Rio de Janeiro and all other Brazilian cities by the 1960s and buses had become the main form of public transportation. The opening of more tunnels and the construction of highways sustained urban expansion in the newer southern areas of São Conrado and Barra da Tijuca.

Barra da Tijuca is a planned, upper-class community that looks like a mixture of a British New Town and an affluent U.S. residential suburb. There are areas where only individual houses may be built, apartment buildings are scattered, and commercial and service areas are segregated.

The North Zone

Railways and streetcar lines influenced the North Zone's tentacular form, which prevailed until the mid-20th century. As bus service became the main urban mass transportation, the empty areas between the tentacles were settled, and the North Zone became a large mass of streets, houses, and other buildings. Commercial and service activities were established along the streets or avenues served by streetcar lines and on squares where traffic was concentrated. Neighbourhoods such as Tijuca, Méier, and Madureira emerged as local centres for large portions of the North Zone. In the late 1970s the subway was introduced, linking Tijuca in the North Zone to Botafogo in the South Zone, passing through the Centre. A branch of this line continues to the northern suburban area.

Areas of the North Zone are socially differentiated by the average level of income of the inhabitants, which is reflected in the urban infrastructure and the commercial and service activities that are available. Tijuca and Grajaú are wealthy neighbourhoods, while Engenho Nôvo and Lins de Vasconcelos are poor. Moving outward from the Centre, the first ring, corresponding to older settlement, is a deteriorated area. A second ring contains varyingly affluent neighbourhoods. Outwardly from there, in general, as distance from the Centre increases, poverty also increases. As in the South Zone, but more so, *favelas* have been established on the slopes of mountains and hills, but there are also many such areas in the swampy lands along Guanabara Bay and the river plains.

The suburbs

The suburban zone inside the municipality of Rio de Janeiro extends 12 miles (20 kilometres) north and more than 30 miles west from the Centre. Suburbs also lie in neighbouring municipalities. These areas have experienced the most rapid growth of the metropolitan region since the 1950s. This growth has been due to natural reproduction and migrations from the interior of the state, other states of Brazil, and other areas of the metropolitan region.

With the exception of the South Zone, most of the suburbs are poor; the streets are largely unpaved and most of

the areas are without a sewer system. Individual houses dominate, although the number of apartment buildings is growing. Government housing programs of the 1960s attempted to relocate inhabitants of the *favelas* of the Centre and North Zone to the suburbs, but the population resisted such removal from proximity to their places of work. More recent programs concentrate on rebuilding the *favelas* proper. For example, the huts of the Favela da Maré, close to the Federal University of Rio de Janeiro on the shoreline of Guanabara Bay, were replaced with low-income three-story apartment buildings.

The Baixada

Among the suburbs that extend into the municipalities of Greater Rio are Nova Iguaçú, Nilopolis, São João do Meriti, Duque de Caxias, and Paracambi to the north. These five towns, known as the Baixada (Portuguese: "Lowland"), were once small rural centres that grew tremendously after being linked by rail to Rio proper. Along the road to São Paulo, in Nova Iguaçú, and to Belo Horizonte, in Duque de Caxias, are large industrial plants, such as the Petrobras oil refinery in the latter.

Farther to the north of Rio proper is the satellite city of Petrópolis, once the summer residence of the Brazilian royal family and former capital of Rio de Janeiro state (1894–1903). Located in the highlands at an altitude of 2,667 feet (813 metres), it is a summer tourist resort as well as an industrial centre.

To the west are the smaller localities of Itaguaí and Mangaratiba, where truck farming and tourism prevail. A branch line of the railway that links Rio to Minas Gerais passes through the two towns to the port that was built at Mangaratiba especially for iron-ore exports.

On the east shore of Guanabara Bay, a major urban agglomeration includes Niterói, a former capital of the state of Rio de Janeiro; São Gonçalo; and the suburbs of Itaboraí, Maricá, and others. There also is a large, poor, suburban settlement. Voluminous commuting to Rio de Janeiro is made via the Rio-Niterói bridge (built in the 1970s) and by ferries, motorboats, and hydrofoils. Motorboat service also links Rio to the resort island of Paquetá, which lies near the middle of the bay. Industries in Niterói-São Gonçalo include shipyards and textile, food-processing, and metallurgy plants.

THE PEOPLE

Historically, Rio's population grew primarily as a result of internal migration, which in some years accounted for two-thirds of the city's increase. From the 1930s, governmental economic policies included restrictions on foreign immigration and incentives for internal migration to Brazil's large urban areas. The proportion of foreigners in the former Federal District decreased from 30 to 7 percent between 1890 and 1960. By the 1960s almost half of the city's population were Brazilian migrants, most of them born in the states of Rio de Janeiro, Minas Gerais, and Espírito Santo. Among the largest groups of foreign-born in Greater Rio are the Portuguese, Italians, and Spaniards.

With the transfer of the national capital to Brasília in 1960, the rhythm of population growth in Rio declined. Most national migrants were directed to other municipalities of the metropolitan region, leaving Rio to rely more upon the birth rate within its own boundaries for further growth; persons born in Rio are called cariocas.

Rio's inhabitants are primarily Roman Catholic, although many simultaneously follow the religious folk cult called Umbanda. The population is composed of whites, blacks, and people of mixed race, the whites largely predominant in the wealthy neighbourhoods of Flamengo, Copacabana, Ipanema-Leblon, Jardim Botânico, and Tijuca-Grajaú.

THE ECONOMY

Industry. Greater Rio is the second most important industrial area of Brazil after São Paulo. Large shipyards and an electronics-computer sector have been added to the older industries of metallurgy, engineering, wearing apparel and footwear, textiles, nonmetallic mineral products, food and beverages, chemicals, pharmaceuticals, and printing and publishing.

To attract industry, the state government has designated six areas as industrial districts (Fazenda Botafogo, Palmares, Paciência, Santa Cruz, Campo Grande, and Jacarepaguá) where infrastructure is provided and the sale of land is made under special conditions. The discovery and exploitation of oil and natural gas fields off the coast of the state of Rio de Janeiro also has opened new possibilities for the development of industrial activities in Rio's metropolitan region.

Rio's attraction for tourists and seasonal residents is an important stimulus for trade. The establishment of second residences in Rio by many wealthy people from other Brazilian cities and from abroad has been a factor in the city's active building industry, which is a significant source of employment for large numbers of unskilled labourers.

Management, finance, and trade. Because it was once the national capital, Rio de Janeiro was chosen as the site for the headquarters of many large private, national, multinational, and state corporations, even when their factories were located in other cities or states. Despite the transfer of the capital to Brasília, many of these headquarters remained within the Rio metropolitan area, as did, for instance, those of Petrobrás, the state oil company; Rio Doce Valley Company, a state mining enterprise; the National Economic and Social Development Bank, a federal investment bank; and Esso and Shell, multinational oil corporations.

Rio is an important financial centre, second only to São Paulo in volume of business in the stock market or in banking. Its securities market is also of major importance. The port of Rio has a large market area, exporting automobiles produced in Belo Horizonte, for example, and is among the nation's leading ports by tons moved.

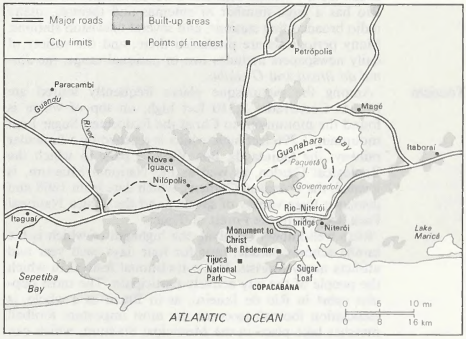
In Greater Rio, which has one of the highest per capita incomes in Brazil, the retail trade is substantial. Many of the most important retail stores are located in the Centre, but others are scattered throughout the commercial areas of the other districts, where shopping centres, supermarkets, and other retail businesses handle a large volume of consumer trade.

Transportation. Rio de Janeiro is still the primary centre for air services in Brazil, with flights to major cities of North America and Europe. There are two modern airports—Galeão for domestic and international services and Santos Dumont for domestic lines only.

Most surface transport with other places is negotiated by bus, truck, and automobile; there are railway links to São Paulo and Belo Horizonte, but their daily service is somewhat limited. Privately owned bus services are the main means of public transportation inside the urban agglomeration. The city began building its subway in 1972, and by 1979 the first station was opened, initiating the first step in the development of an underground system marked to alleviate Rio's serious traffic congestion.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. The city and municipality of Rio de Janeiro are governed by a *prefeito* (Portuguese: "mayor")



Rio de Janeiro and surrounding area.

with the assistance of secretaries who head administrative departments. Since 1984 the *prefeito* has been elected to a four-year term. Legislative power is held by the members of the Municipal Chamber, who are elected proportionally from Rio de Janeiro's 24 administrative regions.

Services. As frequently happens in rapidly developing metropolitan areas, Rio faces serious difficulties in providing urban services and facilities, especially in the growing suburban areas. The water supply and sewer system belong to a state company, and a state company also supplies domestic gas. The water supply system was modernized during the 1960s, but the extension of the water supply and the sewage system to new urbanized areas and to the *favelas* has been a continuing problem.

Telephone service is operated by two public enterprises, one for the central area and another for the suburbs. The demand for home telephone equipment is greater than the system's capacity, and, consequently, the trunk lines are frequently overloaded. Electricity is distributed by a former Canadian-owned private enterprise bought by the federal government during the 1970s. Rio is connected to an electrical system that extends throughout south-central Brazil.

The city has a healthful climate. Although pollution is a growing concern, there are no serious health problems, except within the *favelas,* where diseases related to lack of sanitation, poor diet, and inadequate health facilities prevail. The Rio de Janeiro area as a whole has one of the nation's better ratios of population to hospital beds and doctors.

Education. The literate population of the city makes up approximately 90 percent of those 10 years or older, a proportion that is higher than the national average. The city's two most important universities, the Federal University of Rio de Janeiro and the Pontifical Catholic University of Rio de Janeiro, offer graduate courses. The state also administers the State University of Rio de Janeiro and a large network of secondary schools. A number of governmental national research centres in Rio de Janeiro conduct studies in such fields as physics, atomic energy, geology, geography, and statistics.

CULTURAL LIFE

As the country's cultural capital, Rio de Janeiro has many prestigious artistic, literary, and scientific institutions. These include the Brazilian Academy of Letters, the Brazilian Academy of Sciences, and numerous museums. Among the museums are the National Museum of Fine Arts, founded in 1818; the National Museum, rich in anthropological objects and located in the former Imperial Palace of the Quinta da Boa Vista; the National Historical Museum; the Museum of Modern Art; and the Indian Museum.

Museums

The most important of the city's many libraries is the National Library; it was founded in 1810 with the remains of the Royal Library of Ajuda, which were brought to Brazil from Portugal after the 1755 earthquake in Lisbon. Rio has a large number of cinemas and theatres, many radio broadcasting stations, and several television stations. Many periodicals are published there, and the variety of daily newspapers includes two of national range, the *Jornal do Brasil* and *O Globo.*

Among the picturesque places frequently toured are Mount Corcovado, 2,310 feet high, on top of which is found the monument to Christ the Redeemer; Sugar Loaf mountain, 1,296 feet high, which is reached by a funicular railway; the Quinta da Boa Vista, a park in which the Zoological Garden, as well as the National Museum, is located; the Botanical Gardens, which date from 1808 and display a huge variety of species; and the Tijuca National Park, located in the Forest of Tijuca.

Tourism

Rio's well-known Carnival, the highlight of which is the samba schools parade, lasts for four days each year and attracts many tourists. It is a traditional festival in which the people of the city actively participate. The most popular sport in Rio de Janeiro, as in Brazil as a whole, is association football (soccer); the most important football matches take place in the Municipal Stadium, which can seat 200,000 spectators.

History

THE COLONIAL PERIOD

Several years after the Portuguese first explored Brazil, French traders in search of *pau-brasil* (a type of brazilwood) explored the rich area extending from the Cape Frio coast to the beaches and islands of Guanabara Bay—the economic and, above all, strategic importance of which was already well-known. On one of these islands, the French founded a colony that was called La France Antarctique (Antarctic France).

Early settlement

The Portuguese wanted to expel the French from Brazil, and the task was given to Estácio de Sá, a nephew of Governor Mem de Sá of Brazil, who in 1565 occupied the plain between Dog Face Hill (Morro Cara de Cão) and the Sugar Loaf and Urca mounts, thus laying the foundations of the future town of Rio de Janeiro. After two years (1565–67) of bloody battles, in which Estácio de Sá was killed and the French expelled, Mem de Sá chose a new site for the town, farther inland on the coast of the bay, at the top of the Hill of Rest (Morro do Descanso), or St. Januarius Hill (São Januário), later called the Castle Hill (Morro do Castelo). In 1568 the settlement was laid out in the form of a medieval citadel, protected by a bulwark and cannons.

The surrounding fertile land, allotted to Portuguese settlers by the Portuguese king in enormous plots called *sesmarias,* was planted with sugarcane, which was to provide the colony with its main source of income. In 1660 the community became the seat of the government of the southern captaincies (Portuguese administrative units) of Brazil. In the second half of the 17th century, the captaincy population grew to 8,000 inhabitants, two-thirds of whom were probably Indian and black slaves.

At the beginning of the 18th century, Brazil began to engage in gold and diamond mining, which brought about remarkable changes in the colony's economy and stimulated a great migration from Europe, thereby increasing the white population. The former village became a town of 24,000 in 1749. When the colonial capital was transferred from Bahia to Rio de Janeiro in 1763, the town expanded farther, far beyond its walls. The remains of the monumental Roman-style aqueduct Arcos ("Arches") built at this time still stand in the city.

At the end of the 18th century, the town's economy, as well as that of the colony as a whole, was in a crisis because of the decline of the mines and competition from Central America for the world sugar market. In 1796 the value of exports from Rio's port was less than half of what it had been in 1760.

Coffee production and the resettlement of the Portuguese royal family in Brazil in 1808 again brought prosperity to the colony. By 1815, when Brazil became a kingdom, Rio de Janeiro was large enough to accommodate a foreign population. At about this time the city's initial features were being transformed; from 1808 to 1818, 600 houses and 100 country houses were built, and many older buildings were restored. Many streets were lighted and paved, more land was reclaimed, new roads opened, and new public fountains installed. Among new institutions established were the Royal Press, the Royal Library, the Theatre of Saint John, the Academy of Fine Arts, the Botanical Gardens, and the Bank of Brazil. When King John VI returned to Portugal in 1821, Rio had almost 113,000 inhabitants and 13,500 buildings, and the town had extended both northward and southward. A year later Brazil was independent.

Colonial improvements

THE CITY AFTER INDEPENDENCE

Expansion of coffee plantations in the state of Rio de Janeiro gave a new impulse to the city's development. Nobles and bourgeois moved their residences north to the São Cristóvão district. Merchants and English bankers chose to live around the Outeiro da Glória and Praia do Flamengo areas in the south, or they established their residences in the nearby Botafogo and Laranjeiras districts. The French, on the other hand, lived in country houses scattered in the Tijuca area farther westward.

In this era, as Brazil expanded its world export trade in

such products as coffee, cotton, sugar, and rubber, the city changed its appearance, and the traces of its colonial past were effaced. In 1829 oxcart traffic was banned from the Rua do Ouvidor, then the city's most elegant street. In 1838 the first public transportation—horse-drawn buses—began to run to the districts of São Cristóvão, Engenho Velho, and Botafogo. In 1868 the first tramcars, also drawn by animals, were introduced. A steamboat service to Niterói began to operate in 1835. The first railroad was built in 1852 to Petrópolis, and a line reached Queimados in the Nova Iguaçú area in 1858. In 1854 gas replaced oil for street-lighting, and wireless telegraphy was inaugurated. Sewerage was installed in 1864, and telephone service began in 1877.

THE REPUBLICAN PERIOD

When Rio de Janeiro, which had formerly been the capital of the empire, became capital of the Republic of Brazil in 1889, it was already a considerable community. At the time of the 1890 census, it had more than 520,000 inhabitants on 61 square miles, ranking it as the largest city in Brazil and one of the larger cities in the world. The 1891 constitution made it the Federal District.

During the federal administration of President Francisco de Paula Rodrigues Alves, from 1902 to 1906, Rio de Janeiro was further transformed. A team of administrators and technicians drained swamps, cleared slums, paved and widened streets, and markedly improved health conditions, notably by reducing cases of yellow fever and smallpox. The central avenue (called Avenida Rio Branco from 1912), still the most important of the Centre, was opened during this period; Avenida Beira-Mar, running parallel to part of the south shore, was built on reclaimed land; and several other important avenues were opened.

The population of the Federal District exceeded 1,000,000 by 1920 and increased to 1,750,000 by 1940. During this period the number of industrial establishments in the Federal District nearly trebled. Castle Hill was demolished, land reclamation increased the area in the Centre, and the first skyscrapers appeared. Streetcar lines, now moved by electrical power, multiplied. While settlement spread on the east coast, some areas to the north lost status—such as São Cristóvão, which became an industrial and lower-class residential neighbourhood. (A.P.G./P.P.G.)

In the mid-20th century metropolitan Rio de Janeiro grew beyond its formerly narrow confines because of immigration from other parts of Brazil, as well as a high rate of natural increase. The suburban population expanded from some 450,000 in 1940 to 1,550,000 in 1960, at a time when the city proper exceeded 3,300,000. Following the transfer of the national capital to Brasília in 1960, Rio's rate of population growth slowed somewhat; however, the suburban population continued to mushroom. By 1980 there were more than 5,000,000 residents in the city and 3,600,000 in the suburbs. Although subsequent growth continued at a lesser rate, by 2000 the metropolitan population approached 11,000,000 people, of which some 5,850,000 resided in the city proper.

During the latter part of the 20th century population growth, real estate speculation, and an amplified dependence on automobiles precipitated a series of changes in the city. The highway system was expanded, with improved access to numerous suburban areas; still higher skyscrapers arose in the Centre; and larger residential towers increasingly replaced small apartment buildings and houses. Landowners, the industrial sector, and the financial sector favoured the changes; however, higher rents forced many poorer residents to move from the central city to the periphery or to the rapidly expanding *favelas* on the steeper hillsides. Meanwhile, many affluent families settled in the southwest, notably in Ipanema and Leblon and at sites farther down the coast.

Rio has maintained its status as the second largest manufacturing centre in Brazil, after São Paulo. In addition, Rio's service sector remains one of the strongest in the nation, providing for increasing numbers of jobs in such areas as finance, communications, tourism and entertainment, public and private education, and computer engineering. Rio remains a major centre for cultural exhibitions and international conferences, and many of the more affluent and influential Brazilians keep homes there. Despite desperate conditions in the city's *favelas,* as well as high rates of violent crime throughout the city, grave levels of air and water pollution, and myriad other problems, Rio de Janeiro is still widely recognized as one of the world's more beautiful urban centres.

BIBLIOGRAPHY. Comprehensive guidebooks include MICK DAY and BEN BOX, *Rio de Janeiro* (2001); LOU BYRNES and CHARLES RAWLINGS-WAY, *Rio de Janeiro* (1999), in the Lonely Planet series; and *Fodor's Pocket Rio de Janeiro: The Best of the City* (2000). The physical setting is studied in ALBERTO RIBEIRO LAMEGO, *O Homem e a Guanabara,* 2nd ed. (1964). Urban planning is discussed in NORMA EVENSON, *Two Brazilian Capitals: Architecture and Urbanism in Rio de Janeiro and Brasília* (1973). PEDRO PINCHAS GEIGER, *Evolução da rêde urbana brasileira* (1963); and JÚLIA ADÃO BERNARDES (ed.), *Rio de Janeiro, painel de um espaço em crise* (1986), are overviews of the city's geography and development.

Slums and the urban poor are examined in JANICE E. PERLMAN, *The Myth of Marginality: Urban Poverty and Politics in Rio de Janeiro* (1976, reprinted 1979); JULIO CESAR PINO, *Family and Favela: The Reproduction of Poverty in Rio de Janeiro* (1997); and ALMA GUILLERMOPRIETO, *Samba* (1991), which describes a year of Carnival preparations in one of Rio's *favelas.*

Historical accounts of the city include VIVALDO COARACY, *O Rio de Janeiro no século 17,* 2nd ed. rev. and expanded (1965); GASTÃO CRULS, *Aparência do Rio de Janeiro: Noticia histórica e descritiva da cidade,* 3rd ed. rev., 2 vol. (1965); MARY C. KARASCH, *Slave Life in Rio de Janeiro, 1801–1850* (1987); and JEFFREY D. NEEDELL, *A Tropical Belle Époque: Elite Culture & Society in Turn-of-the-Century Rio de Janeiro* (1988). Ethnic relations and the myth of Brazil's "racial democracy" are traced in MICHAEL GEORGE HANCHARD, *Orpheus and Power: The Movimento Negro of Rio de Janeiro and São Paulo, Brazil, 1945–1988* (1994, reissued 1999). (Ed.)

Sacred Rites and Ceremonies

The sacred (or holy) is the power, being, or realm understood by religious persons to be at the core of existence and to have a transformative effect on their lives and destinies. Other terms, such as divine, transcendent, ultimate being (or reality), mystery, and perfection (or purity) have been used for this domain. "Sacred" is also an important technical term in the scholarly study and interpretation of religions.

Broadly defined, worship is man's response to the sacred. Characteristic modes of response (other than spontaneous, ecstatic behaviour) include primarily cultic and private acts of ritual, the performance of which is prescribed by tradition or sacerdotal decree.

This article will treat various ritualistic, behavioral responses to the sacred. For a discussion of dogmatic interpretations of the Divine as a being or force, see RELIGIOUS AND SPIRITUAL BELIEF and DOCTRINES AND DOGMAS. This article is divided into the following sections:

The sacred or holy 770
- The emergence of the concept of the sacred
- Basic characteristics of the sacred
- Critical problems
- The sacred today

Worship 773
- Nature and significance
- Functions of worship
- Types of worship
- Variations or distinctions within the act of worship
- Times and places of worship
- Focuses of worship
- Conclusion

The concept and forms of ritual 778
- Nature and significance
- Functions of ritual
- Types of ritual
- Conclusion

Prayer 781
- Nature and significance
- Origin and development
- Types of prayer
- Forms of prayer in the religions of the world
- Conclusion

Creed and confession 785
- Creeds in the major religions
- Confessions of the Christian faith

Sacrament 788
- Nature and significance
- Types and variations
- Theology and practice of sacraments in Christianity
- Conclusion

Sacrifice 791
- Nature and origins
- Analysis of the rite of sacrifice
- Sacrifice in the religions of the world

Rites of passage 798
- Nature and significance
- Functions
- Classifications of rites
- Rites of passage in the context of the social system
- Psychological aspects of rites of passage
- Rites of passage in the context of the religious system
- Primary rites of passage

Death rites and customs 804
- Relevant concepts and doctrines
- Patterns of myth and symbol
- Death and funerary rites and customs
- Cults and memorials of the dead
- Psychological and sociological aspects of death
- Modern notions of death

Purification rites and customs 810
- Concepts of purity and pollution
- Categories and theories of pollution and impurity
- Types of purification rites
- Examples of purification rites
- Pollution beliefs in modern society

Dietary laws and food customs 815
- Nature and significance
- Laws and customs at different stages of social development
- Rules and customs in world religions

Ceremonial and ritualistic objects 822
- Varieties
- Types of sacred settings for ceremonial and ritualistic objects
- Forms of ceremonial and ritualistic objects according to their functions
- Conclusion

Religious dress and vestments 830
- Types of dress and vestments in Western religions
- Types of dress and vestments in Eastern religions

Feasts and festivals 835
- Nature and significance
- Types and kinds of feasts and festivals
- Conclusion

Bibliography 839

THE SACRED OR HOLY

The emergence of the concept of the sacred. It was during the first quarter of the 20th century that the concept of the sacred (or holy) became dominant in the comparative study of religions. Nathan Söderblom, an eminent Swedish churchman and historian of religions, asserted in 1913 that the central notion of religion was "holiness" and that the distinction between sacred and profane was basic to all "real" religious life. In 1917 Rudolf Otto's *Heilige* (Eng. trans., *The Idea of the Holy,* 1923) appeared and exercised a great influence on the study of religion through its description of religious man's experience of the "numinous" (a mysterious, majestic presence inspiring dread and fascination), which Otto, a German theologian and historian of religions, claimed, could not be derived from anything other than an a priori sacred reality. Other scholars who used the notion of sacred as an important interpretive term during this period included the sociologist Émile Durkheim in France, and the psychologist-philosopher Max Scheler in Germany. For Durkheim, sacredness referred to those things in society that were forbidden or set apart; and since these sacred things were set apart by society, the sacred force, he concluded, was society itself. In contrast to this understanding of the nature of the sacred, Scheler argued that the sacred (or infinite) was not limited to the experience of a finite object. While Scheler did not agree with Otto's claim that the holy is experienced through a radically different kind of awareness, he did agree with Otto that the awareness of the sacred is not simply the result of conditioning social and psychological forces. Though he criticized Friedrich Schleiermacher, an early 19th-century Protestant theologian, for being too subjective in his definition of religion as "the consciousness of being absolutely dependent on God," Otto was indebted to him in working out the idea of the holy. Söderblom recorded his dependence on the scholarship of the history of religions (*Religionswissenschaft*), which had been a growing discipline in European universities for about half a century; Durkheim had access to two decades

Otto's *The Idea of the Holy*

of scholarship on nonliterate peoples, some of which was an account of actual fieldwork. Scheler combined the interests of an empirical scientist with a philosophical effort that followed in the tradition of 19th-century attempts to relate human experiences to the concept of a reality (essence) that underlies human thoughts and activities.

Since the first quarter of the 20th century many historians of religions have accepted the notion of the sacred and of sacred events, places, people, and acts as being central in religious life if not indeed the essential reality in religious life. For example, phenomenologists of religion such as Gerardus van der Leeuw and W. Brede Kristensen have considered the sacred (holy) as central and have organized the material in their systematic works around the (transcendent) object and (human) subject of sacred (cultic) activity, together with a consideration of the forms and symbols of the sacred. Such historians of religions as Friedrich Heiler and Gustav Mensching organized their material according to the nature of the sacred, its forms and structural types. Significant contributions to the analysis and elaboration of the sacred have been made by Roger Caillois, a sociologist, and by Mircea Eliade, an eminent historian of religions.

Basic characteristics of the sacred. *Sacred–profane and other dichotomies.* The term sacred has been used from a wide variety of perspectives and given varying descriptive and evaluative connotations by scholars seeking to interpret the materials provided by anthropology and the history of religions. In these different interpretations, however, common characteristics were recognized in the sacred, as it is understood by participant individuals and groups: it is separated from the common (profane) world; it expresses the ultimate total value and meaning of life; and it is the eternal reality, which is recognized to have been before it was known and to be known in a way different from that through which common things are known.

Sacer and *numen*

The term sacred comes from Latin *sacer* ("set off, restricted"). A person or thing was designated as sacred when it was unique or extraordinary. Closely related to *sacer* is *numen* ("mysterious power, god"). The term numinous is used at present as a description of the sacred to indicate its power, before which man trembles. Various terms from different traditions have been recognized as correlates of *sacer:* Greek *hagios,* Hebrew *qadosh,* Polynesian *tapu,* Arabic *ḥaram;* correlates of *numen* include the Melanesian *mana,* the Sioux *wakanda,* the old German *haminja* (luck), and Sanskrit *Brahman.*

Besides the dichotomy of sacred–profane the sacred includes basic dichotomies of pure–unpure and pollutant–"free." In ancient Rome the word *sacer* could mean that which would pollute someone or something that came into contact with it, as well as that which was restricted for divine use. Similarly, the Polynesian *tapu* ("tabu") designated something as not "free" for common use. It might be someone or something specially blessed because it was full of power, or it might be something accursed, as a corpse. Whatever was tabu had special restrictions around it, for it was full of extraordinary energy that could destroy anyone unprotected with special power himself. In this case the sacred is whatever is uncommon and may include both generating and polluting forces. On the other hand there is the pure–impure dichotomy, in which the sacred is identified with the pure and the profane is identified with the impure. The pure state is that which produces health, vigour, luck, fortune, and long life. The impure state is that characterized by weakness, illness, misfortune, and death. To acquire purity means to enter the sacred realm, which could be done through purification rituals or through the fasting, continence, and meditation of ascetic life. When a person became pure he entered the realm of the divine and left the profane, impure, decaying world. Such a transition was often marked by a ritual act of rebirth.

The elements of fear and fascination

Ambivalence in man's response to the sacred. Because the sacred contains notions both of a positive, creative power and a danger that requires stringent prohibitions, the common human reaction is both fear and fascination. Otto elaborated his understanding of the holy from this basic ambiguity. Only the sacred can fulfill man's deepest needs and hopes; thus, the reverence that man shows to the sacred is composed both of trust and terror. On the one hand, the sacred is the limit of human effort both in the sense of that which meets human frailty and that which prohibits human activity; on the other hand, it is the unlimited possibility that draws mankind beyond the limiting temporal–spacial structures that are constituents of human existence.

Not only is there an ambivalence in the individual's reaction to the numinous quality of the sacred but the restrictions, the tabus, can be expressive of the creative power of the sacred. Caillois has described at length the social mechanism of nonliterate societies, in which the group is divided into two complementary subgroups (moieties), and has interpreted the tabus and the necessary interrelationship of the moieties as expressions of sacredness. Whatever is sacred and restricted for one group is "free" for the other group. In a number of respects—*e.g.,* in supplying certain goods, food, and wives—each group is dependent on the other for elemental needs. Here the sacred is seen to be manifested in the order of the social–physical universe, in which these tribal members live. To disrupt this order, this natural harmony, would be sacrilege, and the culprit would be severely punished. In this understanding of the sacred, a person is, by nature, one of a pair; he is never complete as a single unit. Reality is experienced as one of prescribed relationships, some of these being vertical, hierarchical relationships and others being horizontal, corresponding relationships.

Another significant ambiguity is that the sacred manifests itself in concrete forms that are also profane. The transcendent mystery is recognized in a specific concrete symbol, act, idea, image, person, or community. The unconditioned reality is manifested in conditioned form. Eliade has elucidated this "dialectic of the sacred," in which the sacred may be seen in virtually any sort of form in religious history: a stone, an animal, or the sea. The ambiguity of the sacred taking on profane forms also means that even though every system of sacred thought and action differentiates between those things it regards as sacred or as profane, not all people find the sacred manifested in the same form; and what is profane for some is sacred for others.

Manifestations of the sacred. The sacred appears in myths, sounds, ritual activity, people, and natural objects. Through retelling the myth the divine action that was done "in the beginning" is repeated. The repetition of the sacred action symbolically duplicates the structure and power that established the world originally. Thus, it is important to know and preserve the eternal structure through which man has life, for it is the model and source of power in the present.

Sacred sounds

The recognition of sacred power in the myth is related to the notion that sound itself has creative power—in particular special, sacred sounds. Sometimes these sounds are words, such as the name of god, divine myth, a prayer, or hymn; but sometimes the most sacred sounds are those that do not have a common meaning, for example, the Hindu *oṃ,* the Buddhist *oṃ maṇi padme hūṃ,* or the Jewish and Christian "Hallelujah."

Closely connected with verbal expressions of sacred power are activities done in worship, in sacraments, sacrifices, and festivals. Part of the importance of religious ritual is that in the realm of the sacred all things have their place. In order for human existence to prosper (or even continue) it must correspond as closely as possible to the divine pattern (destiny, or will). Different religious traditions have different theological and philosophical formulations of the meaning of sacraments. In Roman Catholic Christianity, a sacrament is "an outward and visible sign of an inward and invisible grace." In Brahmanic Hinduism a *saṃskāra* (sacrament) is a sacred act that perfects a person and that culminates at the end of a series of *saṃskāras* in a spiritual rebirth, a symbolic "second birth." In both of these cases, the sacred action establishes the relation between the divine and human worlds.

Initiation, sacrifice, and festival

Other sacred activity includes initiation, sacrifice, and festival. Initiation rites among nonliterate societies both expose and establish the world view of the participants. The initiate learns the eternal order of life as proclaimed

in the myth. Life is viewed essentially as the work of supernatural beings, and the initiate in this ritual is taught this secret of life and how to gain access to divine benefits. The initiate learns the tabus and is often given a sacred mark—*e.g.,* circumcision, tattoo, or incisions—to express physically that he is part of the sacred (original) community. In other religions, such as Christianity, Buddhism, and Hinduism, an initiate to a special holy (often monastic) community within the larger religious community is designated by a change in name and wearing apparel, denoting his special relation to the sacred.

In festivals and sacrifices two religious functions are often combined: (1) to provide new power (energy, life) for the world, and (2) to purify the corrupted, defiled existence. Religious festivals are a return to sacred time, that time prior to the structured existence that most people commonly experience (profane time). Sacred calendars provide the opportunity for the profane time to be rejuvenated periodically in the festivals. These occasions symbolically repeat the primordial chaos before the beginning of the world; and just as the world was created "in the beginning," so in the repetition of that time the present world is regenerated (see also *Dimensions of the sacred,* below). The use of masks and the suspension of normal tabus express the unstructured, unconditioned nature of the sacred. Dancing, running, singing, and processions are all techniques for re-creation, for stimulating the original power of life. Ritual activity moves power in two directions: (1) it concentrates it in one place, time, and occasion, and (2) it releases power into the everyday stream of events through its self-abundance—the primal vibration reverberates throughout existence. The new energy dispels the old, depleted, polluted energy; it cleanses the constricted, clogged, hardened channels of life.

One of the most important forms in which man has access to the sacred is in the sacrifice. The central procedure in all sacrifices is the use of a victim or substitute to serve as a mediator between the sacred and profane worlds. The sacrifice (Latin *sacri-ficium,* "making sacred") is a consecration of an offering through which the profane world has access to the sacred without being destroyed by the sacred. Instead, the sacrificial object (victim) is destroyed in serving as a unique, extraordinary channel between these two realms. In sacrificial rites it is important to duplicate the original (divine) act; and because creation is variously conceived in different religious traditions, different forms are preserved: the burning or crushing of the "corn mother," the crushing of the *soma* stalks, the slaughter of the lamb without blemish, the blood spilling of a sacred person, such as the firstborn.

Sacredness is manifested in sacred officials, such as priests and kings; in specially designated sacred places, such as temples and images; and in natural objects, such as rivers, the sun, mountains, or trees. The priest is a special agent in the religious cult, his ritual actions represent the divine action. Similarly, the king or emperor is a special mediator between heaven and earth and has been called by such names as the "son of heaven," or an "arm of god."

Just as certain persons are consecrated, so specific places are designated as the "gate of heaven." Temples and shrines are recognized by devotees as places where special attitudes and restrictions prevail because they are the abode of the sacred. Likewise, certain images of God (and sacred books) are held to be uniquely powerful and true (pure) expressions of divine reality. The image and the temple are, in traditional societies, not simply productions by individual artists and architects; they are reflections of the sacred essence of life, and their measurements and forms are specified through sacred communication from the divine sphere. In this same context, natural objects can be imbued with sacred power. The sun, for example, is the embodiment of the power of life, the source of all human consciousness, the central pivot for the eternal rhythm and order of existence. Or, a river, such as the Nile for the ancient Egyptians and the Ganges for the Hindu, gave witness to the power of life incarnated in geography. Sacred mountains (*e.g.,* Sinai for Jews, Kailāsa for Hindus, Fujiyama for Japanese) were particular loci of divine power, law, and truth.

Sacred time and place

Dimensions of the sacred. The sacred, by definition, pervades all dimensions of life. Within the kind of religious apprehension that is expressed in sacred myth and ritual, however, there is a special focus on time, place (cosmos), and active agents (heroes, ancestors, divinities). When existence is seen in terms of the dichotomy of sacred and profane—which assumes that the sacred is wholly other than, yet necessary for, everyday existence—it is very important to know and to get in contact with the sacred. In periodic festivals men celebrate sacred time; a sacred calendar marks off the intervals of man's life, and these sacred festivals provide the pattern for productive and joyous living.

Seasonal sacred calendars are especially important in predominantly agricultural societies. In the very order of nature, people see that different seasons have their distinct values. These differences are celebrated with spring festivals (when the world is re-created through ritual expressions of generation) and harvest festivals (of thanksgiving and of protecting the life force in seeds for the next spring). Here time is regarded as cyclical, and one's life is marked by those rituals in which one continually returns to the divine source.

Similarly, the myths and rituals mark off the world (cosmos) into places that have special sacred significance. The territory in which one lives is real insofar as it is in contact with the divine reality. Within this territory is life; outside it is chaos, danger, and demons. Throughout most of history the "sacred world" was coextensive with a certain territory, and one could speak literally of Christian lands, the Jewish homeland, the Muslim world, the place of the noble people (*Āryāvarta,* Hindu), or the central kingdom (China). Consecrating one's possession of land with certain rituals was equal to establishing an order with divine sanction. In Vedic ritual, for example, the erection of a fire altar (in which the god Agni—fire—was present) was the establishment of a cosmos on a microcosmic scale. Once a cosmos is established, there are certain places that are especially sacred. Certain rivers, mountains, groves of trees, caves, or human constructions such as temples, shrines, or cities provide the "gate," "ladder," "navel," or "pole" between heaven and earth. This sacred place is that which both allows the sacred power to flow into existence and gives order and stability to life.

Another dimension of the sacred is divine or heroic activity: the decisive action done by creative or protective agents. One's spiritual ancestors need not be biologically defined ancestors; they may not even be human. They are the essential forces on which survival depends and can be embodied in animal skills (longevity, rebirth, magical skills), in the "ways of the ancients," or through a special hero who has provided present existence with material and spiritual benefits. If the notion of sacred manifestation is extended to include the social relationships (especially tabus) in a community, then communal relations can be viewed as a dimension through which the sacred is manifested. Here human values are sacralized by social restraints that prescribe—*e.g.,* with whom one can eat or whom one can marry or kill. The establishment of a community requires forming certain relationships; and these relationships are sacred when they bear the power of ultimate, eternal, cosmic force. For example, the consecration of a king or emperor in traditional agricultural societies was the establishment of a system of allegiance and order for society.

Divine or heroic activity and social relations

By extending the notion of "sacralization" to include human reorganization of experience within the context of any absolute norm, the sacred can be seen in such dimensions of life as history, self-consciousness, aesthetics, and philosophical reflection (conceptualization). Each of these modes of human experience can become the creative force whereby some people have "become real" and gained the most profound understanding of themselves.

Critical problems. Phenomenologists of religion who use the concept "sacred" as a universal term for the basis of religion differ in their estimation of the nature of the sacred manifestation. Otto and van der Leeuw hold (in different formulations) that the sacred is a reality that transcends the apprehension of the sacred in symbols or

rituals. The forms (ideograms) through which the sacred is expressed are secondary and are simply reactions to the "wholly other." Kristensen and Eliade, on the other hand, regard the sacred reality to be available through the particular symbols or ways of apprehending the sacred. Thus, Kristensen places emphasis on how the sacred is apprehended, and Eliade describes different modalities of the sacred, while Otto looks beyond the forms toward a meta-empirical source.

Nonuniversalist views of the sacred

A second problem is the continuing question of whether or not the sacred is a universal category. There are religious expressions from various parts of the world that clearly manifest the kind of structure of religious awareness characterized above. It is especially apropos of some aspects in the religion of nonliterate societies, the ancient Near East, and some popular devotional aspects of Hinduism. There is, however, a serious question regarding the usefulness of this structure in interpreting a large part of Chinese religion, the social relationships (*dharma*) in Hinduism, the effort to achieve superconscious awareness in Hinduism (Yoga), Jainism, Buddhism (Zen), some forms of Taoism, and some contemporary (modern) options of total commitment that, nevertheless, reject the notion of an absolute source and goal essentially different from human existence. If one takes the notion of sacred as something above (beyond, different from) the religious structure dominated by divine or transcendent activity (described above), then this suggests that the notion of sacredness should not be limited to that structure. Thus, some scholars have found it confusing to use the notion of sacred as a universal religious quality, for it has been accepted by many religious people and by scholars of religion as referring to only one (though important) type of religious consciousness.

The 20th-century discussion of the nature and manifestation of the sacred includes other approaches than those of scholars in the comparative study of religions. For example, Sri Aurobindo, a Hindu mystic-philosopher, speaks of the supreme reality as the "Consciousness-Force"; and Nishida Kitaro, a Japanese philosopher, expresses his apprehension of universal reality as that of "absolute Nothingness." Martin Heidegger, a German philosopher, speaks of "the holy" as that dimension of existence through which there is the illumination of the things that are, though it is no absolute Being prior to existence; rather it is a creative act at the point of engaging the Nothing (*Nichts*). In contrast, the Protestant theologian Karl Barth rejects philosophical reflection or mystical insight for apprehending the sacred, and insists that personal acceptance of God's self-revelation in a particular historical form, Jesus Christ, is the place to begin any awareness of what philosophers call "ultimate."

Sociologists who study religion have, since Durkheim, usually identified the sacred with social values that claim a supernatural basis. Nevertheless, the sacred has been identified predominantly as found in the social occasions (festivals) that disrupt the common social order (by Caillois), or as the reinforcing of social activities that secure a given social structure (by Howard Becker). During the 1960s, however, the usual definition of religion as those sacred activities which claimed a transcendent source was questioned by some empirical scholars. For example, Thomas Luckmann, a German-American sociologist, described the sacred in modern society as that "strata of significance to which everyday life is ultimately referred"; and this definition includes such themes as "the autonomous individual" and "the mobility ethos."

The sacred today. The problems of defining and investigating religion mentioned above are already expressive of the shifts in modern consciousness regarding the sacred. Both the physical and social sciences have given modern man a new image of himself and techniques for improving his present life. The acceptance of rational and critical perspectives for judging the claims of religious authorities in Europe since the 18th century, plus the development of historical criticism and a sense of historical relativism, has contributed to the affirmation of man as basically a secular person. The once absolute authorities in the West (the Bible, priest, rabbi) are no longer the prime sources for one's self-identity. To a growing extent the cultures in the East are also experiencing a loss of their traditional authorities. Some attempts have been made to resacralize contemporary cosmology, history, and personal experience by (1) extending the scope of religious concerns to "secular" areas such as politics, economics, personality development, and art; and (2) modifying theological positions, ethical norms, and liturgical forms to incorporate new modes of expression and to experiment with new styles of living.

Assertions of the end of the sacred

An important 20th-century development in religious life has been the easy flow of information between religious communities on different continents. This has provided an opportunity for experimenting with religious forms from outside the traditionally acceptable forms in a culture. During the 1950s and 1960s, for example, Yoga and Zen meditation were serious religious options for some Westerners and a form of experimentation for large numbers. The concern to experiment with personal experience and with styles of living during the 1960s in the West has itself been considered an important religious expression by some commentators. These years saw considerable exploration in exotic experience with psychedelic drugs, many attempts to set up new communities for group living (communes)—though few lasted more than a year—and a shift in the values of middle class youth from a concern for personal economic security to social and experiential concerns. These recent activities may be viewed as attempts to recapture the experience of the sacred.

Throughout the past hundred years a number of philosophers and social scientists have asserted the disappearance of the sacred and predicted the demise of religion. A study of the history of religions shows that religious forms change and that there has never been unanimity on the nature and expression of religion. Whether or not man is now in a new situation for developing structures of ultimate values radically different from those provided in the traditionally affirmed awareness of the sacred is a vital question. The suggestion that a radically different kind of reality is possible is, of course, nonsense for those to whom the sacred already has been manifested once and for all in a particular form.

(F.J.S.)

WORSHIP

Broadly defined, worship is man's response to the appearance of that which is accepted as the holy—that is, to the sacred, transcendent power or being. Characteristic modes of response to the holy include cultic acts of all kinds: ritual drama, prayers of many sorts, dancing, ecstatic speech, veneration of various persons and objects, sermons, silent meditation, and sacred music and song. Also included in worship are acts of private response: spoken or unspoken prayers, silence, the assumption of particular postures, ritual acts and gestures, and individual acts of veneration of persons or objects.

Nature and significance. The performance of acts of worship rests upon the assumption that there is a realm of being that transcends the ordinary (*i.e.*, secular or profane) "world" of the worshipper. Acts of worship serve to unite, temporarily at least, the ordinary and the transcendent realms through one or more of a variety of possible means. According to the imagery of this assumption, the heavenly world is above and apart from the earthly, and the reality and powers of the heavenly realm are made to be effectively present on earth through acts of worship. The worshipper may thus find himself transported from the earthly to the heavenly world or may perceive the heavenly to descend to the earthly through the movement of worship.

The act of uniting the sacred and profane realms, in

effect, transforms the situation of the worshipper into one that means health, fresh understanding, renewal of life, or salvation. The situation that prompts worship thus calls for change or for the acknowledgment of change. Frequently, life is recognized to be in need of renewal, and worship is viewed as offering the path to such renewal. Some acts of worship arise from the need of the worshipper to exult in praise of the holy and to express his joy or gratitude that his situation, in fact, has changed for the better.

In both instances, the change is widely believed to take place through the worshipper's return, by means of the acts of worship, to primordial time (as in primitive religions), to the realm in which unity and blessedness obtain. Public acts of ritual often include the recitation of myths of creation or of origin; such recitation transports the worshipper from ordinary time and circumstance back to the beginnings of things. The result is the reconstitution of the world itself and of the worshipper within the world.

Significance in maintaining life and order

Worship, especially in ancient societies, was no matter of indifference to the society at large, for the very continuation of life demanded it. In hunting and food-gathering societies, the continuance of sources of food depended upon the performance of ritual acts through which the means of sustaining life were preserved or secured. In agricultural societies, the fructification of the soil took place in relation to acts of worship focussed upon fertility (*e.g.,* in Syrian and Palestinian religions). In the religion of the state (*e.g.,* ancient Rome), the preservation of the society in times of danger depended upon appropriate acts of worship through which the power of the holy was focussed upon the community's particular need.

In ancient societies (and in some contemporary communities) worship was viewed as affecting all aspects of the life of the community, since it was recognized to provide the means for preserving and renewing life itself. Most of the arts developed in relation to worship and to statecraft and law, and the practical (technical) arts generally gained legitimacy and continuing force through their place in the ritual and liturgical acts of the community. In many ancient societies, the chief institutions (*e.g.,* monarchies) and customs of the society were understood to be derivative from their prototypes or archetypes in the realm of the gods. Kingship was patterned upon divine kingship; worship itself had its heavenly archetype; and the representations of the gods and goddesses were modelled upon the divine beings themselves or upon replicas of them in the heavenly place of worship.

Functions of worship. *Primary functions.* The basic function of worship—the establishment and maintenance of the relation between man and the holy—includes many facets. The relation between the holy and the earthly has, however, a noteworthy ambivalence. On the one hand, man's life is enriched and renewed through ever closer relations with the divine. On the other hand, the holy represents threatening, potentially damaging power, for the force of the holy so greatly transcends that of man that its coming is recognized as a grave danger. This double relationship to the gods has been summed up in various ways. The Latin expression *do ut des,* "I give, that you may give," voices some of the dimensions. The worshipper turns to the gods with his gifts (*e.g.,* sacrifices, prayers, words of praise and adoration, and petitions), and the gods receive these and bestow the gifts on which human life depends. The other dimension of the relationship is signalled by the Latin *do ut abeas,* "I give, that you may go (and stay) away." The divine power must be averted in order to preserve human life. The gods can become the enemies of man, and worship can function to keep the gods at a safe distance.

The ambivalent nature of the holy

The rites of worship well document this double attitude of worshippers toward the holy. The sacred precincts are most holy because at them the holy once appeared and continues to appear. Thus, the precincts must be guarded, worship must be performed in the right manner, and the sanctity of the site identified and maintained.

Acts of sacrifice include gifts to the gods in exchange for gifts received or anticipated. They also include offerings entirely devoted to the gods, none of which is touched again by the worshipper; these are sacrifices intended to avert the wrath of the gods or to express the worshipper's complete dependence upon them. The most characteristic sacrifice, however, is one in which both the beneficence and the danger of the holy are affirmed: sacrifices that relate the divine and the human, that express and create communion between God and man. These communion sacrifices generally take the form of a meal (*e.g.,* in Mithraism) that worshipper and deity share. Care must still be taken not to infringe upon the deity's rights or desires, but the mood of such sacrificial meals is one of sacramental participation in the life and beneficence and power of the god.

Secondary functions. Secondary functions of worship—highly significant for the social and personal life of the community—are distinguishable, although their interrelationship is evident. An important function of worship is the creation and maintenance of social concord in societies dominated by one religion. The understandings expressed in worship bind the members of the society together. The acts of worship celebrate and symbolize this unity when the majority of the members of the society regularly engage in common worship. In Muslim lands, for example, the regular division of the day into five parts through the call of the muezzin (official proclaimer) to prayer and the daily gatherings in the mosque unite the society and express its common commitments and character.

A second function of worship is the creation and maintenance of views and attitudinal stances that identify the members of the society to each other and in relation to other groups. Worship thus involves social learning: the members of the community, through their common worship, learn how to plant, to cultivate the soil, to hunt game, to engage in warfare, to settle disputes, to relate to the various strata of the society. Worship displays and reinforces the character of the society; the traditions are passed along through the worship of the community. In this way, acts of worship sum up and reinforce the moral and cultural commitments and understandings of the community. In a situation in which one religion predominates, such social learning pervades the entire society. This situation formerly pertained, for example, in Ethiopia, where the Ethiopian Orthodox Church was closely identified with the character and objectives of the state, and in Buddhist Tibet before the Chinese invasion in 1950.

Worship as societal reinforcement

The worship of a particular group within a society performs the same purposes for that group. Group concord is effected and maintained through the rites and formal acts of group worship. The celebration of fundamental understandings and values through worship bestows solidity and substance on them so that they become a part of the divinely ordained system of laws, customs, and social practices. Doctrines or dogmas are significantly strengthened and reinforced through worship, and religious truths become a part of the very existence of a group as they are embodied in the art, music, drama, and public rites of the worshipping community. Group learning also is effected, since occasions for worship offer opportunity for the group to reflect upon the significance of its history, rites, and traditions and to celebrate the import of these. The remembrance of deliverance from Egypt (*i.e.,* the Exodus, 13th century BC) became a regular part of the celebration of Passover for the Jews.

A third function of worship is that of providing personal support to individual members of a group or of a society. Because life is marked by anxiety, disasters, and dangers from natural and historical happenings, the individual is provided with a sense of well-being through acts of worship. Worship is viewed as relating the disparate elements of life to the life, purposes, and plans of the divine; and in this way the worshipper is enabled to believe that the burden has been shared, or taken over, by the gods.

Tendency toward conservatism in worship

Since acts of worship need to be performed in the "right" way in order to be efficacious, there is a strong tendency toward conservatism regarding the forms and understandings of worship. The desire for release of personal or group anxiety also makes it likely that the practice of worship may support conservatism within the group, since the tendency is to rely upon past solutions to personal or group problems long after the time when such solutions appear

to be entirely satisfactory. This conservatism belongs to the nature of religion, inasmuch as religion deals precisely with those issues of life that yield no easy resolution or solution: the mystery of life itself, the travails of birth, initiation into the community of adults, marriage, sickness, public disasters, and death. It is of great social importance that worship brings to the group and its members a sense of the enduring qualities of life, hope for a doubtful future, and a sense of well-being and health in the midst of trials and illness and danger. The result, however, can also be that the forward movement of the society is inhibited by the religious traditions and claims of a past age that are regularly reinforced through acts of worship.

Types of worship. The forms and types of worship are extraordinarily rich and varied. Three types may be distinguished: corporate exclusive worship; corporate inclusive worship; and personal worship.

Corporate exclusive worship. Exclusive corporate worship is worship that belongs to the group alone. Such exclusive groups may understand their distinct status over against other groups on the basis of a divine mission in the world (*e.g.,* Judaism, Christianity, and Islām), of clan, social, or initiatory distinctions (*e.g.,* totemic societies, Gnostic groups), or by reference to certain ritual or ethical commitments and practices (*e.g.,* Seventh-day Adventists) characteristic of the group. Study of contemporary religious groups discloses many similarities of belief among these exclusive communities, and distinctions considered unique by the group may not be unique at all—but they are perceived to be unique.

Among the exclusive types, the mystery religions (*e.g.,* Eleusinian) of the Mediterranean world are particularly well-known. The worship of such communities (also including Gnostic sects—*i.e.,* Christian dualistic heretical groups) centred in the sharing of secret knowledge concerning the origin of the world, the true nature of mankind, and man's proper vocation and destiny. An elaborate system of initiation brought the new member into the community. The community maintained its exclusiveness through the passing on of the secret lore to new members through rites designed to free the devotee from the hold of the material world and thus prepare the way for his ascent to the realm of the divine, from which he had been separated.

Totemic societies are drawn and held together by the recognition of the significance of the animal or object that embodies and displays the holy in their midst. Signs worn or placed on the body identify the adherents of the society. Groups otherwise quite similar in language, customs, religious rites, social behaviour, and culture are held distinct from one another through the power of the totem, and worship underscores and helps to maintain this exclusivism.

Social or ethical exclusiveness

Certain social or ethical commitments operate to single out the exclusiveness of religious communities and to define their form of worship. The commitment of Mennonites (a Christian group originating during the 16th century) to refuse participation in acts of war, for example, affects the character of their worship as well as of their general religious life. The worship is conducted over against the larger society, especially the power and commitments of the state. It centres upon peace and reconciliation, upon the moral demands laid upon each worshipper, underscoring the need for worship that issues in the service of fellowmen caught up in the evil of warfare.

The worship of the "gathered church" has a similar character: Baptists, Congregationalists, and many of the free churches—*i.e.,* those not connected with the state (including Mennonites)—engage in a form of worship that stresses the need for each member to make his own confession of faith and to identify personally the character of his religious commitment. The new member must enter such a community on the basis of personal testimony and commitment. Those who do not share these commitments are, in principle (if not in actual fact), not to be accepted into membership.

Racial, ethnic, and language distinctions also can operate to create and maintain exclusive communities. The Negro churches of the United States, though open to members who are not Negroes, have become, in many sections of the United States, exclusive communities, largely through the exclusion of Negroes from white churches. The worship of Negro communities has incorporated elements from African religions and has focussed upon forms of worship appropriate to a people oppressed by the larger society and excluded from many of its benefits. The service generally is "freer" than that of the white churches, including a more significant place for congregational singing and responses and more active participation by the congregation than has become customary in most white churches.

Many other exclusive communities could be mentioned: new religious groups in Japan and African countries (*e.g.,* nativistic religious movements), in the United States and Canada (*e.g.,* the Churches of Christ, the Nazarenes, the Black Muslims), and in western European countries. These communities give prominence in worship to those features that called them into existence: sectarian religious concerns, nationalism, dissatisfaction with the worship and ideas of the dominant religious communities, or other distinctive commitments.

Women have been excluded from full participation in acts of worship by several religious communities, though this exclusivism, or discrimination, is being challenged in the 20th century, and changes have begun to occur.

Corporate inclusive worship. The second type is corporate inclusive worship, which probably has been numerically the largest throughout human history. Members of a society are, in virtue of birth, included as members of the worshipping community (*e.g.,* the Lutheran churches of Scandinavia) or at least potential members. Though there may be rites of entrance that are to be observed, these frequently become no more than conventional acts, placing few demands upon the initiate. The ancient Greek and Roman city-states observed acts of worship that were open to the entire populace, since they were a fixed part of the ceremonial and political life of the state. The sanctuaries and ceremonies often were cared for at state expense, and the leaders of worship were officials of the state.

In American Christianity, many churches engage in such corporate inclusive worship, even though they may have their fixed doctrines and requirements for membership. The Holy Communion (Lord's Supper) often is open to all who wish to communicate. Some congregations have been organized as community churches; *i.e.,* not belonging to any of the recognized denominations; in these, worship is, of course, inclusive. Participation in worship is generally open to those who wish to take part, irrespective of creed or religious commitment. Members of one church may become members of another and take active part in the life of the new congregation, including its sacramental life, at will. Many of the distinctions marking off the large number of denominations or churches from one another have lost most of their significance.

Openness of U.S. Christianity

Participation in worship also is much less restricted than formerly in the Roman Catholic Church, especially as a result of the "opening to the world" that followed the second Vatican Council (1962–65). The same inclusiveness is evident in other religions, especially with regard to participation in worship. Non-Jews in many areas are welcome to participate in Jewish worship even though they may not wish to become converts to Judaism. Worship in the temples and shrines of Hinduism or Buddhism or in Shintō shrines in Japan is not restricted to adherents of the religion. In general, the major religions of the world welcome nonmembers to their public acts of worship; special rites or ceremonies, however, may be reserved to members or initiates.

Personal worship. The third type of worship is that of the individual. The individual's worship may centre in public events and ceremonies, but there is ample place in most religious communities for the devotions, prayers, and religious exercises of the individual, either lay or religious. In corporate acts of worship, fixed prayers, confessions, ritual acts, processions, and participation by empathy in the acts of the leader of worship all enrich the individual's own worship. Some persons are best able to worship in the company of fellow-worshippers, finding little meaning in acts of devotion done in solitude or even in the family

circle. Some individuals may well find that special times and places, special rubrics and ceremonies, and a properly enclosed and framed setting for the appearance of the holy are necessary for their worship. Other individuals find the opposite to be the case; for them, the public and fixed occasions for worship lack meaning and intimacy, and they thus need to frame their own prayers, engage in their own devotions, and anticipate the appearance of the holy to themselves alone. In most religious communities and for most persons, a combination of public and private worship appears to be desirable. In the public acts of worship, the range and depth of the religious tradition are represented and affirmed; the power of the holy is made the more palpable. Especially in the modern Western world—in which erosion of traditional religion has occurred to a great extent—individuals are aided in acts of worship by the gathering of members of the community for public worship. They affirm together the faith that becomes increasingly difficult for the individual believer himself to affirm as genuinely his own faith.

Disciplines and techniques

Personal worship, whether public or private, is often aided by the observance of disciplines and techniques that focus the attention of the worshipper upon the sacred or holy. Silence, devotional readings, set prayers, the rosary (beads used as a devotional aid in Roman Catholicism), bodily postures and attitudes, music, and works of art, including the icons (images) of Eastern Christian churches, all serve to help the worshipper to concentrate his apprehension of the power of the holy and to intensify his sense of the presence of the holy. The act of worship may aim at a temporary leaving behind of the ordinary concerns and activities of the individual, so that meditation on the divine may occur. It may have as its aim the union of the self with fellowmen and with the divine, as an act that brings the divine powers effectively into the life of the worshipper for that time and for coming days. Or it may be a part of an act of religious devotion aimed at continuing union with the divine (as in mysticism), the sloughing off of natural existence, or a move toward "divinization" (becoming divine).

Variations or distinctions within the act of worship. Worship may be distinguished with regard to the kind of devotion extended to the holy. Worship (Greek *latreia*) in the narrow sense is considered by many religions to be directed to the divine alone: to God in Judaism, Christianity, and Islām and to Amitābha (Buddha) in Mahāyāna (the Greater Vehicle) Buddhism. To worship any being or object other than God alone is thus understood to be an engagement in idolatry, though other beings, persons, or objects may be shown lesser forms of veneration because of their special relationship to the divine.

Certain persons are viewed as being entitled to major veneration (Greek *hyperdoulia*). Among these, the best known are the Virgin Mary in Christianity, especially Roman Catholic Christianity, the *bodhisattvas* (Buddhas-to-be) in Mahāyāna Buddhism, the prophet Muḥammad in Islām, and Jesus in Christian churches that do not emphasize Jesus as the divine Son of God in their worship.

Lesser, or minor, veneration (Greek *doulia*) is extended to the saints of the church in many Christian groups, but especially in the Roman Catholic Church and in Eastern Orthodox churches. The saints are understood to participate in the power of God in virtue of their holy lives and (often) their martyrdom. The saints make intercession in behalf of the worshipper before God and, joining their voices with his, bring about the blessing sought. The relics of the saints are shown veneration as well and are sometimes believed to effect cures or to perform miracles. The forefathers (patriarchs) Abraham, Isaac, and Jacob were venerated in ancient Israel and were named frequently in prayers to God. Veneration of saints also occurs in Buddhism, Jainism, and Islām.

Worship that places emphasis upon the Virgin Mary or upon the lives or relics of the saints has been called idolatry by reforming groups. The danger of idolatry is held to be its tendency to disperse the commitment of the worshipper, to detract from the glory and honour due to God alone. No person or object in the world of God's creation, according to ancient Israel, was entitled to worship; images of the deity were dangerous because of this fact. According to reforming critics the tendency is to slip from religion into magic whenever worship is not centred upon God alone. Magic and religion are difficult to distinguish, but the operational difference in worship is recognizable: worship is response to the holy, the divine, the powers of which are not controllable. Magic represents an act designed precisely to control the power of the holy and to direct it to one's own ends.

But devotion to the Virgin Mary, to the *bodhisattvas*, to the saints or their relics in various religions, to the icons of the saints in Eastern Christianity should not be considered idolatry. Rather, such devotion is intended to acknowledge the power of the divine and the beauty, nobility, and moral excellence of those who stand in an intimate relationship to God or the sacred realm. Thus, worship of God is accomplished by way of devotion to those whose lives have been touched by the sacred or holy in special ways.

Importance of the religious calendar

Times and places of worship. *Sacred seasons.* Worship takes place at appointed seasons and places. The religious calendar is thus of great importance for the worshipping community, since communities associate worship with critical times in the life of the society. The hunting, planting, and harvesting seasons are of special importance. The beginning of the year (at the time of the spring or fall equinox or of the summer or winter solstice, normally), of the new moon (occasionally, the full moon), or of the week is viewed as an especially auspicious time for acts of worship. Special festivities peculiar to the community's geographical or historical existence also provide fixed occasions for worship.

In communities with an elaborate structure for worship, the day frequently is divided into appointed periods for worship (*e.g.*, in Christianity among monastic communities and in Islām). Days commemorating the birth (*e.g.*, December 25 in Christianity) or death of the founder of the religion may be of special significance for worship. Commemoration of the lives of the saints also involves special prayers and acts of devotion for certain communities.

In the ordering of time for worship, the recognition that the holy appears most powerfully on fixed occasions is important. On New Year's Day in many ancient societies (and in some contemporary communities), the act of worship is viewed as actually recreating the cosmos itself. Through the recitation of the myth of the world's creation, the worshippers are drawn back into primordial time, to the fount of natural and historical existence, and participate in the renewal of the world order. In the ancient Near East, such celebrations were of fundamental significance for the society. The Akitu festival of the Babylonians occurred in the spring, marking the rebirth of nature, the reestablishment of the kingship by divine authority, and the securing of the life and destiny of the people for the coming year. The agricultural rhythm of preparing the soil, planting, watering, harvesting, and waiting for the earth to become ready for planting again was the decisive natural factor in many of these seasonal festivals. The world grew old, its fertility languished, but, at the appointed time, new life began to stir and nature was ready once again to produce its bounty.

Ancient Israelite festivities were, for the most part, nature festivals originally, but they came to be associated with historical events in the life of the community. The barley harvest in the early spring was related to the deliverance (the Passover) of the Israelites from slavery in Egypt. The wheat harvest (Pentecost, or the Feast of the Weeks), about seven weeks later, commemorated the giving of the divine Law (the Ten Commandments) at Mt. Sinai. The celebration of the harvest of the summer fruits and the olives in the early fall (Sukkot, or Feast of Tabernacles) was associated with the period of wanderings in the wilderness, prior to the entrance of the Israelites into the Promised Land (Canaan, or Palestine). In this way, the worship of the community was tied to events in its early history, the powerful attraction of worship connected with natural fertility was held in check, and the community's worship was thereby enabled to focus upon the moral and social demands of the deity. A similar "historicizing" of

seasonal festivals occurred in other religious communities (*e.g.,* Iranian religion, Christianity, Islām).

Sacred places. Worship has its appointed places. A place of worship became sacred and suitable by virtue of the holy's appearing at that place. Sacred places were also sites of natural and historical significance for the community: springs, river crossings, threshing places, trees or groves where the community gathered for public business, hills or mountains where there was safety from enemies, and other such areas. Mountains were of particular importance, since they were understood to bring the worshipper into closer relationship with the heavenly realm.

Significance of sacred sites

A centre for worship takes on a special character, once it has come to be recognized as the place where the holy regularly appears. In some religions it represents the centre of the earth, often called the "navel" of the earth, the place that constitutes the meeting place of God and man, heaven and earth. The sanctity of such a place must be preserved. Thus, the need arises for officials to guard the holy place and to instruct worshippers regarding the kind of acts of worship suitable to the gods at that place. Also, the site must be marked off and its sacred precincts identified. A holy place that once was marked by no more than a sacred stone on which gifts were placed and sacrifices made would thus become the location of a house for the god, a temple.

Places are selected for worship for other reasons. Shrines, temples, and mosques have been built to commemorate a particular experience of an individual leader of the community. Places also become holy because of the association of a holy man with the locality. The home of the shaman (a medicine man with psychic and healing powers), for example, is viewed as holy simply because he, a spirit-filled person, resides there. The place of retreat of a hermit may become a place of pilgrimage and of worship, and the site of a miracle is often commemorated because miracles continue to occur there.

Established places of worship came to be characteristic of the major religions. Temples, mosques, and churches were erected at state expense or through the beneficence of kings, merchants, bankers, or religious leaders. Architectural patterns became established, with the result that mosques, churches, or temples would normally be built in a set style, with a fixed orientation. Many temples and churches were oriented toward the rising sun so that its rays at sunrise would enter the door of the building from the east.

Sacred time and space provide the structure within which worshippers respond to the holy in orderly ways. The danger exists, of course, that such acts of worship at precisely the right time and place may make of worship a routine thing, debilitating the spontaneity of the act or the openness to fresh perspectives and experiences. Orderly and timely worship places bounds upon the fear with which worshippers approach the holy. It provides an established mode of approaching God that can evoke from worshippers genuine spontaneity while offering a setting that is rich in aesthetic and intellectual, as well as spiritual, powers.

Focuses of worship. *Objects.* Religious communities are aided in worship through a variety of objects and activities. The power of the holy is focussed not only in sacred spots and on special occasions but also in animate and inanimate objects. Altars of earth, stone, or metal are extremely common. Some altars are quite simple, formed of beaten earth or consisting of natural stone unshaped by tools. Others are formed of clay or metal or carved from stone, with grain, animals, incense, plants, and flowers the most common offerings at the altar. The altar and the sacrifice both participate in the sacredness of the act of worship and thus are removed from the ordinary realm. The ashes of sacrificed victims must be disposed of with care, just as the altar and the victims must be prepared carefully before the offering occurs. One of the chief duties of the leader of worship is to assist the worshipper in making a proper sacrifice: inspecting the offering, guiding the worshipper as he makes the offering, or performing the act in the worshipper's behalf.

Importance of altars

The sacred scriptures of the religious community, the pulpit or stand from which readings and preaching take place, beads or other objects used by the worshipper as he performs his devotions also focus attention upon the holy and participate in its powers. Images of the gods, totems, or other religious objects—in a variety of forms and materials—also have been employed in worship. Such objects must be understood to represent, not to be identical with, the divine being or power that they portray. Some religious communities (Judaism and Islām in particular) have placed severe limits on the making and use of such representations of the deity. For many religious communities, however, worship without objects representing the gods is impoverished (as in Hinduism); worshippers apparently need such portrayals of the presence of the divine among them. The plastic arts (*e.g.,* sculpture) have flourished as a result of such religious usage, despite the danger that the representation can indeed become identified with the holy and worshippers come to believe that they are enabled to exercise control over the gods.

Activities. Activities likewise have had a significant import in focussing attention on the holy. The divine liturgy of Eastern Orthodox churches provides a dramatic portrayal of the view that God works for the salvation of mankind. Incense, vestments, icons, music, and the processional and ritual movements of the liturgy are united into a re-enactment of Christian deliverance from the powers of sin and death and move the congregation toward active participation in the divine life.

Sacred dances

The sacred dance also has occupied a large place in worship, including dances in connection with hunting, marriage, fertility rites, Islāmic mysticism (dervishes), and the Christian liturgy. Dancing serves in particular to open the way for religious ecstasy, a phenomenon known in many religions. The shaman of Central Asia, the medicine men among the American Indians and Australian Aborigines, and many other leaders in worship are susceptible to ecstatic seizure. Ecstatic utterance was characteristic of the priestesses at Delphi in ancient Greece and of the sibyls (prophetesses) at a number of Greek and Roman cult sites, as well as of participants of Pentecostal worship services in Christianity in the 20th century. Evidence of a person's being overwhelmed and overpowered by the Spirit has been highly valued in many religions and continues to be honoured among some.

Other activities include prayers (public and private, which are a part of almost all acts of worship), the preaching or teaching that accompanies many services of worship, and the active silence of worship (*e.g.,* the Quakers of Christianity). Music is another of the most widespread activities of worship. Certain forms of music are considered unsuitable for worship—the group of free churches known as Churches of Christ, for example, prohibit instrumental music in worship.

Other focuses. Other means for focussing attention on the presence of the holy have a long and significant place in worship. The veneration of ancestors is known in many religious communities (*e.g.,* Confucianism, Shintō); shrines in honour of the ancestors were maintained in Greek and Roman homes in antiquity. Heroes of the tribe, the region, or the city were also focuses for acts of devotion in many religions.

The most noteworthy focus of worship in a vast number of religious communities, however, was the king or the emperor. The king was viewed in ancient Egypt as the incarnate deity, entitled to be worshipped along with the other gods. In early Mesopotamian religion, the king was viewed as the adopted son of God and was venerated along with the high god. Kingship was believed to be a gift of the gods; the king represented the god on earth and partook of his divine powers.

Motivations behind opulent forms of worship

The desire of worshippers to have an example of strength, beauty, wisdom, and riches appears to be the motive behind the great honour lavished upon kings and emperors. Impoverished persons apparently took pleasure in the rich dress, the many wives, the corpulence, and the lavish expenditures of their kings, even as they resented their own deprivation. Worship was believed to be enriched by the indications of excess, the overabundance of vitality and riches. These were pointers to the heavenly world, to

the richness of life for which the worshipper longed and prayed. Thus, much of the trappings of worship and the lavishness of temples, churches, and shrines is accounted for by this longing for opulence on the part of those denied it.

Priests and ministers of religion also serve as focuses for worship. The leader may wish not to be associated too closely with the power of the holy, but, even so, worshippers tend to attach to such persons a special quality of holiness, or a special capacity to mediate the divine powers through acts of worship and through their counsel. The leader's primary function is, in fact, to enable the worshipper to participate more actively in the act that is designed to produce communion between the divine and the human.

Conclusion. It is not necessary to believe in a personal God or a transcendent heavenly power in order to worship. Essential to an act of worship is the belief that there are powers outside of one's present experience that can be brought to bear upon that experience through prayer, meditation, or some other act of worship. A full human life may often require acts and modes of celebration—activities that bring into focus the heights and depths of man's being and experience—that offer a way to transcend and understand ordinary existence and provide renewal of life for man and for the world itself. (W.Ha.)

THE CONCEPT AND FORMS OF RITUAL

The performance of ceremonial acts, or rituals, prescribed by tradition or by sacerdotal decree is a specific, observable mode of behaviour exhibited by all known societies. It is thus possible to view ritual as a way of defining or describing man.

Nature and significance. Man is sometimes described or defined as a basically rational, economic, political, or playing species. Man may, however, also be viewed as a ritual being, who exhibits a striking parallel between his ritual and verbal behaviour. Just as language is a system of symbols that is based upon arbitrary rules, ritual may be viewed as a system of symbolic acts that is based upon arbitrary rules.

The intricate, yet complex, relation between ritual and language can be seen in the history of various attempts to explain ritual behaviour. In most explanations, language becomes a necessary factor in the theory concerning the nature of ritual, and the specific form of language that is tied to explanations of ritual is the language of myth. Both myth and ritual remain fundamental to any analysis of religions.

Three general approaches to a theory about the nature and origin of ritual prevail.

The origin approach. The earliest approach was an attempt to explain ritual, as well as religion, by means of a theory concerned with historical origin. In most cases, this theory also assumed an evolutionary hypothesis that would explain the development of ritual behaviour through history. The basic premise, or law, for this approach is that ontogeny (development of an individual organism) recapitulates phylogeny (evolution of a related group of organisms), just as the human embryo recapitulates the stages of human evolutionary history in the womb—*e.g.*, the gill stage. The solution to explaining the apparently universal scope of ritual depended upon the success in locating the oldest cultures and cults. Scholars believed that if they could discover this origin, they would be able to explain the contemporary rituals of man.

There are almost as many solutions as authors in this approach. In the search for an origin of ritual, research turned from the well-known literate cultures to those that appeared to be less complex and preliterate. The use of the terms primitive religion and primitive cultures comes from this approach in seeking an answer to the meaning of ritual, myth, and religion. Various cultures and rituals were singled out, sacrifice of either men or animals becoming one of the main topics for speculation, though the exact motivation or cause of sacrificial ritual was disputed among the leading authors of the theory. For W. Robertson Smith, a British biblical scholar who first published his theory in the ninth edition of *Encyclopædia Britannica* (1875–89), sacrifice was motivated by the desire for communion between members of a primitive group and their god. The origin of ritual, therefore, was believed to be found in totemic (animal symbolic clan) cults; and totemism, for many authors, was thus believed to be the earliest stage of religion and ritual. The various stages of ritual development and evolution, however, were never agreed upon. Given this origin hypothesis, rituals of purification, gift giving, piacular (expiatory) rites, and worship were viewed as developments, or secondary stages, of the original sacrificial ritual. The Christian Eucharist (Holy Communion), along with contemporary banquets and table etiquette, were explained as late developments or traits that had their origin and meaning in the totemic sacrifice.

Early emphasis on sacrifice

The influence of Robertson Smith's theory on the origin of ritual can be seen in the works of the British anthropologist Sir James Frazer, the French sociologist Émile Durkheim, and Sigmund Freud, the father of psychoanalysis. Although they were not in complete agreement with Smith, sacrifice and totemism remained primary concerns in their search for the origin of religion. For Frazer, the search led to magic, a stage preceding religion. Both Smith and Frazer led Durkheim to seek the origin of ritual and religion in totemism as exemplified in Australia. Durkheim believed that in totemism scholars would find the original form of ritual and the division of experience into the sacred and the profane. Ritual behaviour, they held, entails an attitude that is concerned with the sacred; and sacred acts and things, therefore, are nothing more than symbolic representations of society. In his last major work, *Moses and Monotheism,* Freud also remained convinced that the origin of religion and ritual is to be found in sacrifice.

The functional approach. The second approach to explaining ritual behaviour is certainly indebted to the work of such men as Smith, Freud, and Durkheim. Yet very few, if any, of the leading contemporary scholars working on the problems of religion, ritual, and myth begin with a quest for origins. The origin-evolutionary hypothesis of ritual behaviour has been rejected as quite inadequate for explaining human behaviour because no one can verify any of these bold ideas; they remain creative speculations that cannot be confirmed or denied.

Turning from origin hypotheses, scholars next emphasized empirical data gathered by actual observation. Contemporary literature is rich in descriptions of rituals observed throughout the world. If the term origin can be used as central to the first approach, the term function can be used as indicative of the primary focus of the second approach. The nature of ritual, in other words, is to be defined in terms of its function in a society.

Individual needs and social equilibrium

The aim of functionalism is to explain ritual behaviour in terms of individual needs and social equilibrium. Ritual is thus viewed as an adaptive and adjustive response to the social and physical environment. Many leading authorities on religion and ritual have taken this approach as the most adequate way to explain rituals. Bronisław Malinowski, A.R. Radcliffe-Brown, E.E. Evans-Pritchard, Clyde Kluckhohn, Talcott Parsons, and Edmund Leach, all English or American anthropologists, adopted a functional approach to explain ritual, religion, and myth.

Most functional explanations of ritual attempt to explain this behaviour in relation to the needs and maintenance of a society. The strengths of this approach are dependent upon a claim that it is both logical and empirical. It is a claim, however, that is open to serious criticism. If the aim of functionalism is to explain why rituals are present in a society, it will be necessary to clarify such terms as need, maintenance, and a society functioning adequately, and this becomes crucial if they are to be

taken as empirical terms. From a logical point of view, functionalism remains a heuristic device, or indicator, for describing the role of ritual in society. If it is asserted that a society functions adequately only if necessary needs are satisfied; and if it is further asserted that ritual does satisfy that need, scholars cannot conclude that, therefore, ritual is present in that society without committing the logical fallacy of affirming the consequent. To assert that the need is satisfied "if and only if" ritual is present is a tautology and a reversal of the claim to be empirical.

The history of religions approach. A third approach to the study of ritual is centred on the studies of historians of religion. The distinction between this approach and the first two is that though many historians of religions agree with functionalists that the origin-evolutionary theories are useless as hypotheses, they also reject functionalism as an adequate explanation of ritual. Most historians of religions, such as Gerardus van der Leeuw in The Netherlands, Rudolf Otto in Germany, Joachim Wach and Mircea Eliade in the United States, and E.O. James in England, have held the view that ritual behaviour signifies or expresses the sacred (the realm of transcendent or ultimate reality). This approach, however, has never been represented as an explanation of ritual. The basic problem with it remains that it cannot be confirmed unless scholars agree beforehand that such a transcendent reality exists (see also RELIGIONS).

Functions of ritual. Ritual behaviour, established or fixed by traditional rules, has been observed the world over and throughout history. In the study of this behaviour, the terms sacred (the transcendent realm) and profane (the realm of time, space, and cause and effect) have remained useful in distinguishing ritual behaviour from other types of action.

The sacred and the profane

Although there is no consensus on a definition of the sacred and the profane, there is common agreement on the characteristics of these two realms by those who use the terms to describe religions, myth, and ritual. For Durkheim and others who use these terms, ritual is a determined mode of action. According to Durkheim, the reference, or object, of ritual is the belief system of a society, which is constituted by a classification of everything into the two realms of the sacred and the profane. This classification is taken as a universal feature of religion. Belief systems, myths, and the like, are viewed as expressions of the nature of the sacred realm in which ritual becomes the determined conduct of the individual in a society expressing a relation to the sacred and the profane. The sacred is that aspect of a community's beliefs, myths, and sacred objects that is set apart and forbidden. The function of ritual in the community is that of providing the proper rules for action in the realm of the sacred as well as supplying a bridge for passing into the realm of the profane.

Although the distinction between the sacred and profane is taken as absolute and universal, there is an almost infinite variation on how this dichotomy is represented—not only between cultures but also within a culture. What is profane for one culture may be sacred to another. This may also be true, however, within a culture. The relative nature of things sacred and the proper ritual conducted in relation to the sacred as well as the profane varies according to the status of the participants. What is set apart, or holy, for a sacred king, priest, or shaman (a religious personage having healing and psychic transformation powers), for example, will differ from the proper ritual of others in the community who are related to them, even though they share the same belief systems. The crucial feature that both sustains these relations and sets their limits is the ritual of initiation.

Three further characteristics are generally used to specify ritual action beyond that of the dichotomy of sacred and profane thought and action. The first characteristic is a feeling or emotion of respect, awe, fascination, or dread in relation to the sacred. The second characteristic of ritual involves its dependence upon a belief system that is usually expressed in the language of myth. The third characteristic of ritual action is that it is symbolic in relation to its reference. Agreement on these characteristics can be found in most descriptions of the functions of ritual.

The scholarly disputes that have arisen over the functions of ritual centre around the exact relation between ritual and belief or the reference of ritual action. There is little agreement, for example, on the priority of ritual or myth. In some cases, the distinction between ritual, myth, and belief systems is so blurred that ritual is taken to include myth or belief (see also MYTH AND MYTHOLOGY).

The variety of ritual references

The function of ritual depends upon its reference. Once again, although there is common agreement about the symbolic nature of ritual, there is little agreement with respect to the reference of ritual as symbolic. Ritual is often described as a symbolic expression of actual social relations, status, or the role of individuals in a society. Ritual is also described as referring to a transcendent, numinous (spiritual) reality and to the ultimate values of a community.

Whatever the referent, ritual as symbolic behaviour presupposes that the action is nonrational. That is to say, the means–end relation of ritual to its referent is not intrinsic or necessary. Such terms as latent, unintended, or symbolic are often used to specify the nonrational function of ritual. The fundamental problem in all of this is that ritual is described from an observer's point of view. Whether ritual man is basically nonrational or rational, as far as his behaviour and his belief system are concerned, is largely dependent upon whether he also understands both his behaviour and belief to be symbolic of social, psychological, or numinous realities. It is difficult to imagine a Buddhist, a Christian, or an Australian aborigine agreeing that his ritual action and beliefs are nothing but symbols for social, psychological, or ultimate realities. The notion of the sacred as a transcendent reality may, however, come closest to the participant's own experience. The universal nature of the sacred–profane dichotomy, however, remains a disputed issue.

What is needed is a new theory that will overcome the basic weaknesses of functional descriptions of ritual and belief. Until such a time, ritual will remain a mystery. The progress made in the study of language may be of help in devising a more adequate explanation of nonverbal behaviour in general and of ritual in particular.

Types of ritual. Because of the complexities inherent in any discussion of ritual, it is often useful to make distinctions by means of typology. Although typologies do not explain anything, they do help to identify rituals that resemble each other within and across cultures.

Imitative. All rituals are dependent upon some belief system for their complete meaning. A great many rituals are patterned after myths. Such rituals can be typed as imitative rituals in that the ritual repeats the myth or an aspect of the myth. Some of the best examples of this type of ritual include rituals of the New Year, which very often repeat the story of creation. In a passage from an Indian *Brāhmaṇa* (a Hindu scripture) the answer to the question of why the ritual is performed is that the gods did it this way "in the beginning." Rituals of this imitative type can be seen as a repetition of the creative act of the gods, a return to the beginning.

The "myth-ritual" school

This type of myth has led to a theory that all rituals repeat myths or basic motifs in myths. A version of this line of thought, often called "the myth-ritual" school, is that myth is the thing said over ritual. In other words, myths are the librettos for ritual. The works of such scholars as Jane Harrison and S.H. Hooke are examples of this theory. Although it cannot be denied that some rituals explicitly imitate or repeat a myth (*e.g.*, a myth of creation), it cannot be maintained that all rituals do so. The ritual pattern of the ancient Near East, which Hooke considers basic to the festival celebrating the creation, is itself a typological construction. In any case, although there is a combat and killing narrated in the festival myth, no known evidence exists of ritual killing or of king-sacrifice in the ancient Near East. Nevertheless, some rituals do repeat the story of a myth and represent an important type of ritual behaviour, even though the type cannot be universalized as a description of all ritual action.

Positive and negative. Rituals may also be classified as

positive or negative. Most positive rituals are concerned with consecrating or renewing an object or an individual, and negative rituals are always in relation to positive ritual behaviour. Avoidance is a term that better describes the negative ritual; the Polynesian word *tabu* (English, taboo) also has become popular as a descriptive term for this kind of ritual. The word taboo has been applied to those rituals that concern something to be avoided or forbidden. Thus, negative rituals focus on rules of prohibition, which cover an almost infinite variety of rites and behaviour. The one characteristic they all share, however, is that breaking the ritual rule results in a dramatic change in ritual man, usually bringing him some misfortune.

Variation in this type of ritual can be seen from within a culture as well as cross-culturally. What is prohibited for a subject, for example, may not be prohibited for a king, chief, or shaman. Rituals of avoidance also depend upon the belief system of a community and the ritual status of the individuals in their relation to each other. Contact with the forbidden or transgression of the ritual rules is often offset by rituals of purification.

Negative ritual, as noted above, is always in polarity with positive ritual. The birth of a child, the consecration of a king, a marriage, or a death are ritualized both positively and negatively. The ritual of birth or death involves the child or corpse in a ritual that, in turn, places the child or the corpse in a prohibitive status and thus to be avoided by others. The ritual itself, therefore, determines the positive or negative characteristic of ritual behaviour.

Sacrificial. Another type of ritual is classified as sacrificial. Its importance can be seen in the assessment of sacrificial ritual as the earliest or elementary form of religion. See below under *Sacrifice.*

The significance of sacrifice in the history of religions is well documented. One of the best descriptions of the nature and structure of sacrifice is to be found in *Essai sur la nature et le fonction du sacrifice,* by the French sociologists Henri Hubert and Marcel Mauss, who differentiated between sacrifice and rituals of oblation, offering, and consecration. This does not mean that sacrificial rituals do not at times have elements of consecration, offering, or oblation but these are not the distinctive characteristics of sacrificial ritual. Its distinctive feature is to be found in the destruction, either partly or totally, of the victim. The victim need not be human or animal; vegetables, cakes, milk, and the like are also "victims" in this type of ritual. The total or partial destruction of the victim may take place through burning, dismembering or cutting into pieces, eating, or burying.

Differentiation between sacrifice and other ritual acts

Hubert and Mauss have provided a very useful structure for dividing this type of ritual into subtypes. Though sacrificial rituals are very complex and diverse throughout the world, nevertheless, they can be divided into two classes: those in which the participant or participants receive the benefit of the sacrificial act and those in which an object is the direct recipient of the action. This division highlights the fact that it is not just individuals who are affected by sacrificial ritual but in many instances objects such as a house, a particular place, a thing, an action (such as a hunt or war), a family or community, or spirits or gods that become the intended recipients of the sacrifice. The variety of such rituals is very extensive, but the unity in this type of ritual is maintained in the "victim" that is sacrificed.

Life crisis. Any typology of rituals would not be complete without including a number of very important rites that can be found in practically all religious traditions and mark the passage from one domain, stage of life, or vocation into another. Such rituals have often been classified as rites of passage, and the French anthropologist Arnold van Gennep's study of these rituals remains the classic book on the subject. See below under *Rites of passage; Death rites and customs.*

The basic characteristic of the life-crisis ritual is the transition from one mode of life to another. Rites of passage have often been described as rituals that mark a crisis in individual or communal life. These rituals often define the life of an individual. They include rituals of birth, puberty (entrance into the full social life of a community), marriage, conception, and death. Many of these rituals mark a separation from an old situation or mode of life, a transition rite celebrating the new situation, and a ritual of incorporation. Rituals of passage do not always manifest these three divisions; many such rites stress only one or two of these characteristics.

Rituals of initiation

Rituals of initiation into a secret society or a religious vocation (viz., priesthood, ascetic life, medicine man) are often included among rites of passage as characteristic rituals of transition. The great New Year's rituals known throughout the world also represent the characteristic passage from old to new on a larger scale, that includes the whole society or community.

One of the dominant motifs of the life-crisis ritual is the emphasis on separation, as either a death or a return to infancy or the womb. In India, a striking example is the Hindu rite of being "twice born." The young boy who receives the sacred thread in the *upanayana* ritual, a ceremony of initiation, goes through an elaborate ritual that is viewed as a second birth. Rituals such as Baptism in early Christianity, Yoga in India, and the complex puberty rituals among North American Indian cultures exemplify this motif of death and rebirth in rites of passage.

Rituals of crisis and passage are often classified as types of initiation. An excellent description of such rites is found in *Birth and Rebirth* by Mircea Eliade. From Eliade's point of view, rituals, especially initiation rituals, are to be interpreted both historically and existentially. They are related to the history and structure of a particular society and to an experience of the sacred that is both transhistorical and transcendent of a particular social or cultural context. Culture, from this perspective, can be viewed as a series of cults, or rituals, that transform natural experiences into cultural modes of life. This transformation involves both the transmission of social structures and the disclosure of the sacred and spiritual life of man.

Initiation rituals can be classified in many ways. The patterns emphasized by Eliade all include a separation or symbolic death, followed by a rebirth. They include rites all the way from separation from the mother to the more complex and dramatic rituals of circumcision, ordeals of suffering,,or a descent into hell, all of which are symbolic of a death followed by a rebirth. Rites of withdrawal and quest, as well as rituals characteristic of shamans and religious specialists, are typically initiatory in theme and structure. Some of the most dramatic rituals of this type express a death and return to a new period of gestation and birth and often in terms that are specifically embryological or gynecological. Finally, there are the actual rituals of physical death itself, a rite of passage and transition into a spiritual or immortal existence.

The various typologies of ritual that can be found in texts on religion and culture often overlap or reveal a common agreement in the way in which ritual behaviour can be classified. There is a striking contrast in the use of these typologies to interpret the meaning of ritual. In general, this contrast can be described in terms of two positions: the first emphasizes the sociopsychological function of ritual; the second, although not denying the first, asserts the religious value of ritual as a specific expression of a transcendental reality.

Nonverbal communication and meaning

Conclusion. Ritual behaviour is obviously a means of nonverbal communication and meaning. This aspect of ritual is often overlooked in the stress on the relation of ritual to myth. Thus, the meaning of ritual is often looked for in the verbal, spoken, or belief system that is taken as its semantic correlate. The spoken elements in a ritual setting do often reveal the meaning of a ritual by reference to a belief system or mythology, but not always. Such a connection has led to an overemphasis on the importance of the belief system or myth over ritual. To assert that myths disclose more than ritual ever can is an oversimplification of the complex correlation of these two important aspects of religion. A partial explanation of this emphasis is undoubtedly the fact that a vast amount of data, both primary and secondary, is literary in form. Theories about ritual are either deduced from the primary literature of a religious tradition or are translated into written language as a result of observation.

Ritual can be studied as nonverbal communication disclosing its own structure and semantics. Scholars have only recently turned to a systematic analysis of this important aspect of human behaviour; and progress in kinesics, the study of nonverbal communication, may provide new approaches to the analysis of ritual. This development may well parallel the progress in linguistics and the analysis of myth as an aspect of language.

A complete analysis of ritual would also include its relation to art, architecture, and the specific objects used in ritual such as specific forms of ritual dress. All of these components are found in ritual contexts, and all of them are nonverbal in structure and meaning.

Significance of "sacred time" and "sacred space"

Most rituals mark off a particular time of the day, month, year, stage in life, or commencement of a new event or vocation. This temporal characteristic of ritual is often called "sacred time." What must not be forgotten in the study of ritual is a special aspect of ritual that is often described as "sacred space." Time and place are essential features of ritual action, and both mark a specific orientation or setting for ritual. Time and space, whether a plot of ground or a magnificent temple, are ritually created and become, in turn, the context for other rituals. Examples of ritual time and ritual space orientation can be found in the rituals for building the sacrifice in Brahmanic Indian ritual texts; for the building of a Hindu temple or a Christian cathedral; and for consecrating those structures that symbolize a definite space–time orientation in which rituals are enacted. The shape, spatial orientation, and location of the ritual setting are essential features of the semantics of ritual action.

When particular ritual objects, dances, gestures, music, and dress are included in the study of ritual, the total structure and meaning of ritual behaviour far exceed any one description or explanation of ritual man. Most descriptions are selective and are dependent upon the theory and intent with which rituals are to be studied.

In recent years there has been little consensus among scholars on an adequate theory, or framework, for explaining or describing ritual. Though the term has often been used to describe the determined, or fixed, behaviour of both animals and men, the future study of ritual may disclose that this behaviour, found throughout history and cultures, is as unique to man as his capacity for speaking a language and that change in ritual behaviour is parallel to, or correlated with, change in language. Although great progress has been made in the analysis of man as the species who speaks, the syntax and semantics of ritual man are yet to be discovered.

(Ha.P.)

Prayer

Prayer is an act of communication by man with the sacred or holy—God, the gods, the transcendent realm, or supernatural powers. Found in all religions in all times, prayer may be a corporate or personal act utilizing various forms and techniques. Prayer has been described in its sublimity as "an intimate friendship, a frequent conversation held alone with the Beloved" by St. Teresa of Ávila, a 16th-century Spanish mystic.

NATURE AND SIGNIFICANCE

Prayer as the primary mode of expression in religion

Prayer is a significant and universal aspect of religion, whether of primitive peoples or of modern mystics, that expresses the broad range of religious feelings and attitudes that command man's relations with the sacred or holy. Described by some scholars as religion's primary mode of expression, prayer is said to be to religion what rational thought is to philosophy; it is the very expression of living religion. Prayer distinguishes the phenomenon of religion from those phenomena that approach it or resemble it, such as religious and aesthetic feelings.

Historians of religions, theologians, and believers of all faiths agree in recognizing the central position that prayer occupies in religion. According to the American philosopher William James, without prayer there can be no question of religion. An Islāmic proverb states that to pray and to be Muslim are synonymous, and Sadhu Sundar Singh, a modern Christian mystic of India, stated that praying is as important as breathing.

Of the various forms of religious literature, prayer is considered by many to be the purest in expressing the essential elements of a religion. The Islāmic Qurʾān is regarded as a book of prayers, and the book of Psalms of the Bible is viewed as a meditation on biblical history turned into prayer. The *Confessions* of the great Christian thinker St. Augustine (354–430) are, in the final analysis, a long prayer with the Creator. Thus, because religion is culturally and historically ubiquitous, if prayer were removed from the literary heritage of a culture, that culture would be deprived of a particularly rich and uplifting aspect.

From its primitive to its mystical expression, prayer expresses a desire on the part of men to enter into contact with the sacred or holy. As a part of that desire, prayer is linked to a feeling of presence (of the sacred or holy), which is neither an abstract conviction nor an instinctive intuition but rather a volitional movement conscious of realizing its higher end. Thus, prayer is described not only as meditation about God but as a step, a "going out of one's self," a pilgrimage of the spirit "in the presence of God." It has, therefore, a personal and experiential character that goes beyond critical analysis.

Prayer is also linked to sacrifice, which seems to support prayer as a cultic—as well as a personal—act and as a supplement to the bare word of man in his attempts to relate to the sacred or holy. In any case, the sacrificial act generally precedes the verbal act of prayer. Thus, the presentation of an offering often prolongs prayer and is viewed as a recognition of the sovereignty and beneficence of the deity or supernatual powers. The word of man (in prayer), however, apart from a concomitant sacrificial act, is itself viewed as the embodiment of sacred action and power.

When prayer becomes dominating and manipulative in its intent, it becomes magic. With words and songs, man thus believes that he can ask, conjure, and threaten the sacred or supernatural powers. Imprecation and incantation become, in effect, "oral talismans" (charms). The effectiveness of such magical prayer is believed to depend on the recitation of a precise formula, or rhythm, or on the saying and repeating of the divine name. Manipulation by magic, however, is neither the explanation nor the essence of prayer but rather its deviation and exploitation, a tendency that is to be noticed whenever prayer departs from its basic and essential meaning—*i.e.*, the expression of a desire to enter into contact with the sacred or holy.

ORIGIN AND DEVELOPMENT

Theories of the origin of prayer

During the 19th century, when various evolutionary theories were in vogue, prayer was viewed as a stage in the development of religion from a magical to a "higher" stage. Such theories, which saw in prayer no more than a development of magic or incantation, failed to recognize the strictly personal characteristics of prayer. Even if a scholar could prove the chronological precedence of magical incantations to prayer—which has thus far not been done—he would be derelict in his scholarly duty if he saw in such a precedence the only explanation of prayer. The origin of prayer is to be found—essentially and existentially—in the recognition and invocation of the creator-god, the god of heaven.

Though some scholars, such as Costa Guimaraens, a French psychologist in the early 20th century, have attempted to trace prayer back to a biological need, the attempt, on the whole, has been unsuccessful. If sometimes—especially with exceptional subjects or subjects with fragile nervous systems—the act of prayer is accompanied by corporal phenomena (*e.g.*, bleeding, shaking), such phenomena can accompany it without having provoked it and without explaining its deep inspiration. In order to analyze normal prayer psychologically it is especially important to choose normal subjects. Affective sources such as fear, joy, and sadness doubtless play a role in prayer. Such affectations are expressed in prayers recorded in various religions and particularly in the Psalms of the Bible; but they do not explain the recourse to prayer itself, which is explained by a motivation deeper than affective

elements. The cause and occasion of prayer must not be confused.

Moral sentiments also are integrating elements, but they are accidental to the development of prayer; virtue is not necessarily expressed in the act of praying because there exist atheists of incontestable morality. Morality is more a consequence than a cause of prayer; and it follows more than it prepares for the development of the religious man.

William James and psychologists such as Joseph Segond describe prayer as a "subconscious" and "emotional effusion," an outburst of the mind that desires to enter into communication with the invisible. Experiences of prayer very often, in fact, do include "cries from the heart," "inexpressible laments," and "spiritual outbursts." The psychological explanation has the advantage of probing the subconscious, of describing the various forces that act within men's psyches; but the emergence of the subconscious in the act of prayer is not the essence of prayer since it minimizes the role of intelligence and the will. Among what are called the higher religions (*e.g.,* Judaism, Christianity, Islām, Hinduism, Buddhism), divine action, which is the object of the human action of prayer, violates neither man's consciousness nor his freedom.

Sociologists often explain prayer in terms of the religious environment, which plays an indubitable role in spiritual behaviour. Though prayer supposes a personal belief, that belief is, to a great extent, provided by society. Society creates and regulates social and religious rites and liturgies to express its beliefs, but to explain the origin of prayer solely in terms of an environmental context would be to neglect the inner, personal origins of prayer. That belief is transmitted by society is incontestable, but the channel is not to be viewed as the source. Society itself is, so to speak, a tributary of beliefs that are both received from and given to the collective whole and also from and to each of its members. The collective forms may influence personal prayer, but they do not explain it.

The vertical (divine–human) as well as horizontal (social) dimension of prayer is also expressed in the alternation between speech and silence. Whereas magical formulas are used to coerce the supernatural, liturgical language, even when incomprehensible to the congregation, seeks to lead the participants into an apprehension of the mystery of the divine. In the presence of the mystery of the divine, man often discovers that he can only stammer or that his speech often falters. When this occurs, he frequently expresses his "fear and love" (Luther) or "*tremendum et fascinans*"—*i.e.,* fear and attraction (according to Rudolf Otto, a modern German historian of religion), in apophatic (negative) formulas. Speech with the divine is, in such cases, followed by silence before men, as one apprehends the inexpressible (*i.e.,* the sacred or holy). Religious language, like silence, thus expresses the distance and inadequacy of man in relation to the divine mystery.

TYPES OF PRAYER

Because the various types of prayer are connected and permit a flow from one type to another, it is difficult to conceive of them in terms of rigid classifications. They are enumerated here more on the basis of psychology than on history.

The central position of request in religion

Petition. The role of the request in religion has played such a central part that by metonymy (using a word for another expected word) it has given its name to prayer. However contestable this may sometimes be, it is impossible to refuse to recognize the importance of request, whether it be for a material or spiritual gift or accomplishment. The requests that occur most often are for preservation of or return to health, the healing of the sick, long life, material goods, prosperity, or success in one's undertakings. Request for such goals may be tied to a magical invocation; it may also be a deviation from prayer when it takes the form of a bargain or of a request for payment due: "In payment of our praise, give to the head of the family who is imploring you glory and riches" (from the Rigveda (Rgveda), a sacred scripture of Hinduism). Christianity has never condemned material requests but rather has integrated them into a single providential order while at the same time subordinating them to spiritual values. Thus, in essence though not always in practice, requests are only on the fringe of prayer. As a religion adopts more spiritual goals, the requests become more spiritual: in the *Choephori,* a play written by Aeschylus (a Greek tragic poet of the 6th–5th centuries BC), Electra, the daughter of King Agamemnon, prays, "Grant that I may be a more temperate and a more pious wife than was my mother." Other examples of the transformation to spiritual goals may be seen in the prayers of the ancient Babylonian and Assyrian kings who asked for the fear of God, rather than material benefits, and that of a priest of the Ewe (a West African people) who even asks of his god "That I remain near you and that you remain near me."

Confession. The term confession expresses at the same time an affirmation of faith and a recognition of the state of sin. In Mazdaism (Zoroastrianism and Parsiism), as in ancient Christianity, the confession of faith accompanies the renunciation of demons. The *Confessions* of St. Augustine also illustrate this dual theme. In a similar fashion, ancient and primitive men recognize that their sins unleash the anger of the gods. To counter the divine wrath, a Ewe, for example, throws a little bundle of twigs—which symbolizes the confessor's sins—into the air and he says words symbolizing the deity's response, "All your sins are forgiven you."

The admission of sin cannot be explained only by anguish or by the feeling of guilt; it is also related to what is deepest in man—*i.e.,* to what constitutes his being and his action (as noted by Karl Jaspers, a 20th-century philosopher). The awareness of sin is one of the salient features of religion, as, for example, in Hinduism: "Varuṇa is merciful even to him who has committed sin" (Rigveda). Confession is viewed as the first step toward salvation in both Judaism and Christianity; in Buddhism, monks confess their sins publicly before the Buddha and the congregation two times every month.

Situated at the most personal level of man, sin places him directly before God, who alone is able to grant pardon and salvation. The Miserere ("Lord, have mercy," Psalm 51) of the ancient Israelite king David expresses repentance for sin with an intensity and depth that has a universal value. One of the results of such a dialogue with God is the discovery of the dark depths of sin.

Intercession. Members of primitive societies have a clear sense of their solidarity in the framework of the family, the clan, and the tribe. This solidarity is often expressed in intercessory prayer, in which the needs of others are expressed. In such societies, the head of the family prays for the other members of the family, but his prayers also are extended to the whole tribe, especially to its chief; the primitive may pray even for those who are not members of his tribe (*e.g.,* strangers or Europeans).

The purposes of prayer

Intercessory prayers are also significant in Eastern and ancient religions. In the hymns of the Rigveda the father implores the god Agni (god of fire) for all of those who "owe him their lives and are his family." In the Greek play *Alcestis* by Euripides (5th century BC), the mother, on her death, entrusts the orphans she is about to leave to Hestia, the goddess of the home. Among the Babylonians and the Assyrians, a priesthood was established primarily to say prayers of intercession.

Prayers of intercession to the divine are supported by mediatory minor gods or human protectors (alive or dead), marabouts (dervishes, or mystics, believed to have special powers) in Islām, or saints in Christianity, whose mediation ensures that the prayer will be efficacious.

In biblical religion, intercession is spiritualized in view of a consciousness of the messianic (salvatory) mission. Moses views himself as one with his people even when they fail in their duty: "Pardon your people," he prays, "or remove me from the Book of Life." Such solidarity finds its supreme form in the prayer of Christ on the cross—"Father, forgive them, for they know not what they do"—which St. Stephen (the first recorded Christian martyr) and other martyrs repeated in the course of their sufferings.

Praise and thanksgiving. Praise, in the prayer of primitive peoples, can be traced to salutations, such as in the prayer of the Hottentots (of South Africa) to the New Moon—"Welcome." Praise among most of the ancient

peoples was expressed in the hymn, which was primarily a prayer of praise (whether ritual or personal) for the gift of the created world. Israel praises its Creator for "his handiwork," as does the Qur'ān. Contemplation of the majesty of the universe thus often gives rise to a prayer that is not always completely free from pantheism (the divine in all things) and that can be found all the way from the nature hymns of Oriental religions to the effusions of J.-J. Rousseau, the 18th-century French moralist, embracing the trees and contemplating the sunrise.

Praise—in addition to concerns for the created world—plays an important role in the prayer of mystics, for whom it is a form of adoration. Praise in this instance constitutes an essential element of the mystic experience and celebrates God, no longer for his works, but for himself, his greatness, and his mystery.

When the great deeds of God are the theme of praise, it becomes benediction and thanksgiving. Even when words denoting thanksgiving are not present, the substance of thanksgiving is manifest, even, for example, for the Pygmy of Central Africa, who says to his god, "Waka [meaning God], you gave me this buffalo, this honey, this wine." Mealtime prayers, frequently enunciated in both ancient and modern religions, give thanks for the goods of the earth and are linked to the giving of an offering.

In Christianity, Christ is discovered as the gift of God and in his mission the economy (or mode of operation) of salvation. Thus, the giving of thanks is viewed as man's response, as a spiritual reaction to the benefit received—*i.e.,* the mediatory work of Christ. Because of the cultivation of this expected response, praise and thanksgiving occupy a central position in Christian prayer and in the liturgy, so much so that its name is given to the Eucharistic Prayer (*i.e.,* the Prayer of Thanksgiving).

Motivation behind adoration

Adoration. Adoration is generally considered the most noble form of prayer, a kind of prostration of the whole being before God. Even if the prayer of request is predominant among primitives, they are seized with the feeling of fear and trembling before the numen (spiritual power) of all that is mana (endowed with the power of the sacred or holy) or taboo (forbidden because of association with the sacred). Names given to the divinity in prayers of adoration express dependency and submission, as, for example, in the prayer of the Kekchí Indians of Central America: "O God, you are my lord, you are my mother, you are my father, the lord of the mountains and the valleys." To express his adoration man often falls to the ground and prostrates himself. The feeling of submissive reverence also is expressed by body movements: raising the hands, touching or kissing a sacred object, deep bowing of the body, kneeling with the right hand on the mouth, prostration, or touching the forehead to the ground. The gesture often is accompanied by cries of fear, amazement, or joy; *e.g., has* (Hebrew), *hū* (Islām), or *svāhā* (Hindu).

Adoration takes on its fullest meaning in the presence of the transcendental God who reveals himself to man in the religions of revelation (Judaism, Christianity, and Islām). In the Old Testament prophet Isaiah's vision of the holy (Isa. 6:3), the seraphim (winged creatures) chant to Yahweh: "Holy, holy, holy is the Lord of hosts; the whole earth is full of his glory." This hymn of adoration became a part of the Christian liturgy. The supreme form of adoration, however, is generally considered to be holy silence, which can be found in primitive religion and in ancient religions, as well as in the "higher" religions, and among mystics it expresses the most adequate attitude toward the immeasurable mystery of God: "I am in a dark sanctuary, I pray in silence; O silence full of reverence" (Gerhard Tersteegen, an 18th-century Protestant mystic). Silent adoration is often viewed as the introduction or the response to an encounter with the sacred or holy.

Prayer as an expression of the mystical experience

Unitative: mystical union or ecstasy. Ecstasy is literally a departure from, a tearing away from, or a surpassing of human limitations and also a meeting with and embracing of the divine. It is a fusion of being with being, in which the mystic experiences a union that he characterizes as a nuptial union: "God is in me and I am in him." The mystic experiences God himself in an inexpressible encounter because it is beyond the ordinary experiences of man. The mystical union may be a lucid and conscious progression of contemplative prayer, or it may take a more passive form of a "seizing" by God of the one who is praying.

The mystic, by his goals and actions, is removed from both the world and himself. He discovers in the light and majesty of the divine his own poverty and nothingness and is thus torn between the contemplation of the greatness of God and his own meagreness. St. Francis of Assisi exemplified this dichotomy in his prayer: "Who are you, O God of sweetness, and who am I, worm of the earth and your lowly servant?" Ecstatic prayer goes beyond the frame of ordinary prayer and becomes an experience in which words fail. Mystics speak in turn of unity (*e.g.,* the 3rd-century-AD Greek philosopher Plotinus), of great pleasure (Augustine), or of intoxication (Philo). It is found in the accounts of Hindu, Persian, Hellenistic, and Christian mystics. "You are me, supreme divinity, I am you," says Nimbāditya. The Ṣūfī (mystic) of Islām Jalāl ad-Dīn ar-Rūmī sighs in the same words as a Christian mystic, Angela da Foligno: "I am you and you are me." Mechthild von Magdeburg develops the same kind of reciprocity: "I am in you and you are me. We cannot be closer. We are two united, poured into a single form by an eternal fusion." Such reciprocity that is so complete that it becomes identity is the supreme expression of ecstatic prayer. It is found in all of the mystic writings, from the Orient to the West.

FORMS OF PRAYER IN THE RELIGIONS OF THE WORLD

The forms that prayer takes in the religions of the world, though varied, generally follow certain fixed patterns. These include: benedictions (blessings), litanies (alternate statements, titles of the deity or deities, or petitions and responses), ceremonial and ritualistic prayers, free prayers (in intent following no fixed form), repetition or formula prayers (*e.g.,* the repetition of the name of Jesus in Eastern Christian Hesychasm, a quietistic monastic movement, or the repetition of the name of Amitābha Buddha in Japanese Buddhism), hymns, doxologies (statements of praise or glory), and other forms.

Religions of nonliterate peoples. Prayer is one of the most ancient expressions of religion. Practically the only evidence of early forms left, however, is that to be noted among the most primitive peoples of today. Together with his dependency in relation to his tribe, the primitive man is aware of his dependency in relation to the Supreme Being. He often addresses his prayers, however, to various numina (spiritual powers): the dead, the divinities of nature, protective gods or actor gods, the Supreme Being localized somewhere in heaven, or a feminine divinity linked to the earth (*i.e.,* the great mother). It is impossible to determine the historical precedence of one over the others, and it is difficult to describe the most primitive prayer because certain forms escape modern scholars, so much so that it has been assumed by some that prayer was absent in earliest religion. The first form may have been a cry, then brief formulas repeated as incantations, such as "Come . . . hear me . . . have pity." (*e.g.,* Algonkin Indians of North America).

Internalized prayer is found among the Eskimos of North America, the Algonkin tribes, the Semangs (of the Andaman Islands in the Indian Ocean), and the Aboriginals of Australia. Prayer in gestures is also found among the Semangs. Another form is spontaneous prayer, without any precise formulation, which is found, for example, among the Negritos of the Philippines and the Alacaluf (Halakwulups) of Tierra del Fuego. More developed liturgies and prayer vigils are found among the Negritos and the Pygmies of Gabon.

The concerns of prayers of primitive peoples

The prayers of peoples of a nonliterate society generally are concerned with the self (egoistic) and concerned with well-being (eudaemonic) at the same time; they are clearly pragmatic, concerned above all with food, protection, and posterity. But the higher forms of adoration and recognition of obligations, of confidence, and self-abandonment are to be noted. Among the Australian Aboriginals are prayers on tombs for the dead, so they may be received in heaven, and prayers are also addressed to the spirits of ancestors. Request and pardon accompany sacrifices in

the propitiatory rites of the Semangs. Of special interest is the fact that the Wiradjuri-Kamilaroi of Australia practice public prayer on only two occasions: the burial of a man and the consecration of puberty. They believe that excessive prayer serves no purpose.

Ancient civilizations. From the 3rd millennium BC to the beginning of the Christian Era, forms of prayer changed little among the Assyrians and Babylonians and their descendants. The oldest forms are composed of hymns and litanies to the moon goddess Sin and to the god Tammuz. Though some songs of joy have been found, most are adjurations. Some hymns of thanksgiving tell of gratitude to the divinity for victory over an enemy. One such hymn, addressed to Marduk (the Babylonian sun god), apparently goes back to the 12th century BC. A number of hymns of later date celebrate the king, but their intent is to request divine protection first for him and his country. Preserved in the library of Ashurbanipal (7th-century-BC Assyrian king) at Nineveh is a rather long hymn to the goddess Nana (queen of the world and giver of life), the consort of the god Nabu, son of Marduk and a god of wisdom and science. There also is a long acrostic poem in praise of the god Marduk, creator of heaven and earth, and hymns that the Babylonians recited at the new year, at the beginning of spring, and at the celebration of Marduk.

Other hymns accompany sacrifices, such as in the offering of a young gazelle in place of humans. A most important form of prayer, however, is found in the conjurations and exorcisms of a priest or believer and in lamentations, which are particularly numerous and which often end in a refrain similar to a litany.

Ancient Egyptian piety is preserved in numerous precepts engraved on the backs of scarabs. These engravings sometimes include praises of the divinity ("All good fates are in the hand of God"), statements of confidence, or requests for protection for the one praying and for his whole family ("God is the protector of my life; the house of one favoured by God fears nothing"). Hymns of thanksgiving, such as that of the artist Nebre, who obtained from the god Amon the healing of his son who had been struck with illness because of Nebre's fault, are numerous in ancient Egyptian religion. Protective magic, widely practiced, also utilized formulas of incantation, recited or written, and amulets (charms). Some of the incantation formulas (anonymously written) come from the earliest times, and others, more recent but no less efficient, were composed by magicians. In order to increase their authority and efficacy, several, such as those composed by pharaoh Ramses III (12th century BC) and preserved in Cairo, were attributed in origin to the gods themselves.

Collections of formulas, such as the Egyptian Pyramid Texts and the Book of the Dead, were compilations of magical prayers that allowed the dead to forestall all the dangers and meet all the eventualities. In particular, they contain negative confessions in which the dead man justifies himself before the court of Osiris (god of the dead). The funeral liturgies of the ancient Egyptians have preserved lamentations that echo the family in mourning. Hymns written on papyrus that are compositions in honour of a divinity and that were recited during sacred ceremonies have also been preserved. Such are the hymns of the pharaoh Akhenaton (Amenophis IV, 14th century BC) to the god Aton and the hymns in honour of the god Amon-Re that boast of divine benefits and sometimes confess misery and sin.

In Greece, poetic prayer can be distinguished from ceremonial prayer. The first, like all of the liturgical prayers, contains three essential parts: the invocation of the god, a justification for fulfillment (*e.g.*, sacrifices offered, favours given and received), and a conclusion that formulates the request, such as in the prayer of Diomedes to the goddess Athena in the *Iliad* (written by the Greek poet Homer in the 8th century BC). Generally, the ceremonial prayer followed a ritual pattern: washing of the hands, the prayer proper, then sacrifice and libations. The prayer initiated the liturgical action; without it there could be no ceremony. Prayers often were transformed into hymns, a characteristic of Greek religion. One of the oldest known Greek hymns is that of women devotees of Dionysus, the god of wine and fertility. That such hymns were not always sublime in character is attested to by the comment of a 6th-century-BC Greek philosopher in regard to a Dionysiac festival.

> If they did not hold a procession and sing a hymn to the genitals, it would be an outrageous performance. Hades and Dionysus, in whose honour they rage and celebrate the Bacchus rites, are one and the same.

Another ancient hymn is a morning hymn to Asclepius, the god of healing. All the hymns begin with invocations of the names of the gods to whom they are addressed. The invocation was believed to have an almost magic value. Though there are many individual Greek hymns in existence, the only official collection remaining contains the Orphic Hymns (addressed to the ancient hero Orpheus); it dates from the Greco-Roman period (*c.* 3rd century BC–*c.* 4th century AD).

Roman prayers begin with an invocation to the divinity. Addressing the god is of capital importance and one must be careful not to address the wrong god. In order to avoid this error, there were litanies of 15 gods and goddesses. The prayer itself generally takes two forms, depending on whether it implies a request or is simply limited to praise. The prayer of request has a juridical pattern in which the offering, as a contractual element, dominates. The offering is what jurists call bail bond, a guarantee. The prayer of request's effectiveness depends on a precise formulation, with parallelisms, solemn repetition, and accumulation of synonyms. The verb *precor* ("I pray") is reinforced by many synonyms. Prayers of praise developed out of meditation or experiences of religious elevation and utilized various patterns in both public and private ceremonies. An example of collections of prayers of praise is preserved in the *Verba pontificalia* ("Priestly Words").

Another form of prayer is the *votum* ("vow"), in which a person undertakes to offer to the divinity, in exchange for divine favour, a sacrifice, the building of a temple, or other such offerings. It is a kind of bargain in which is still felt the prudence of the peasant who has experienced failure. These *vota* ("vows") become more numerous than other prayers the farther one goes from the historical origins of Rome. The most solemn form of the vow is the *devotio* ("act of devotion"), by which a chief offers himself to the divinity in order to obtain victory.

Religions of the East. Although the religion of the Vedas contains private prayers, it gives importance and hieratic stature to liturgical prayer, which may or may not include sacrifice. There exists a whole series of hymns, such as the morning hymn addressed to Agni (the god of fire), who brings light, and to the two Aśvin (twin gods of light). There is also an evening prayer, the *sāvitū*, more precisely a prayer for dusk, which the disciple of the Brahmins (priestly teachers) says at nightfall until the stars appear, and a benediction formula. The gestures of adoration (*upasthāna*) in effect give more intensity to the prayer. The prayers that accompany sacrifices and the numerous hymns of the Rigveda (Ṛgveda; a collection of sacred ritualistic lyrics), which were composed by the members of the priestly caste according to a stereotyped and schematic form, are addressed to the greatness of the divinity in exaltation of his great deeds.

In Hinduism there is an elementary form of prayer—*i.e.*, an affirmation of homage and refuge with the divinity. More frequent is a more elaborate prayer in two forms: *dhyāna* ("meditation") and the *stotra* ("praise"). The *stotra* occurs in a variety of subforms and generally opens with an invocation. It is often characterized by a sort of litany of the titles given, for example, to Vishnu (Viṣṇu; the preserver god) or Śiva (the destroyer god). The *Śivasahasranaman* ("The Thousand Names of Śiva") lists 1,008 titles. In this hymn, each strophe ends with the same refrain. When recited with concentration and pure heart, these prayers are believed to achieve remission of sins.

Ceremonial prayers

Hindu mysticism gives great importance to spoken prayer, which, by progressive absorption, leads to ecstasy. The scale of the prayer of Hindu mystics is exemplified in the five stages of *bhakti* ("devotion") as taught by the Hindu mystic Caitanya (15th–16th century AD), who uses

Hindu mystical prayer

the metaphor of love in social relationships: *śānta* (peaceful love), *dāsya* (servant of God love), *sakhya* (friendship with God), *vātsalya* (filial attitude toward God), and *mādhurya* (love of God as one's lover). "When I was no longer capable of recognizing, I said me and mine. I am you and you are mine" (*Nalayiram*).

In Chinese Buddhism and Taoism, in addition to prayer that accompanies sacrifice, there is the monastic prayer (mu-yu), which is practiced morning, noon, and night to the sound of a small bell. There is also a prayer for the dead, related to the transmigration of souls, which is recited at funerals, the 30th day, the anniversary of the death, and the celebration of the deceased's day of birth. Taoism gave increased importance to this latter form.

Private prayer prepares the way for liberation and illumination. The *tsai-fei* is a prayer—to accompany abstinence—that monks will recite for a believer on payment of alms. Other prayers accompany vows and pilgrimages. Both Buddhist monks and laymen use a string with 108 beads, which monks always carry in their hands.

Religions of the West. In Judaism is one of the best known collections of prayers, the 150 psalms in the Bible. In these psalms, which always presuppose a collective witness, though they may be used by an individual privately, praise is descriptive (God is . . .) or narrative (God does . . .) in nature. Also included are hymns, exhortations to praise God, and supplications. The psalms of request include lamentations and songs of confidence or gratitude. Whether individual or collective, the psalms have a rather similar structure: a cry to God, a confession of sins, a protestation of innocence, and imprecations against one's enemies.

To the prayers of the Bible, the rabbis (religious teachers and leaders) added the Shema ("Hear"), which is a confession composed of three quotes from the Bible (Deut. 6:4–9, 11:13–21, Num. 15:37–41) with attendant blessings and which the Israelite recites daily. At the time of Christ, there appears the prayer par excellence, the *tefilla* or *'amida* (standing prayer), also called *shemone 'esre* ("18 Benedictions"), which every Israelite recites two or three times a day. To these must be added the benediction before eating that raises the meal to the level of the dignity of a religious act.

Relationship of Christian to Jewish prayers

Christianity preserves the doxologies and benedictions from its Jewish heritage. It adds to them the Lord's Prayer, psalms, hymns, and canticles, the first specimens of which are furnished by the New Testament (*e.g.*, the *Nunc dimittis*, "Now let your servant depart"). Christian prayer, like that of other religions, includes liturgical prayer and personal prayer. Liturgical prayer frames and explains more especially the sacraments of Baptism and the Eucharist (Lord's Supper).

The liturgical collection, for Sundays as well as other days, includes readings from the Bible, collects (brief prayers including an invocation, petition, and conclusion in which the name of Jesus is called upon), and a litany (general prayer) for the intentions of the universal church. During the Eucharist, there is a consecration of the bread and wine to be used in the sacred meal. This consecration prayer is called the Eucharistic (Thanksgiving) Prayer, a long prayer in which the element of thanksgiving is dominant. Addressed to the Father, through the mediation of the Son, and in the Holy Spirit, this prayer develops, like the Jewish liturgies, from praise, to thanksgiving, to the memorial (or anamnesis), and finally to an invocation of the Spirit (epiclesis). Originally improvised and spontaneous, this liturgical prayer became fixed in stereotyped forms, first in the West, then—though with more flexibility—in the East.

The first Christians retained the custom of praying three times a day, reciting the "Our Father" (Lord's Prayer). Special times for prayer are morning and evening. Christ's custom of praying at meals (as a devout Jew) is also maintained. This framework can and does favour the life and spirit of prayer that make a Christian existence, according to the words of Clement of Alexandria, a 2nd–3rd-century theologian, "an uninterrupted celebration." Bible readings, silent prayer (in the West especially), brief, fervent invocations, and the repetition of formulas like the Kyrie eleison ("Lord, have mercy") in the East have enriched spiritual life and have led monks and laymen to contemplative prayer, as is shown by the growth of mysticism in both the West and East.

From its beginning in the 7th century AD, the most important part of Islāmic liturgy has been the ritual prayer called the *ṣalāt* (daily prayer), in which both Christian and Jewish influences can be seen. This minutely detailed prayer is recited while the suppliant turns toward Mecca (in Saudi Arabia) five times a day. On Friday, the *ṣalāt al-jum'ah* (Friday prayer) replaces the noon prayer. It is celebrated by the community in the principal mosque and includes preaching and a *ṣalāt* of two ritual bowings. Twice a year, at the end of Ramaḍān and the 10th month, a solemn *ṣalāt* is celebrated, similar to Friday's.

Islāmic prayer is an act of adoration of Allāh (God) and thus it would not be suitable to add a request. Before adoring God the believer must purify himself by means of ablutions in pure water or, failing this, in sand. The prayer is accompanied by a meticulous ceremonial with prostration of the body (*rak'ah*). The sense of adoration and conversation with Allāh has led many spiritual Muslims to the heights of mysticism (Ṣūfism).

In Mazdaism, Avestan (scriptural) prayer, sacerdotal prayer, and the prayer common to priests and laymen alike can be distinguished. In the very first poem of the Avesta, Zoroaster (Zarathushtra) presents himself to Ahura Mazdā (the Good Lord) in a prayer that ends with these words: "I will sing for you again praises of great value." What is characteristic of these hymns is that man proceeds almost exclusively by questions and answers. Only priests can understand the ceremony of the *Yasna* (the sacrifice), during which they recite verses from the Avesta, adding to it the *Visp-rat* (shorter liturgy), with or without the *Vidēvdāt* ("Law against the Demons"), which is concerned with ritual purity. Songs (involving light symbolism) accompany the five fire ceremonies that are celebrated daily. There are also ceremonies in which both priests and laymen participate. The great Bāj, a ritual offering of consecrated bread, grain, and butter, begins with a long preface: "In the name of God, Lord Ormazd, may your power and glory increase." The Satum, in praise of the dead, is recited at the beginning of a meal prepared in their honour every month for the first year after a death and then on each anniversary. Other prayers accompany benedictions, especially those used at the consecration of fire, initiation, and marriage. To these must be added the prayers of great purification.

CONCLUSION

Though historians of religion, psychologists, and anthropologists debate various theories concerned with the origin of prayer, the act of prayer itself is of great significance to the believers of all religions, whatever their inspiration, revealed or otherwise. Ludwig Feuerbach, a 19th-century German philosopher, summed up the significance of prayer when he stated, "The most intimate essence of religion is revealed by the most simple religious act: prayer."

As a religious phenomenon, prayer—in terms of its evolution—appears to be neither universally progressive nor progressively regressive. Its great moments and the appearance of men of prayer at various times, whether simple men or men of genius, are found throughout its long history, which thus marks it as a significant and characteristic element of most, if not all, religions. Whether halting or mystical, ceremonial or personal, prayer expresses the experience of a mystery that envelopes and surpasses man. In the presence of that mystery, prayer seeks and establishes dialogue.

(A.G.Ha.)

Creed and confession

Creeds and confessions of faith are authoritative formulations of the beliefs of religious communities (or, by transference, of individuals). The two terms are sometimes used interchangeably, but when distinguished a "creed" refers to a brief affirmation of faith employed in public worship or initiation rites, while a "confession" is generally a longer,

more detailed, and systematic doctrinal declaration; the latter term is usually restricted to such declarations within the Christian faith. Both creeds and confessions were historically called symbols, and the teachings they contain are termed articles of faith or, sometimes, dogmas.

Role of belief

The role of belief within religion is interpreted differently in the various empirical disciplines and by the proponents of particular theological or philosophical positions. Traditionally, it has been considered the primary factor in religion, but some modern scholars often regard beliefs as rationales for ritual, that is to say, as secondary expressions of religious experience or as a posteriori ideological sanctions for social and cultural patterns. The present article follows a current anthropological and sociological tendency to define religion as a symbolic system in which ideas and their concomitant attitudinal aspects and actions provide to an individual or group a model of itself and its world. From this perspective, every religion involves distinctive views or beliefs regarding the nature of ultimate reality.

CREEDS IN THE MAJOR RELIGIONS

Origins and functions of creeds. These beliefs, however, need not be explicitly articulated but may be wholly embedded and transmitted in rituals, myths, and social structures and practices. This is especially true in primitive religions. Even when differentiated from other factors, beliefs are frequently not stated in creedal form but are diffusely expressed in sacred writings, legal codes, liturgical formulas, and theological and philosophical reflection. This was true in the ancient cultural religions of Egypt, Mesopotamia, Greece, and Rome, and in traditional Hinduism, Confucianism, and Taoism. When, however, a religion is transmitted from one culture to another (as from Semitic to Hellenistic; *i.e.,* Palestine to Rome) or claims some degree of universal or exclusive truth, formal creeds often develop as aids in maintaining continuity and identity. They serve this purpose because the relative abstractness, comprehensiveness, and concentration of the verbal expressions of beliefs enable them to serve better than most other forms of religious symbolism as stable identifying marks in pluralistic, changing, proselytizing, and missionary situations.

Purpose of creeds

Creeds in the full sense are therefore found only in so-called universal religions, such as Zoroastrianism, Buddhism, Judaism, Christianity, Islām, and certain modern Hindu movements (*e.g.,* Brahmo Samaj). Even here they are of variable importance, with some groups rejecting all formal creeds. Confessions are less common. They function to define the distinctive beliefs of opposing or uniting groups within a given religion or to formulate doctrines appropriate to new circumstances, and are chiefly a Christian phenomenon during the period from the Reformation to the present.

Religions of the East. Related to creeds in the full sense are certain words and phrases which have partially creedal functions. Terms like *tao* (literally, the "way") in Taoism or *li* (rules of propriety) and *hsiao* (filial piety) in Confucianism summarize fundamental emphases of the religious systems of which they are a part. The endlessly repeated *mantra* (evocatative sacred syllables) of magic invocation, *Oṃ maṇi padme hūṃ* ("O, the jewel in the lotus"), especially popular in Tibetan Buddhism, is in one sense a profession of belief in the Avalokiteśvara (jewel's) presence in the world (lotus). Various Hindu *mantras,* most notably the Gāyatrī prayer from the Rigveda (Rgveda) (3.62.10) that is learned as part of the initiation rites of Brahmin youth, also serve in part as professions of faith. Indeed, it is primarily through liturgical utterances (*e.g.,* the Lord's Prayer in Christianity), that religious identity is signalized and faith confessed in most religions.

More specifically creedal is the early thrice-repeated *triratna* of Hinayāna Buddhism: "I take my refuge in the Buddha. I take my refuge in the *dharma* (doctrine). I take my refuge in the *saṅgha* (monastic community)."

Religions of the West. Even earlier perhaps are such Zoroastrian formulations as "I profess myself a Mazdāh-worshipper, a Zarathustrian, enemy of the demons, servant of the Lord" (*Yasna* 12.1), whereby the believer declared himself a monotheist, a member of a specific community, and a dualist.

Islām. The intensely anti-polytheistic faith of Islām is summed up in the *shahādah:* "there is no God but God; Muḥammad is the Prophet of God." This is proclaimed in the daily calls to prayer from every mosque, and every Muslim must recite it aloud with full comprehension and assent at least once in his life, and profess it without hesitation until his death. Doctrinal disputes have contributed to the development of additional creedal formulations called *ʿaqāʾid* (singular, *ʿaqīdah*), but these do not divide Islām into clearly marked confessional groupings or denominations such as exist in Christianity.

Judaism. In Judaism, the central affirmations of belief are parts of worship; *e.g.,* the confessions of the oneness of God in the Shema (Deut. 6:4 "Hear, O Israel: The Lord our God is one Lord") and of the resurrection of the dead in the *amidah* (standing prayer). Of the various medieval attempts to formulate creeds, the most enduring has been Maimonides' Thirteen Principles of Faith, but these have never become formally binding. The Reform movement's doctrinal declarations, such as the Pittsburgh Platform (1885), have been without lasting influence. The reason for this paucity of creeds is that Jewish identity has been chiefly defined in terms of the observance of the commandments and of the Oral Law, not the acceptance of doctrines.

Christianity. In Christianity, in contrast, there are over 150 officially recognized creeds and confessions. In part this is because the church was from the beginning doctrinally oriented, making the acceptance of a specific kerygma (proclamation) a condition for membership. The faith of the community was expressed in acclamations such as "Jesus is Lord" (*e.g.,* Rom. 10:9, I Cor. 12:3) and in longer, partly stereotyped summaries of essential beliefs (*e.g.,* I Cor. 15:3 ff.) For the New Testament community, in contrast to some Christian groups in later times, a creedless Christianity was inconceivable.

Fully formed creeds first developed for use in baptismal rites and catechetical instruction. They generally had three sections concerned with God the Father, Jesus Christ, and the Holy Spirit, but were variable in wording and content and only gradually became standardized.

The Ecumenical Creeds: Apostles', Nicene, and Athanasian

This process culminated in the West in the Apostles' Creed, which is now almost universally recognized by Western churches, and is still used in baptismal rites as well as public worship by Catholics and most Protestants. This creed is wholly derived from New Testament affirmations, but the 5th-century legend that the Twelve Apostles were its authors is without foundation. Not until the 8th century is it quoted in its present wording. Its sources, however, are to be found in earlier baptismal creeds, most probably in the Old Roman Symbol, which appears to go back in its essentials to the 2nd century. As is true of other creeds, it is in part intended to exclude heretical views. For example, against Gnosticism and Marcionism (dualistic heresies), it emphasizes that God, not an evil demiurge, is the creator of the world, and against docetic views that Jesus was a heavenly being with a phantom body, it insists that he was born of the Virgin Mary and actually suffered and died and was buried.

The Nicene Creed exists in two versions and represents a new type of doctrinal statement. It was first formulated at Nicaea in 325 by the first of the universal, or ecumenical, councils, after Christianity became the official religion of the Roman Empire, and was designed not as a baptismal confession but as a binding standard of orthodox teachings. Its second version has become the most fully ecumenical of Christian creeds, accepted in East and West alike, including the major Protestant bodies. In Eastern churches, it is regularly employed in both Baptism and eucharistic worship; in the West, only in the Eucharist, and chiefly by Roman Catholics, Anglicans, and Lutherans.

The first version of this formulary is that promulgated at the Council of Nicaea in 325, but the second version, the "Niceno-Constantinopolitan Creed," which has everywhere become standard and is generally referred to as the Nicene Creed, was affirmed at the Council of Chalcedon (451) as the Nicene "faith of the 150 fathers" (*i.e.,* the

Council of Constantinople of AD 381). In 4th-and 5th-century usage, "the Nicene faith" did not refer to the creed of Nicaea as such, but rather to its teaching.

Both versions make the same fundamental affirmations against the Arian heresy that denied the equality of the Father and the Son, asserting that Jesus Christ, the Son of God, is *homoousios* ("of one substance") with the Father. They are also both derived from Eastern baptismal formulas, though which ones is in dispute.

The *filioque* clause, affirming that the Spirit proceeds "from the Son" as well as the Father, was inserted into the text in Spain during the 6th century and gradually spread to all Western churches, but was probably not used in Rome itself until 1014. Eastern Christians continue to reject this addition, though now they do not generally regard it as heretical, especially if it is understood in the sense of "through the Son."

The Athanasian Creed, also called the *Quicumque vult* from its initial words, is the last of what in the West are regarded as the three catholic or ecumenical creeds. It has received some slight recognition in the East, but only since the 16th century. While officially accepted in the Roman Catholic, Anglican, and Lutheran communions, its liturgical use has greatly declined in recent centuries. In part this is because it is in form more a theological exposition than a creed, and in part because of the damnatory clauses that exclude from salvation all those who do not accept every detail of its teaching. The main themes are the nature of Christ and the Trinity, and these are developed in opposition not only to Arianism but also apparently to later heresies such as Nestorianism and Eutychianism. While its doctrine can in general be attributed to the 4th-century Church Father Athanasius, he was not its author. It probably originated in southern France about 450–500, although there is no scholarly consensus on this point.

CONFESSIONS OF THE CHRISTIAN FAITH

Origins and functions of confessions. Official doctrine has chiefly developed during later periods of church history by the formulation of confessions of faith, rather than new creeds. This process did not begin, however, until the 16th-century Reformation. During the Middle Ages, dogmas evolved slowly, almost unconsciously, and then were ratified from time to time by decisions of the church councils, such as the decision on the seven sacraments at the Council of Ferrara-Florence in 1439. The Protestant Reformers, however, were confronted with the need to define and make legitimate their views over against the established system, and thus issued comprehensive manifestos that, much more than the early creeds, were not only catalogues of beliefs but also interpretations and apologies for them. The Roman Catholic and Eastern Orthodox churches responded with their own confessional statements.

The Augsburg Confession

Lutheran confessions. The Augsburg Confession (1530) was the first of these statements, and still remains the most authoritative standard in Lutheran churches. It (as well as the Apology of the Augsburg Confession of 1531) was written by Philipp Melanchthon and approved by Martin Luther, and presents an irenic statement aiming to show that the pope and his allies, not the Reformers, had departed from Scripture and the tradition of the early Fathers. Luther's Small Catechism also enjoys official status in all Lutheran churches and has been determinative for most Lutheran preaching and instruction. The Formula of Concord (1577) further defined the Lutheran position in reference to controversies both within and outside the ranks. These four writings, together with the Large Catechism (1529), the Schmalkald Articles, and the Treatise were assembled into the *Book of Concord* (1580), which has official status in many Lutheran churches.

The Heidelberg Catechism

Reformed churches confessions. In the Reformed tradition stemming from John Calvin (1509–64) and Huldrych Zwingli (1484–1531), each national church produced its own confessional documents. No one of these is authoritative for all, though some (*e.g.,* the Heidelberg Catechism; 1563) are widely esteemed and used. In Switzerland, the First (1536) Helvetic Confession and the Second (1566) Helvetic Confession are the most generally accepted. The French Gallican Confession of 1559 is much admired, and in the Low Countries, the Belgic Confession of 1561 is important. The Netherlands was also the site of the international Synod of Dort (1619) that presented an especially rigid statement of Calvinism against Arminianism (a view that asserted the compatibility of God's sovereignty and man's free will). This same emphasis, combined with Puritan covenantal theology, is reflected in the English Westminster Confession of 1646 that in Scotland replaced the Scots Confession in 1560, was adopted with modifications by Congregationalists and many Baptists, and still remains standard for American Presbyterian churches, though with some revisions.

The Anglican Communion. The Thirty-nine Articles (1563) is the only doctrinal formulation other than the early creeds recognized in the Church of England and its offshoots, but its authority is not great. In the Anglican Communion, *The Book of Common Prayer* plays the identity-sustaining role served by confessions in Lutheran and Reformed churches. The Thirty-nine Articles, abbreviated to 25, are also the chief doctrinal standard in the Methodist churches, but their authority is uncertain.

Confessions of other Protestant groups. Confessional documents are of little significance for most of the radical groups (*e.g.,* Anabaptists) coming out of the Reformation. To be sure, the Anabaptist Schleitheim Confession (1527) was historically important, the Dordrecht Confession (1632) still has some standing in Mennonite churches, and various Baptist and Congregationalist statements could also be mentioned. The general tendency in these churches, however, has been to oppose formal creeds and confessions for fear of stifling the workings of the Holy Spirit or imperilling the sole authority of the Bible or, in theologically liberal circles, endangering freedom of thought and conscience.

Roman Catholic doctrinal statements. Roman Catholic doctrinal statements are not usually called confessions, but the presentation of the distinctive points of Catholic dogma in the Decrees and Canons of the Council of Trent (1564) is as fully elaborated as are Protestant confessional writings. The dogmatic constitutions of the first Vatican Council (1869–70) and papal definitions of the dogmas of the Immaculate Conception (1854) and of the Assumption (1950) also have some of the character of confessions.

The Decrees and Canons of the Council of Trent

Eastern Orthodox doctrinal statements. Eastern Orthodoxy responded to Protestant and Roman Catholic challenges with the confessions of Peter Mogila, Metropolitan of Kiev, in 1643 and of Dosítheos, the Patriarch of Jerusalem, in 1672, both adopted by the Synod of Jerusalem (1672), as well as with the Catechism of Philaret, Metropolitan of Moscow, revised and approved by the Holy Synod in 1839. The Orthodox, however, place little emphasis on these documents, for they regard only the Nicene Creed with its Chalcedonian additions as fully authoritative, and in practice also treat their historic liturgies as doctrinally more important than later statements.

Creeds and confessions today. Recently new types of confessions have begun to emerge. With the decline of state churches, confessions are no longer legally established norms and can once again regain their original function of witnessing to basic convictions. Especially notable in this respect is the Barmen Declaration, formulated in 1934 by a group of Reformed and Lutheran churchmen in opposition to the Nazi-influenced "German Christians." Because of the advance of the ecumenical movement, recent confessional statements have usually been unitive rather than divisive. The doctrinal basis of the World Council of Churches is limited to the affirmation that it is "a fellowship of churches which accept our Lord Jesus Christ as God and Savior" (1961). Preparation of joint Protestant and Roman Catholic official translations into English of the Apostles and Nicene Creeds commenced in 1969. Another characteristic of contemporary doctrinal statements, such as those of the Roman Catholic second Vatican Council (1962–64) and the Presbyterian (U.S.A.) Confession of 1967, is the attempt to reformulate traditional beliefs in ways appropriate to modern circumstances.

The Barmen Declaration

Despite these developments, creeds and confessions are losing influence in both Christian and non-Christian

groups. They are, among other things, often attacked as obstacles to the individual's freedom of thought. This objection applies with special force against a fideistic attitude, such as is illustrated in extreme form by the well-known saying attributed traditionally, though not altogether correctly, to the 2nd-century North African Church Father Tertullian, *credo quia absurdum est,* "I believe because it is absurd." It is less applicable to another ancient and theologically more common approach summed up in the 11th- and 12th-century theologian Anselm's (and, in a somewhat different wording, Augustine's) classic phrase, *credo ut intelligam,* "I believe in order that I may understand." The latter view claims that true faith promotes rather than suppresses inquiry and intellectual liberty.

Yet, whatever the merits of such views, doctrinal convictions are clearly weakening, even in traditionally creedal and confessional bodies. The search for creedless religion is widespread. There is the possibility, however, that this trend may be eventually reversed because the quest for religious community is also strong, and may require the formation or re-affirmation of community-identifying beliefs; *i.e.,* of creeds or confessions. (G.A.L.)

Sacrament

The Latin word *sacramentum,* which etymologically is an ambiguous theological term, was used in Roman law to describe a legal sanction in which a man placed his life or property in the hands of the supernatural powers that upheld justice and honoured solemn contracts. It later became an oath of allegiance taken by soldiers to their commander when embarking on a new campaign, sworn in a sacred place and using a formula having a religious connotation.

NATURE AND SIGNIFICANCE

When *sacramentum* was adopted as an ordinance by the early Christian Church in the 3rd century, the Latin word *sacer* ("holy") was brought into conjunction with the Greek word *mystērion* ("secret rite"). *Sacramentum* was thus given a sacred mysterious significance that indicated a spiritual potency. The power was transmitted through material instruments and vehicles viewed as channels of divine grace and as benefits in ritual observances instituted by Christ. St. Augustine defined sacrament as "the visible form of an invisible grace" or "a sign of a sacred thing." Similarly, St. Thomas Aquinas wrote that anything that is called sacred may be called *sacramentum.* It is made efficacious by virtue of its divine institution by Christ in order to establish a bond of union between God and man. In the Anglican catechism it is defined as "an outward and visible sign of an inward and spiritual grace."

The term sacrament has become a convenient expression for a sign or symbol of a sacred thing, occasion, or event imparting spiritual benefits to participants; and such signs or symbols have been associated with eating, drinking, lustration (ceremonial purification), nuptial intercourse, or ritual techniques regarded as "means of grace" and pledges of a covenant relationship with the sacred order. In this way the material aspects have become the forms of the embodied spiritual reality.

TYPES AND VARIATIONS

Types. The several types of sacraments (*i.e.,* initiatory, purificatory, renewal, communion, healing, cultic elevation) are well exemplified in Christianity, though they also may be found in other Western religions, the Eastern religions, and preliterate religions. In the 12th century the several sacraments of the Western Christian Church were narrowed by Peter Lombard (12th-century theologian and bishop) to seven rites, viz., Baptism, confirmation, the Eucharist (the Lord's Supper), penance, holy orders, matrimony, and extreme unction. This enumeration was accepted by St. Thomas Aquinas, the Council of Florence in 1439, and subsequently by the Council of Trent (1545–63). All these rites were thus affirmed by the Roman Catholic Church as being sacraments instituted by Christ. The number, however, has been modified by modern theologians since the precise origins of some of the seven sacraments are uncertain. Protestant Reformers of the 16th century accepted two or three sacraments as valid: Baptism, the Lord's Supper, and, in some fashion, penance. Both Roman Catholicism and Eastern Orthodoxy accept the sevenfold enumeration. In addition to these, any ceremonial actions and objects related to sacraments that endow a person or thing with a sacred character have been designated "sacramental," though they are differentiated from those of dominical (*i.e.,* Christ's) institution in conveying divine grace *ex opere operato* (it works by itself) or in conferring an indelible character on the recipient, such as Baptism, confirmation, and holy orders. Sacramentals include the use of holy water, incense, vestments, candles, exorcisms, anointing and making the sign of the cross, fasting, abstinence, and almsgiving.

Christian sacraments

"Sacramentals"

The word sacrament, in its broadest sense as a sign or symbol conveying something "hidden," mysterious, and efficacious, has a wider application and cosmic significance than that used in Christianity. For example, the evolutionary process is viewed by some as a graded series in which the lower stratum provides a basis for the one next above it. The lower, indeed, seems to be necessary to the growth of the higher. This view has introduced concepts of new powers and potentialities in organic evolution culminating in the human synthesis of mind transcending the process. The entire universe, therefore, can be said to have a sacramental significance in which the "inward" (or spiritual) and the "outward" (or material) elements meet in a higher unity that guarantees for the latter its full validity. Thus, the sacred meal has been at once a sacramental communion and a sacrificial offering (*e.g.,* wine, bread, or animal as a sign or symbol of a divine death and resurrection for the benefit of man) in which the two fundamental and complementary rites have been closely combined throughout their long and varied histories.

Variations. *Sacramental ideas and practices in preliterate societies.* In preliterate society everyday events have been given sacramental interpretations by being invested with supernatural meanings in relation to their ultimate sources in the unseen divine or sacred powers. The well-being of primitive society, in fact, demands the recognition of a hierarchy of values in which the lower is always dependent on the higher and in which the highest is regarded as the transcendental source of values outside and above mankind and the natural order. To partake of the flesh of a sacrificial victim or of the god himself or to consume the cereal image of a vegetation deity (as was done among the Aztecs in ancient Mexico), makes the eater a recipient of divine life and its qualities. Similarly, portions of the dead may be imbibed in mortuary sacramental rites to obtain the attributes of the deceased or to ensure their reincarnation. To give the dead new life beyond the grave, mourners may allow life-giving blood to fall upon the corpse sacramentally. In this cycle of sacramental ideas and practices, the giving, conservation, and promotion of life, together with the establishment of a bond of union with the sacred order, are fundamental. In Paleolithic hunting communities this sacramental idea appears to have been manifested in the sacramental rites performed to control the fortunes of the chase, to promote the propagation of the species on which the food supply depended, and to maintain right relations with the transcendental source of the means of subsistence, as exemplified in paintings—discovered in the caves at Altamira, Lascaux, Les Trois Frères, Font-de-Gaume and elsewhere in France and Spain—that show men with animal masks (illustrating a ritual or mystical communion of men and animals that were sources of food).

Mortuary sacramental rites

Sacramental ideas and practices in the ancient Near East. When agriculture and herding became the basic type of food production, sacramental concepts and techniques were centred mainly in the fertility of the soil, its products, and in the succession of the seasons. This centralization was most apparent in the ancient Near East in and after the 4th millennium BC. A death and resurrection sacred drama arose around the fertility motif, in which a perpetual dying and rebirth in nature and humanity was enacted. In this sequence birth, maturity, death, and rebirth were ritually repeated and renewed through sacramental tran-

Fertility motifs

sitional acts, such as passage rites, ceremonies ensuring passage from one status to another. In passage rites the king often was the principal actor in the promotion of the growth of the crops and the propagation of man and beast and in the promotion of the reproductive forces in nature in general at the turn of the year.

Sacramental ideas and practices in the Greco-Roman world. In the Greco-Oriental mystery cults the sacramental ritual based on the fertility motif was less prominent than in the Egyptian and Mesopotamian religions. It did, nevertheless, occur in the Eleusinia, a Greek agricultural festival celebrated in honour of the goddess Demeter and her daughter Kore. The things spoken and done in this great event have remained undisclosed, though some light has been thrown upon them by the contents of the museum at Eleusis, such as the vase paintings, and by later untrustworthy references in the writings of the early Church Fathers (*e.g.,* Clement of Alexandria) and some Gnostics (early Christian heretics who held that matter was evil and the spirit good). The drinking of the *kykeōn*—a gruel of meal and water—can hardly be regarded as a sacramental beverage since it was consumed during the preparation for the initiation rather than at its climax. There is nothing to suggest that a ritual rebirth was effected by a sacramental lustration, or sacred meal, at any point in the Eleusinian ritual. What is indicated is that the neophytes (mystae) emerged from their profound experience with an assurance of having attained newness of life and the hope of a blessed immortality. From the character of the ritual, the mystery would seem to have been connected with the seasonal drama in which originally a sacred marriage may have been an important feature, centred in Demeter, the corn mother, and Kore (Persephone), the corn maiden.

In the 6th century BC, or perhaps very much earlier, the orgiastic religion of the god Dionysus, probably originating in Thrace and Phrygia, was established in Greece. In the Dionysiac rites the Maenads (female attendents) became possessed by the spirit of Dionysus by means of tumultuous music and dancing, the free use of wine, and an orgiastic meal (the tearing to pieces and devouring of animals embodying Dionysus Zagreus with their bare hands as the central act of the Bacchanalia). Though not necessarily sacramental, these rites enabled the Maenads to surmount the barrier that separated them from the supernatural world and to surrender themselves unconditionally to the mighty powers that transcended time and space, thus carrying them into the realm of the eternal. Ecstatic rites of this nature did not commend themselves to the Greeks of the unemotional nonsacramental Homeric tradition; such rites did appeal, however, to many, some of whom had come under the influences of the Orphic mysteries in which it was possible for them to rise to a higher level in its *thiasoi* (brotherhoods). The purpose of the Orphic ritual was to confer divine life sacramentally on its initiates so that they might attain immortality through regeneration and reincarnation, thereby freeing the soul from its fleshly bondage.

Sacramental ideas and practices in the Indo-Iranian world. To what extent, if at all, metempsychosis (the passing of the soul at death into another body) was introduced into Greece from India can be only conjectural in the absence of conclusive evidence. Though belief in rebirth and the transmigration of souls has been widespread, however, especially in preliterate religions, it was in India and Greece that the two concepts attained their highest development. In post-Vedic (the period after the formulation of the Hindu sacred scriptures, the Veda) India, belief in the transmigration of souls became a characteristic doctrine in Hinduism, and the priestly caste (*i.e.,* the Brahmins) reached their zenith as the sole immolators of the sacrificial offerings; but sacramentalism was not a feature in the *Brāhmaṇas,* the ritual texts complied by the Brahmins. In the earlier Vedic conception of *soma,* the personification of the fermented juice of a plant, comparable to that of *ambros* in Greece, kava in Polynesia, and especially *haoma* in Iran, the sacramental view is most apparent (see HINDUISM).

Rebirth

In Zoroastrianism *haoma* (Sanskrit *soma,* from the root *su* or *bu,* "to squeeze" or "pound") is the name given to the yellow plant, from which a juice was extracted and consumed in the *Yasna* ceremony, the general sacrifice in honour of all the deities. The liturgy of the *Yasna* was a remarkable anticipation of the mass in Christianity. Haoma was regarded by Zoroaster as the son of the Wise Lord and Creator (Ahura Mazdā) and the chief priest of the *Yasna* cult. He was believed to be incarnate in the sacred plant that was pounded to death in order to extract its life-giving juice so that those who consumed it might be given immortality. He was regarded as both victim and priest in a sacrificial-sacramental offering in worship. As the intermediary between God and man, Haoma acquired a place and sacramental significance in the worship of Mithra (an Indo-Iranian god of light) in his capacity as the immaculate priest of Ahura Mazdā with whom he was coequal. The Mithraic sacramental banquet was derived from the *Yasna* ceremony, wine taking the place of the *haoma* and Mithra that of Ahura Mazdā. In the Mithraic initiation rites, it was not until one attained the status of the initiatory degree known as "Lion" that the neophyte could partake of the oblation of bread, wine, and water, which was the earthly counterpart of the celestial mystical sacramental banquet. The sacred wine gave vigour to the body, prosperity, wisdom, and the power to combat malignant spirits and to obtain immortality.

Incarnation of Haoma

The early Christian leaders noticed the resemblances between the Mithraic meal, the Zoroastrian *haoma* ceremony, and the Christian Eucharist; and between Mithraism and Christianity, to some extent, there was mutual influence and borrowing of respective beliefs and practices. But Mithraism's antecedents were different, being Iranian and Mesopotamian with a Vedic background before it become part of the Hellenistic and Christian world (*c.* 67 BC to about AD 385).

Sacramental ideas and practices of pre-Columbian America. The recurrent and widespread practice of holding sacred meals in the sacramental system, in addition to being well documented in the Greco-Roman world, also occurred in the pre-Columbian Mexican calendrical ritual in association with human sacrifice on a grand scale. In the May Festival in honour of the war god Huitzilopochtli, an image of the deity was fashioned from a dough containing beet seed, maize, and honey; then the image was covered with a rich garment, placed on a litter; and carried in a procession to a pyramid-temple. There pieces of paste similarly compounded and in the form of large bones were tranformed by rites of consecration into Huitzilopochtli's flesh and bones. A number of human victims were then offered to him, and the image was broken into small fragments and consumed sacramentally by the worshippers with tears, fears, and reverence, a strict fast being observed until the ceremonies were over and the sick had been given their communion with the particles. This ceremony was repeated at the winter solstice when the dough was fortified with the blood of children, and similar images were venerated and eaten by families in their houses. The main purpose of the sacrament was to secure a good maize harvest and a renewal of the crops, as well as human health and strength. In Peru at the Festival of the Sun, after three days of fasting, llamas, the sacred animals, were sacrificed as a burnt offering, and the flesh was eaten sacramentally at a banquet by the lord of the Incas and his nobles. It was then distributed to the rest of the community with sacred maize cakes. Dogs, regarded as divine incarnations, also were slain and parts of their flesh solemnly eaten by the worshippers.

Mexican sacramental rites

Similar rites were celebrated in North America by Indians at the Feast of Grain among the Natchez of Mississippi and Louisiana and among the Creeks in the Mississippi Valley when the corn was ripe. Among the Plains Indians sacrificial blood was employed sacramentally to make the earth fruitful by the fructifying power of the sun.

THEOLOGY AND PRACTICE OF SACRAMENTS IN CHRISTIANITY

Though the widespread conception of the sacramental principle is an ancient heritage, in all probability going back before the dawn of civilization, it acquired in Christianity a unique significance. There it became the fun-

Significance of the sacramental principle in Christianity

damental system and institution for the perpetuation of the union of God and man in the person of Jesus Christ through the visible organization and constitution of the church, which was viewed as the mystical body of Christ.

Baptism. Baptism, as the initial rite, took the place of circumcision in Judaism in which this ancient and primitive custom was the covenant sign and a legal injunction rather than a sacramental ordinance. Baptismal immersion in water was practiced in Judaism for some time before the fall of Jerusalem in AD 70, and it was adopted by John the Baptist (a Jewish prophet and cousin of Jesus Christ) as the principal sacrament in his messianic movement.

The purificatory lustration of John the Baptist, however, was transformed into the prototype of the Christian sacrament by the baptism of Jesus in the river Jordan and by the imagery of this event combined with the imagery of his death and resurrection. A distinction was made, however, between the water baptism of John and the Christian Spirit Baptism in the apostolic church. Under the influence of the missionary Apostle St. Paul, the Christian rite was given an interpretation in the terms of the mystery religions, and the catechumen (initiate instructed in the secrets of the faith) was identified with the death and Resurrection of Christ (Rom. 6:3–5; Gal. 3:12). The bestowal of the new life constituted a sacramental rebirth in the church in union with the risen Lord as its divine head.

Confirmation. With the development of infant Baptism, the regenerative initial sacrament was coupled with the charismatic apostolic laying on of hands as the seal of the Spirit in the rite of confirmation (Acts 8:14–17). By the 4th century, confirmation became a separate "unction" (rite using oil) administered by a bishop or, earlier and in the Eastern Church, by a priest to complete the sacramental baptismal grace already bestowed at birth or on some other previous occasion. At first, especially in the East, a threefold rite was performed consisting of Baptism, confirmation, and first communion; but in the West, where the consecration of the oil and the laying on of hands were confined to the episcopate, confirmation tended to become a separate event with the growth in the size of dioceses. It was not, however, until the 16th century that Baptism and confirmation were permanently separated. In England Queen Elizabeth I was confirmed when she was only three days old; and infant confirmation is still sometimes practiced in Spain. But the normal custom in Western Christendom has been for confirmation to be administered at or after the age of reason and to be the occasion for instruction in the faith, as in the case of the *mystae* in the Mysteries of Eleusis. But whether or not confirmation conveys a new gift of the Spirit or is the sealing of the same grace bestowed in Baptism, which is still debated, it has come to be regarded in some churches as conferring an indelible quality on the soul. Therefore, it cannot be repeated when it has once been validly performed as a sacrament.

The Eucharist. Together with Baptism the greatest importance has been given to the Eucharist, both of which institutions are singled out in the Gospels as dominical (instituted by Christ) in origin, with a special status and rank. Under a variety of titles (Eucharist from the Greek *eucharistia,* "thanksgiving"; the Latin mass; the Holy Communion; the Lord's Supper; and the breaking of the bread) it has been the central act of worship ever since the night of the betrayal of Christ on the Thursday preceding his crucifixion. It was then that the elements of bread and wine were identified with the body and blood of Christ in his institution of the Eucharist with his disciples and with the sacrifice he was about to offer in order to establish and seal the new covenant. This "real presence" has been variously interpreted in actual, figurative, or symbolical senses; but the sacramental sense, as the *anamnesis,* or memorial before God, of the sacrificial offering on the cross once and for all, has always been accepted.

The Eucharist as the central act of Christian worship

Along these lines a eucharistic theology gradually took shape in the apostolic and early church without much controversy or formulation. In the New Testament, in addition to the three accounts of the institution of the Eucharist in the first three "books" of the New Testament known as Synoptic Gospels because they have a common viewpoint and common sources (Matt. 26:26ff.; Mark 14:22ff.; Luke 22:17–20), St. Paul's earliest record of the ordinance in I Cor. 11:17–29, written about AD 55, suggests that some abuses had arisen in conjunction with the common meal, or *agapē,* with which it was combined. Like the Iranian *haoma* ceremony, it had become an occasion of drunkenness and gluttony. To rectify this, St. Paul recalled and re-established the original institution and its purpose and interpretation as a sacrificial-sacramental rite. Fellowship meals continued in association with the postapostolic Eucharist, as is shown in the *Didachē* (a Christian document concerned with worship and church discipline written *c.* 100–*c.* 140) and in the doctrinal and liturgical development described in the writings of the Early Church Fathers little was changed. Not until the beginning of the Middle Ages did controversial issues arise that found expression in the definition of the doctrine of transubstantiation at the fourth Lateran Council in 1215. This definition opened the way for the scholastic interpretation of the eucharistic Presence of Christ and of the sacramental principle, in Aristotelian terms. Thus, St. Thomas Aquinas maintained that a complete change occurred in the "substance" of each of the species, while the "accidents," or outward appearances, remained the same. During the Reformation, though the medieval doctrine was denied by the continental Reformers, it was reaffirmed by the Council of Trent in 1551. Holy Communion was retained as a sacrament by most of the Protestant groups, except the Society of Friends, the Salvation Army, and some of the Adventist groups, which abandoned the sacramental principle altogether.

Repentance. In its formulation, the Christian doctrine of conciliation, which, as St. Paul contended, required a change of status in the penitent, had to be made sacramentally effective in the individual and in redeemed humanity as a whole. In the Gospel According to Matthew (16:13–20, 18:18) the power to "bind and loose" was conferred on St. Peter and the other Apostles. Lapses into paganism and infidelity in the Roman world by the 3rd century had demanded penitential exercises. These included fasting, wearing sackcloth, lying in ashes and other forms of mortification, almsgiving, and the threat of temporary excommunication. Details of the sins committed were confessed in secret to a priest, who then pronounced absolution and imposed an appropriate penance. In 1215 the sacrament of penance received the authorization of the fourth Lateran Council and was made obligatory at least once a year at Easter on all mature Christians in Western Christendom. When pilgrimages to the Holy Land, to Santiago de Compostela in Spain, or going on a Crusade could be imposed as penitential exercises, commutation by means of payment of money led to abuses and traffic in indulgences and the treasury of merits, a superabundance of merits attributed to Christ and his saints that could be transferred to sinful believers. The abuses opened the way for the Lutheran revolt against the penitential system, before they were abolished by the Council of Trent. The power of absolution was retained in the Anglican ordinal and conferred upon priests at their ordination and in the Order of the Visitation of the Sick. The sacrament of penance, however, ceased to be of obligation in the Anglican Communion, though it was commended and practiced by John Whitgift, Richard Hooker, and, after the Restoration in 1660 by the Nonjurors (Anglican clergy who refused to take oaths of allegiance to William III and Mary II in 1689) and revived by the Tractarians (Anglo-Catholic advocates of High Church ideals) after 1833, who encountered some Protestant opposition notwithstanding its entrenchment in canon law and in *The Book of Common Prayer.*

Penitential exercises

Ordination. The church has claimed that the ministry of bishops, priests, and deacons derives its authority and sacramental efficacy from Christ through his Apostles. In the Roman Catholic Church it has been maintained that a special charismatic sacramental endowment conveying an indelible "character" has been conferred on those who receive valid ordination by the laying on of hands on their heads by bishops (who thus transfer to them the "power of orders"), prayer, and a right intention. In Protestant

Roman Catholic and Protestant views

churches the ministry is interpreted as a function rather than as a status. Just as the sacramental power to ordain, confirm, absolve, bless, and consecrate the Eucharist can be given, so also it can be taken away or suspended for sufficient reason.

Marriage. In the Roman Catholic Church the institution of holy matrimony was raised to the level of a sacrament because it was assigned a divine origin and made an indissoluble union typifying the union of Christ with his church as his mystical body (Matt. 5:27–32; Mark 10:2–12; Luke 16:18; I Cor. 7:2, 10; Eph. 5:23ff.). The adherence of Jesus to a rigorist position in regard to divorce and remarriage (Matt. 19:9; *i.e.,* unchastity the only cause for divorce), similar to that adopted by the rabbinical school headed by the conservative teacher Shammai in Judaism, was made the basis of the nuptial union as taught by St. Paul, except in regard to the dissolution of a marriage contracted between a Christian and a pagan who refused to live with his or her partner (I Cor. 7:2ff., 15ff.).

Apart from this deviation, known as the "Pauline Privilege," which was recognized in canon law in the 13th century, a marriage validly contracted in the presence of a priest, blessed by him, and duly consummated has been regarded as a sacramental ordinance by virtue of the grace given to render the union indissoluble. In Protestant churches, marriage is regarded as a rite, not a sacrament; views on divorce, however, vary, and many traditional notions of marriage and divorce are now being debated.

Healing of the sick. The anointing of the sick, an initiatory rite, conforms to the general pattern of the sacramental principle and is comparable to the other rites of passage, such as those concerned with birth and death, seed time and harvest, and with the securing of supernatural power and spiritual grace against the forces of evil that are looked upon as rampant at these critical junctures.

In Christianity anointing of the sick was widely practiced from apostolic times as a sacramental rite in association with the ceremony of the imposition of hands to convey a blessing, recovery from illness, or with the last communion to fortify the believer safely on his new career in the fuller life of the eternal world. Not until the 8th and 9th centuries, however, did extreme unction, another term for the final anointing of the sick, become one of the seven sacraments. In Eastern Christendom, it has never been confined to those *in extremis* (near death) nor has the blessing of the oil by a bishop been required; the administration of the sacrament by seven, five, or three priests was for the recovery of health rather than administered exclusively as a mortuary rite. Extreme unction is also coupled with exorcism for the restraint of the powers of evil—a practice taken over from Judaism by the early church and still retained by the Orthodox Eastern Church for mental diseases.

CONCLUSION

The ecumenical movement in the 20th century has initiated reforms in liturgical worship and in private devotions within Christianity. Such reforms, involving the celebration of sacraments (primarily the Eucharist), have done much to promote the recovery of a unity among Christians that transcends differences in beliefs and ritual practices. The second Vatican Council (1962–65) has played a significant part in the process of recovery of unity and of renewal. (E.O.J.)

Sacrifice

Sacrifice is a religious rite in which an object is offered to a divinity in order to establish, maintain, or restore a right relationship of man to the sacred order. It is a complex phenomenon that has been found in the earliest known forms of worship and in all parts of the world. The present article will treat the nature of sacrifice and will survey the theories about its origin. It will then analyze sacrifice in terms of its constituent elements, such as the material of the offering, the time and place of the sacrifice, and the motive or intention of the rite. Finally, it will briefly consider sacrifice in the religions of the world.

NATURE AND ORIGINS

Nature of sacrifice. The term sacrifice derives from the Latin *sacrificium,* which is a combination of the words *sacer,* meaning something set apart from the secular or profane for the use of supernatural powers, and *facere,* meaning "to make." The term has acquired a popular and frequently secular use to describe some sort of renunciation or giving up of something valuable in order that something more valuable might be obtained; *e.g.,* parents make sacrifices for their children, one sacrifices a limb for one's country. But the original use of the term was peculiarly religious, referring to a cultic act in which objects were set apart or consecrated and offered to a god or some other supernatural power; thus, sacrifice should be understood within a religious, cultic context.

Religion is man's relation to that which he regards as sacred or holy. This relationship may be conceived in a variety of forms. Although moral conduct, right belief, and participation in religious institutions are commonly constituent elements of the religious life, cult or worship is generally accepted as the most basic and universal element. Worship is man's reaction to his experience of the sacred power; it is a response in action, a giving of self, especially by devotion and service, to the transcendent reality upon which man feels himself dependent. Sacrifice and prayer—man's personal attempt to communicate with the transcendent reality in word or in thought—are the fundamental acts of worship.

In a sense, what is always offered in sacrifice is, in one form or another, life itself. Sacrifice is a celebration of life, a recognition of its divine and imperishable nature. In the sacrifice the consecrated life of an offering is liberated as a sacred potency that establishes a bond between the sacrificer and the sacred power. Through sacrifice, life is returned to its divine source, regenerating the power or life of that source; life is fed by life. Thus, the word of the Roman sacrificer to his god: "Be thou increased (*macte*) by this offering." It is, however, an increase of sacred power that is ultimately beneficial to the sacrificer. In a sense, sacrifice is the impetus and guarantee of the reciprocal flow of the divine life-force between its source and its manifestations.

Often the act of sacrifice involves the destruction of the offering, but this destruction—whether by burning, slaughter, or whatever means—is not in itself the sacrifice. The killing of an animal is the means by which its consecrated life is "liberated" and thus made available to the deity, and the destruction of a food offering in an altar's fire is the means by which the deity receives the offering. Sacrifice as such, however, is the total act of offering and not merely the method in which it is performed.

Although the fundamental meaning of sacrificial rites is that of effecting a necessary and efficacious relationship with the sacred power and of establishing man and his world in the sacred order, the rites have assumed a multitude of forms and intentions. The basic forms of sacrifice, however, seem to be some type of either sacrificial gift or sacramental meal. Sacrifice as a gift may refer either to a gift that should be followed by a return gift (because of the intimate relationship that gift giving establishes) or to a gift that is offered in homage to a god without expectation of a return. Sacrifice as a sacramental communal meal may involve the idea of the god as a participant in the meal or as identical with the food consumed; it may also involve the idea of a ritual meal at which either some primordial event such as creation is repeated or the sanctification of the world is symbolically renewed.

Theories of the origin of sacrifice. Since the rise of the comparative or historical study of religions in the latter part of the 19th century, attempts have been made to discover the origins of sacrifice. These attempts, though helpful for a greater understanding of sacrifice, have not been conclusive.

Theories of Tylor, Smith, and Frazer

In 1871 Sir Edward Burnett Tylor, a British anthropologist, proposed his theory that sacrifice was originally a gift to the gods to secure their favour or to minimize their hostility. In the course of time the primary motive for offering sacrificial gifts developed into homage, in which the sacrificer no longer expressed any hope for a return, and from

homage into abnegation and renunciation, in which the sacrificer more fully offered himself. Even though Tylor's gift theory entered into later interpretations of sacrifice, it left unexplained such phenomena as sacrificial offerings wholly or partly eaten by worshippers.

William Robertson Smith, a Scottish Semitic scholar and encyclopaedist, marked a new departure with his theory that the original motive of sacrifice was an effort toward communion among the members of a group, on the one hand, and between them and their god, on the other. Communion was brought about through a sacrificial meal. Smith began with totemism, according to which an animal or plant is intimately associated in a "blood relationship" with a social group or clan as its sacred ally. In general, the totem animal is taboo for the members of its clan, but on certain sacred occasions the animal is eaten in a sacramental meal that ensures the unity of the clan and totem and thus the well-being of the clan. For Smith an animal sacrifice was essentially a communion through the flesh and blood of the sacred animal, which he called the "theanthropic animal"—an intermediary in which the sacred and the profane realms were joined. The later forms of sacrifice retained some sacramental character: people commune with the god through sacrifice, and this communion occurs because the people share food and drink in which the god is immanent. From the communion sacrifice Smith derived the expiatory or propitiatory forms of sacrifice, which he termed *piaculum,* and the gift sacrifice. There were great difficulties with this theory: it made the totem a sacrificial victim rather than a supernatural ally; it postulated the universality of totemism; and, further, it did not adequately account for holocaust sacrifices in which the offering is consumed by fire and there is no communal eating. Nevertheless, many of Smith's ideas concerning sacrifice as sacramental communion have exerted tremendous influence.

Sir James George Frazer, a British anthropologist and folklorist, author of *The Golden Bough,* saw sacrifice as originating from magical practices in which the ritual slaying of a god was performed as a means of rejuvenating the god. The king or chief of a tribe was held to be sacred because he possessed mana, or sacred power, which assured the tribe's well-being. When he became old and weak, his mana weakened, and the tribe was in danger of decline. The king was thus slain and replaced with a vigorous successor. In this way the god was slain to save him from decay and to facilitate his rejuvenation. The old god appeared to carry away with him various weaknesses and fulfilled the role of an expiatory victim and scapegoat.

Theories of Hubert and Mauss, and van der Leeuw

Henri Hubert and Marcel Mauss, French sociologists, concentrated their investigations on Hindu and Hebrew sacrifice, arriving at the conclusion that "sacrifice is a religious act which, through the consecration of a victim, modifies the condition of the moral person who accomplishes it or that of certain objects with which he is concerned." Like Smith, they believed that a sacrifice establishes a relationship between the realms of the sacred and the profane. This occurs through the mediation of the ritually slain victim, which acts as a buffer between the two realms, and through participation in a sacred meal. The rituals chosen by Hubert and Mauss for analysis, however, are not those of preliterate societies.

Another study by Mauss helped to broaden the notion of sacrifice as gift. It was an old idea that man makes a gift to the god but expects a gift in return. The Latin formula *do ut des* ("I give that you may give") was formulated in classical times. In the Vedic religion, the oldest stratum of religion known to have existed in India, one of the *Brāhmaṇa*s (commentaries on the Vedas, or sacred hymns, that were used in ritual sacrifices) expressed the same principle: "Here is the butter; where are your gifts?" But, according to Mauss, in giving it is not merely an object that is passed on but a part of the giver, so that a firm bond is forged. The owner's mana is conveyed to the object, and, when the object is given away, the new owner shares in this mana and is in the power of the giver. The gift thus creates a bond. Even more, however, it makes power flow both ways to connect the giver and the receiver; it invites a gift in return.

Gerardus van der Leeuw, a Dutch historian of religion, developed this notion of gift in the context of sacrifice. In sacrifice a gift is given to the god, and thus man releases a flow between himself and the god. For him sacrifice as gift is "no longer a mere matter of bartering with gods corresponding to that carried on with men, and no longer homage to the god such as is offered to princes: it is an opening of a blessed source of gifts." His interpretation thus melded the gift and communion theories, but it also involved a magical flavour, for he asserted that the central power of the sacrificial act is neither god nor giver but is always the gift itself.

Theories of Jensen and Freud

German anthropologists have emphasized the idea of culture history, in which the entire history of mankind is seen as a system of coherent and articulated phases and strata, with certain cultural phenomena appearing at specific levels of culture. Leo Frobenius, the originator of the theory that later became known as the *Kulturkreislehre,* distinguished the creative or expressive phase of a culture, in which a new insight assumes its specific form, and the phase of application, in which the original significance of the new insight degenerates. Working within this context, Adolf E. Jensen attempted to explain why men have resorted to the incomprehensible act of killing other men or animals and eating them for the glorification of a god or many gods. Blood sacrifice is linked not with the cultures of the hunter–gatherers but with those of the cultivators; its origin is in the ritual killing of the archaic cultivator cultures, which, in turn, is grounded in myth. For Jensen the early cultivators all knew the idea of a mythic primal past in which not men but Dema lived on the Earth and prominent among them were the Dema-deities. The central element of the myth is the slaying of a Dema-deity, an event that inaugurated human history and gave shape to the human lot. The Dema became men, subject to birth and death, whose self-preservation depends upon the destruction of life. The deity became in some way associated with the realm of the dead; and, from the body of the slain deity, crop plants originated, so that the eating of the plants is an eating of the deity. Ritual killing, whether of animals or men, is a cultic re-enactment of the mythological event. Strictly speaking, the action is not a sacrifice because there is no offering to a god; rather, it is a way to keep alive the memory of primeval events. Blood sacrifice as found in the later higher cultures is a persistence of the ritual killing in a degenerated form. Because the victim is identified with the deity, later expiatory sacrifices also become intelligible: sin is an offense against the moral order established at the beginning of human history; the killing of the victim is an intensified act restoring that order.

Another interpretation of some historical interest is that of Sigmund Freud in his work *Totem und Tabu* (1913; Eng. trans., *Totem and Taboo,* 1918). Freud's theory was based on the assumption that the Oedipus complex is innate and universal. It is normal for a child to wish to have a sexual relationship with its mother and to will the death of its father; this is often achieved symbolically. In the primal horde, although the sons did slay their father, they never consummated a sexual union with their mother; in fact, they set up specific taboos against such sexual relations. According to Freud, the ritual slaughter of an animal was instituted to re-enact the primeval act of parricide. The rite, however, reflected an ambivalent attitude. After the primal father had been slain, the sons felt some remorse for their act, and, thus, the sacrificial ritual expressed the desire not only for the death of the father but also for reconciliation and communion with him through the substitute victim. Freud claimed that his reconstruction of the rise of sacrifice was historical, but this hardly seems probable.

Mid-20th-century theories

In 1963 Raymond Firth, a New Zealand-born anthropologist, addressed himself to the question of the influence that a people's ideas about the control of their economic resources have on their ideology of sacrifice. He noted that the time and frequency of sacrifice and the type and quality of victim are affected by economic considerations; that the procedure of collective sacrifice involves not only the symbol of group unity but also a lightening of the economic burden or any one participant; that the use of

surrogate victims and the reservation of the sacrificial food for consumption are possibly ways of meeting the problem of resources. Firth concluded that sacrifice is ultimately a personal act in which the self is symbolically given, but it is an act that is often conditioned by economic rationality and prudent calculation.

Most social anthropologists and historians of religion in the mid-20th century, however, concentrated less on worldwide typologies or evolutionary sequences and more on investigations of specific historically related societies. Consequently, since World War II there have been few formulations of general theories about the origin of sacrifice, but there have been important studies of sacrifice within particular cultures. For example, E.E. Evans-Pritchard, a social anthropologist at Oxford University, concluded after his study of the religion of the Nuer, a people in the southern Sudan, that for them sacrifice is a gift intended "to get rid of some danger of misfortune, usually sickness." They establish communication with the god not to create a fellowship with him but only to keep him away. Evans-Pritchard acknowledged, however, that the Nuer have many kinds of sacrifice and that no single formula adequately explains all types. Furthermore, he did not maintain that his interpretations of his materials were of universal applicability. Many scholars would agree that, though it is easy to make a long list of many kinds of sacrifice, it is difficult, if not impossible, to find a satisfactory system in which all forms of sacrifice may be assigned a suitable place.

ANALYSIS OF THE RITE OF SACRIFICE

It is possible to analyze the rite of sacrifice in terms of six different elements: the sacrificer, the material of the offering, the time and place of the rite, the method of sacrificing, the recipient of the sacrifice, and the motive or intention of the rite. These categories are not of equal importance and often overlap.

Sacrificer. In general, it may be said that the one who makes sacrifices is man, either an individual or a collective group—a family, a clan, a tribe, a nation, a secret society. Frequently, special acts must be performed by the sacrificer before and sometimes also after the sacrifice. In the Vedic cult, the sacrificer and his wife were required to undergo an initiation (*dīkṣā*) involving ritual bathing, seclusion, fasting, and prayer, the purpose of which was to remove them from the profane world and to purify them for contact with the sacred world. At the termination of the sacrifice came a rite of "desacralization" (*avabhṛta*) in which they bathed in order to remove any sacred potencies that might have attached themselves during the sacrifice.

There are sacrifices in which there are no participants other than the individual or collective sacrificer. Usually, however, one does not venture to approach sacred things directly and alone; they are too lofty and serious a matter.

The role of the intermediary

An intermediary—certain persons or groups who fulfill particular requirements or qualifications—is necessary. In many cases, sacrificing by unauthorized persons is expressly forbidden and may be severely punished; *e.g.*, in the book of Leviticus, Korah and his followers, who revolted against Moses and his brother Aaron and arrogated the priestly office of offering incense, were consumed by fire. The qualified person—whether the head of a household, the old man of a tribe, the king, or the priest—acts as the appointed representative on behalf of a community.

The head of the household as sacrificer is a familiar figure in the Old Testament, particularly in the stories of the patriarchs; *e.g.*, Abraham and Jacob. Generally, in cattle-keeping tribes with patriarchal organization, the *paterfamilias* long remained the person who carried out sacrifices, and it was only at a late date that a separate caste of priests developed among these peoples. In ancient China, too, sacrifices were not presided over by a professional priesthood but by the head of the family or, in the case of state sacrifices, by the ruler.

The old man or the elders of the tribe are in charge of sacrifices among several African peoples. Among the Ila, a people of Zambia, for instance, when hunters have no success, the oldest member of the band leads the others in praying for the god's aid; when the hunters are successful in killing, the old man leads them in offering portions of the meat to the god. Similarly, among peoples in Australia the leading role in all sacrificial acts is filled by the old men as bearers of tradition and authority. In cases in which there is a matriarchal organization, as in some parts of West Africa, the oldest woman of the family acts as priestess.

The king has played an important role as the person active in sacrificing, particularly in those cultures in which he not only has temporal authority but also fulfills a religious function. The fact that the king is the primary sacrificer may stem from two roots. It may be that the most important gods of the state were originally family gods of the rulers, and, thus, the king is simply continuing the task of *paterfamilias*, only now on behalf of the whole community. The second root lies in the notion of sacred kingship, according to which the royal office is sacred and the king set apart from ordinary people is the intercessor with the supernatural world. These two concepts often go together. Thus, in ancient Egypt the pharaoh was divine because he descended from the sun god Re. The pharaoh stood for Horus, the son of Re. The concepts of the god as family ancestor and of sacred kingship were combined. Although worship in ancient Egypt was controlled by a powerful priesthood, officially all sacrifices were regarded as made by the pharaoh.

Most frequently, the intermediary between the community and the god, between the profane and the sacred realms, is the priest. As a rule, not everyone can become a priest; there are requirements of different kinds to be satisfied. Usually, the priest must follow some training, which may be long and severe, There is always some form of consecration he has to undergo. For communities in which a priest functions, he is the obvious person to make sacrifices.

Gods as sacrificers

The sacrificer is not always man, however; at times gods also make sacrifices. Examples of this are found chiefly in India and are set down particularly in the *Brāhmaṇa* texts; *e.g.*, it is said in the *Taittirīya Brāhmaṇa:* "By sacrifice the gods obtained heaven." The idea of gods making sacrifice, however, is found in the older Ṛgveda-Saṃhitā, a collection of sacred Vedic hymns: "With offerings the gods offered up sacrifice." In this conception man makes sacrifices in imitation of a divine model inaugurated by the gods themselves. Another instance is the Iranian primordial god Zurvān (Time), who offered sacrifice for 1,000 years in order to obtain a son to create the world.

Material of the oblation. Any form under which life manifests itself in the world or in which life can be symbolized may be a sacrificial oblation. In fact, there are few things that have not, at some time or in some place, served as an offering. Any attempt to categorize the material of sacrifice will group together heterogeneous phenomena; thus, the category human sacrifice includes several fundamentally different sacrificial rites. Nevertheless, for convenience sake, the variety of sacrificial offerings will be treated as (1) blood offerings (animal and human), (2) bloodless offerings (libations and vegetation), and (3) a special category, divine offerings.

Blood offerings. Basic to both animal and human sacrifice is the recognition of blood as the sacred life-force in man and beast. Through the sacrifice—through the return of the sacred life revealed in the victim—the god lives, and, therefore, man and nature live. The great potency of blood has been utilized through sacrifice for a number of purposes; *e.g.*, earth fertility, purification, and expiation. The letting of blood, however, was neither the only end nor the only mode of human and animal sacrifice.

A wide variety of animals have served as sacrificial offerings. In ancient Greece and India, for example, oblations included a number of important domestic animals, such as the goat, ram, bull, ox, and horse. Moreover, in Greek religion all edible birds, wild animals of the hunt, and fish were used. In ancient Judaism the kind and number of animals for the various sacrifices was carefully stipulated so that the offering might be acceptable and thus fully effective. This sort of regulation is generally found in sacrificial cults; the offering must be appropriate either to the deity to whom or to the intention for which it is to

be presented. Very often the sacrificial species (animal or vegetable) was closely associated with the deity to whom it was offered as the deity's symbolic representation or even its incarnation. Thus, in the Vedic ritual the goddesses of night and morning received the milk of a black cow having a white calf; the "bull of heaven," Indra, was offered a bull, and Sūrya, the sun god, a white, male goat. Similarly, the ancient Greeks sacrificed black animals to the deities of the dark underworld; swift horses to the sun god Helios; pregnant sows to the earth mother Demeter; and the dog, guardian of the dead, to Hecate, goddess of darkness. The Syrians sacrificed fish, regarded as the lord of the sea and guardian of the realm of the dead, to the goddess Atargatis and ate the consecrated offering in a communion meal with the deity, sharing in the divine power. An especially prominent sacrificial animal was the bull (or its counterparts, the boar and the ram), which, as the representation and embodiment of the cosmic powers of fertility, was sacrificed to numerous fertility gods (*e.g.*, the Norse god Freyr; the Greek "bull of the Earth," Zeus Chthonios; and the Indian "bull of heaven," Indra).

Human sacrifice

The occurrence of human sacrifice appears to have been widespread and its intentions various, ranging from communion with a god and participation in his divine life to expiation and the promotion of the earth's fertility. It seems to have been adopted by agricultural rather than by hunting or pastoral peoples. Of all the worldly manifestations of the life-force, the human undoubtedly impressed men as the most valuable and thus the most potent and efficacious as an oblation. Thus, in Mexico the belief that the sun needed human nourishment led to sacrifices in which as many as 20,000 victims perished annually in the Aztec and Nahua calendrical maize ritual in the 14th century AD. Bloodless human sacrifices also developed and assumed greatly different forms: *e.g.*, a Celtic ritual involved the sacrifice of a woman by immersion, and

By courtesy of the National Museum, Copenhagen



Celtic sacrifice by immersion, detail of the Gundestrup Caldron, c. 1st century BC. In the Nationalmuseet, Copenhagen.

among the Maya in Mexico young maidens were drowned in sacred wells; in Peru women were strangled; in ancient China the king's retinue was commonly buried with him, and such internments continued intermittently until the 17th century.

In many societies human victims gave place to animal substitutes or to effigies made of dough, wood, or other materials. Thus, in India, with the advent of British rule, human sacrifices to the Dravidian village goddesses (*grāma-devīs*) were replaced by animal sacrifices. In Tibet, under the influence of Buddhism, which prohibits all blood sacrifice, human sacrifice to the pre-Buddhist Bon deities was replaced by the offering of dough images or reduced to pantomime. Moreover, in some cults both human and animal oblations could be "ransomed"—*i.e.*, replaced by offerings or money or other inanimate valuables.

Bloodless offerings. Among the many life-giving substances that have been used as libations are milk, honey, vegetable and animal oils, beer, wine, and water. Of these, the last two have been especially prominent. Wine is the "blood of the grape" and thus the "blood of the earth," a spiritual beverage that invigorates gods and men. Water is always the sacred "water of life," the primordial source of existence and the bearer of the life of plants, animals, human beings, and even the gods. Because of its great potency, water, like blood, has been widely used in purificatory and expiatory rites to wash away defilements and restore spiritual life. It has also, along with wine, been an important offering to the dead as a revivifying force.

Vegetable offerings have included not only the edible herbaceous plants but also grains, fruits, and flowers. In both Hinduism and Jainism, flowers, fruits, and grains (cooked and uncooked) are included in the daily temple offerings. In some agricultural societies (*e.g.*, those of West Africa) yams and other tuber plants have been important in planting and harvest sacrifices and in other rites concerned with the fertility and fecundity of the soil. These plants have been regarded as especially embodying the life-force of the deified earth and are frequently buried or plowed into the soil to replenish and reactivate its energies.

Divine offerings. One further conception must be briefly mentioned: a god himself may be sacrificed. This notion was elaborated in many mythologies; it is fundamental in some sacrificial rituals. In early sacrifice the victim has something of the god in itself, but in the sacrifice of a god the victim is identified with the god. At the festival of the ancient Mexican sun god Huitzilopochtli, the statue of the god, which was made from beetroot paste and kneaded in human blood and which was identified with the god, was divided into pieces, shared out among the devotees, and eaten. In the Hindu *soma* ritual (related to the *haoma* ritual of ancient Persia), the soma plant, which is identified with the god Soma, is pressed for its intoxicating juice, which is then ritually consumed. The Eucharist, as understood in many of the Christian churches, contains similar elements. In short, Jesus is really present in the bread and wine that are ritually offered and then consumed. According to the traditional eucharistic doctrine of Roman Catholicism, the elements of bread and wine are "transubstantiated" into the body and blood of Christ: *i.e.*, their whole substance is converted into the whole substance of the body and blood, although the outward appearances of the elements, their "accidents," remain.

Gods as victims

Time and place of sacrifice. In many cults, sacrifices are distinguished by frequency of performance into two types, regular and special. Regular sacrifices may be daily, weekly, monthly, or seasonal (as at planting, harvest, and New Year). Also often included are sacrifices made at specific times in each man's life—birth, puberty, marriage, and death. Offerings made on special occasions and for special intentions have included, for example, sacrifices in times of danger, sickness, or crop failure and those performed at the construction of a building, for success in battle, or in thanksgiving for a divine favour.

In the Vedic cult the regular sacrifices were daily, monthly, and seasonal. The daily rites included fire offerings to the gods and libations and food offerings to the ancestors and the earth divinities and spirits. The monthly sacrifices, conducted at the time of New and Full Moons, were of cakes or cooked oblations to sundry deities, especially the storm god Indra. Some daily and monthly sacrifices could be celebrated in the home by a householder, but only the official priesthood could perform the complex seasonal sacrifices, offered three times a year—at the beginning of spring, of the rainy season, and of the cool weather—for the purpose of expiation and of abundance. Of the occasional sacrifices, which could be celebrated at any time, especially important were those associated with kingship, such as the royal consecration and the great "horse sacrifice" performed for the increase of the king's power and domain.

In ancient Judaism the regular or periodic sacrifices included the twice daily burnt offerings, the weekly sabbath sacrifices, the monthly offering at the New Moon, and annual celebrations such as Pesaḥ (Passover), Yom Kippur (Day of Atonement), and Sukkot (Feast of Tabernacles). Special sacrifices were usually of a personal nature, such as thank and votive offerings and "guilt offerings."

The common place of sacrifice in most cults is an altar. The table type of altar is uncommon; more often it is only a pillar, a mound of earth, a stone, or a pile of stones. Among the Hebrews in early times and other Semitic peoples the altar of the god was frequently an upright stone

(*matzzeva*) established at a place in which the deity had manifested itself. It was *bet el,* the "house of God."

Frequently, the altar is regarded as the centre or the image of the universe. For the ancient Greeks the grave marker (a mound of earth or a stone) was the earth altar upon which sacrifices to the dead were made and, like other earth altars, it was called the omphalos, "the navel" of the Earth—*i.e.,* the central point from which terrestrial life originated. In Vedic India the altar was regarded as a microcosm, its parts representing the various parts of the universe and its construction being interpreted as a repetition of the creation of the cosmos.

Burning, burying, and ritual killing

Method of sacrifice. Along with libation and the sacrificial effusion of blood, one of the commonest means of making an oblation available to sacred beings is to burn it. In both ancient Judaism and Greek religion the major offering was the burnt or fire offering. Through the medium of the fire, the oblation was conveyed to the divine recipient. In ancient Greece the generic term for sacrifice (*thysia*) was derived from a root meaning to burn or to smoke. In Judaism the important sacrifices (*ʿola* and *zevaḥ*) involved the ritual burning, either entirely or in part, of the oblation, be it animal or vegetation. For the Babylonians, also, fire was essential to sacrifice, and all oblations were conveyed to the gods by the fire god Girru-Nusku, whose presence as intermediary between the gods and men was indispensable. In the Vedic cult the god of fire, Agni, received the offerings of men and brought them into the presence of the gods.

As burning is often the appropriate mode for sacrifice to celestial deities, so burial is often the appropriate mode for sacrifice of earth deities. In Greece, for example, sacrifices to the chthonic or underworld powers were frequently buried rather than burned or, if burned, burned near the ground or even in a trench. In Vedic India the blood and entrails of animals sacrificed on the fire altar to the sky gods were put upon the ground for the earth deities, including the ghosts and malevolent spirits. In West Africa yams and fowls sacrificed to promote the fertility of the earth are planted in the soil.

In sacrifice by burning and by burial, as also in the effusion of blood, the prior death of the human or animal victim, even if ritually performed, is in a sense incidental to the sacrificial action. There are, however, sacrifices (including live burial and burning) in which the ritual killing is itself the means by which the offering is effected. Illustrative of this method was the practice in ancient Greek and Indian cults of making sacrifices to water gods by drowning the oblations in sacred lakes or rivers. Similarly, the Norse cast human and animal victims over cliffs and into wells and waterfalls as offerings to the divinities dwelling therein. In the Aztec sacrifice of human beings to the creator god Xipe Totec, the victim was lashed to a scaffold and shot to death with bow and arrow.

"Fireless oblations"

There are also sacrifices that do not involve the death or destruction of the oblation. Such were the sacrifices in ancient Greece of fruits and vegetables at the "pure" (*katharos*) altar of Apollo at Delos, at the shrine of Athena at Lindus, and at the altar of Zeus in Athens. These "fireless oblations" (*apura hiera*) were especially appropriate for the deities of vegetation and fertility; *e.g.,* Demeter and Dionysus. In Egypt bloodless offerings of food and drink were simply laid before the god on mats or a table in a daily ceremony called "performing the presentation of the divine oblations." In both Greek and Egyptian cults such offerings were never to be eaten by the worshippers, but they were probably surreptitiously consumed by the priests or temple attendants. In ancient Israel, on the other hand, the food offerings of the "table of the shewbread" (the "bread of the presence" of God) were regarded as available to the priests and could be given by them to the laity. In Hinduism the daily offering of cooked rice and vegetable, after its consecration, is distributed by the priests to the worshippers as the deity's "grace" (*prasāda*). In some cases the sacrificial gifts are put out to be eaten by an animal representative of the deity. In Dahomey wandering dogs consume, on behalf of the trickster deity Eshu (Elegba), the consecrated food oblations presented to the god each morning at his shrines.

Recipient of the sacrifice. Sacrifices may be offered to beings who can be the object of religious veneration or worship. They will not be made to human beings unless they have first been deified in some way. In some cases sacrifice is made only to the god or gods; in others it is made to the deity, the spirits, and the departed; in others it is made only to the spirits and the departed, who are considered intermediaries between the deity and men. The Nkole people of Uganda, for example, are said to make no sacrifices to God, thinking he does not expect any. But, on the third day following the New Moon, they make offerings to the guardian spirits (*emandwa*), and they also make offerings at the shrines of ancestors (*emizimu*) of up to three generations back. Worship of spirits and of ancestors, often including the offering of sacrifices, occurs in widely distributed cultures; in fact, according to some scholars, probably the major recipients of sacrifice in non-Western traditions are the ancestors.

Intentions. Sacrifices have been offered for a multiplicity of intentions, and it is possible to list only some of the most prominent. In any one sacrificial rite a number of intentions may be expressed, and the ultimate goal of all sacrifice is to establish a beneficial relationship with the sacred order, to make the sacred power present and efficacious.

Propitiation and expiation. Serious illness, drought, pestilence, epidemic, famine, and other misfortune and calamity have universally been regarded as the workings of supernatural forces. Often they have been understood as the effects of offenses against the sacred order committed by individuals or communities, deliberately or unintentionally. Such offenses break the relationship with the sacred order or impede the flow of divine life. Thus, it has been considered necessary in times of crisis, individual or communal, to offer sacrifices to propitiate sacred powers and to wipe out offenses (or at least neutralize their effects) and restore the relationship.

Among the Yoruba of West Africa, blood sacrifice must be made to the gods, especially the earth deities, who, as elsewhere in Africa, are regarded as the divine punishers of sin. For the individual the oblation may be a fowl or a goat; for an entire community it may be hundreds of animals (in former days, the principal oblation was human). Once consecrated and ritually slain, the oblations are buried, burnt, or left exposed but never shared by the sacrificer.

Atonement in Judaism and Christianity

In ancient Judaism the *ḥaṭṭaʾt,* or "sin offering," was an important ritual for the expiation of certain, especially unwittingly committed, defilements. The guilty laid their hands upon the head of the sacrificial animal (an unblemished bullock or goat), thereby identifying themselves with the victim, making it their representative (but not their substitute, for their sins were not transferred to the victim). After the priest killed the beast, blood was sprinkled upon the altar and elsewhere in the sacred precincts. The point of the ritual was to purify the guilty and to re-establish the holy bond with God through the blood of the consecrated victim. It was as such an expiatory sacrifice that early Christianity regarded the life and death of Christ. By the shedding of his blood, the sin of mankind was wiped out and a new relationship of life—eternal life—was effected between God and man. Like the innocent and "spotless" victim of the *ḥaṭṭaʾt,* Christ died for men—*i.e.,* on behalf of but not in place of them. Also, like the *ḥaṭṭaʾt,* the point of his death was not the appeasement of divine wrath but the shedding of his blood for the wiping out of sin. The major differences between the sacrifice of Christ and that of the *ḥaṭṭaʾt* animal are that (1) Christ's was regarded as a voluntary and effective sacrifice for all men and (2) his was considered the perfect sacrifice, made once in time and space but perpetuated in eternity by the risen Lord.

There are sacrifices, however, in which the victim does serve as a substitute for the guilty. In some West African cults a person believed to be under death penalty by the gods offers an animal substitute to which he transfers his sins. The animal, which is then ritually killed, is buried with complete funeral rites as though it were the human person. Thus the guilty person is dead, and it is an innocent man who is free to begin a new life.

Finally, some propitiatory sacrifices are clearly prophylactic, intended to avert possible misfortune and calamity, and as such they are really bribes offered to the gods. Thus, in Dahomey libations and animal and food offerings are frequently made to a variety of Earth spirits to ensure their good favour in preventing any adversity from befalling the one making the offering.

Gift sacrifices. Although all sacrifice involves the giving of something, there are some sacrificial rites in which the oblation is regarded as a gift made to a deity either in expectation of a return gift or as the result of a promise upon the fulfillment of a requested divine favour. Gift sacrifices have been treated above. Here, it can be briefly noted that numerous instances of the votive offering are recorded. In ancient Greece sacrifices were vowed to Athena, Zeus, Artemis, and other gods in return for victory in battle. The solemnity and irrevocability of the votive offering is seen in the Old Testament account of the judge Jephthah's sacrifice of his only child in fulfillment of a vow to Yahweh.

Thank offerings. One form of thank offering is the offering of the first fruits in agricultural societies. Until the first fruits of the harvest have been presented with homage and thanks (and often with animal sacrifices) to the deity of the harvest (sometimes regarded as embodied in the crop), the whole crop is considered sacred and thus taboo and may not be used as food. The first-fruits sacrifice has the effect of "desacralizing" the crops and making them available for profane consumption. It is a recognition of the divine source and ownership of the harvest and the means by which man is reconciled with the vegetational, chthonic powers from whom he takes it.

Offerings of the first fruits of harvest

Fertility. Another distinctive feature of the first-fruits offering is that it serves to replenish the sacred potencies of the earth depleted by the harvest and to ensure thereby the continued regeneration of the crop. Thus, it is one of many sacrificial rites that have as their intention the seasonal renewal and reactivation of the fertility of the earth. Fertility rites usually involve some form of blood sacrifice—in former days especially human sacrifice. In some human sacrifices the victim represented a deity who "in the beginning" allowed himself to be killed so that from his body edible vegetation might grow. The ritual slaying of the human victim amounted to a repetition of the primordial act of creation and thus a renewal of vegetational life. In other human sacrifices the victim was regarded as representing a vegetation spirit that annually died at harvest time so that it might be reborn in a new crop. In still other sacrifices at planting time or in time of famine, the blood of the victim—animal or human—was let upon the ground and its flesh buried in the soil to fertilize the earth and recharge its potencies.

Building sacrifices. Numerous instances are known of animal and human sacrifices made in the course of the construction of houses, shrines, and other buildings, and in the laying out of villages and towns. Their purpose has been to consecrate the ground by establishing the beneficent presence of the sacred order and by repelling or rendering harmless the demonical powers of the place. In some West African cults, for example, before the central pole of a shrine or a house is installed, an animal is ritually slain, its blood being poured around the foundations and its body being put into the posthole. On the one hand, this sacrifice is made to the earth deities and the supernatural powers of the place—the real owners—so that the human owner may take possession and be ensured against malevolent interferences with the construction of the building and its later occupation and use. On the other hand, the sacrifice is offered to the cult deity to establish its benevolent presence in the building.

Mortuary sacrifice. Throughout the history of man's religions, the dead have been the recipients of offerings from the living. In ancient Greece an entire group of offerings (*enagismata*) was consecrated to the dead; these were libations of milk, honey, water, wine, and oil poured onto the grave. In India water and balls of cooked rice were sacrificed to the spirits of the departed. In West Africa, offerings of cooked grain, yams, and animals are made to the ancestors residing in the Earth. The point of such offerings is not that the dead get hungry and thirsty, nor are they merely propitiatory offerings. Their fundamental intention seems to be that of increasing the power of life of the departed. The dead partake of the life of the gods (usually the chthonic deities), and sacrifices to the dead are in effect sacrifices to the gods who bestow never-ending life. In Hittite funeral rites, for example, sacrifices were made to the sun god and other celestial deities—transcendent sources of life—as well as to the divinities of Earth.

Communion sacrifices. Communion in the sense of a bond between the worshipper and the sacred power is fundamental to all sacrifice. Certain sacrifices, however, promote this communion by means of a sacramental meal. The meal may be one in which the sacrificial oblation is simply shared by the deity and the worshippers. Of this sort were the Greek *thysia* and the Jewish *zevaḥ* sacrifices in which one portion of the oblation was burned upon the altar and the remainder eaten by the worshippers. Among the African Yoruba special meals are offered to the deity; if the deity accepts the oblation (as divination will disclose), a portion of the food is placed before his shrine while the remainder is joyfully eaten as a sacred communion by the worshippers. The communion sacrifice may be one in which the deity somehow indwells the oblation so that the worshippers actually consume the divine; *e.g.*, the Hindu *soma* ritual. The Aztecs twice yearly made dough images of the sun god Huitzilopochtli that were consecrated to the god and thereby transubstantiated into his flesh to be eaten with fear and reverence by the worshippers.

The sacramental meal

SACRIFICE IN THE RELIGIONS OF THE WORLD

The constituent elements of sacrifice have been incorporated into the particular religions and cultures of the world in various and often complex ways. A few brief observations that may illustrate this variety and complexity are given here.

Religions of India. Speculations regarding sacrifice and prescribed rituals seem to have been worked out more fully in the Vedic and later Hindu religion in India than anywhere else. These rites, laid down in a complicated system known mainly from the *Brāhmaṇa* texts, included obligatory sacrifices following the course of the year or the important moments in the life of an individual and optional sacrifices occasioned by the special wishes of a sacrificer. Yet cultic sacrifice has not developed in Buddhism, another religion that arose in India. Ritual sacrifice was judged to be ineffective and in some of its forms to involve cruelty and to run counter to the law of *ahiṃsā*, or non-injury. There are, however, in the *Jātaka* stories of the Buddha's previous births accounts of his self-sacrifices. Furthermore, Buddhism emphasizes the notion of ethical sacrifices, acts of self-discipline; and there are instances of devotional offerings, such as burnt incense, to the Buddha.

Religions of China. In China sacrifice, like other aspects of religion, has existed at a number of different levels. The essential feature of Imperial worship in ancient China was the elaborate sacrifices offered by the emperor himself to Heaven and Earth. There are also records of sacrifice, including human sacrifice, associated with the death of a ruler because it was thought proper for him to be accompanied in death with those who served him during life. But, because the common people were excluded from participation in Imperial sacrifices, they had lesser gods—some universal, some local—to whom sacrifices were made. Furthermore, ancestor worship has been the most universal form of religion throughout China's long history; it was the responsibility of the head of a household to see to it that sacrificial offerings to the dead were renewed constantly. The blending of these elements with such established religions as Buddhism and Taoism influenced the great diversification of sacrificial rites in China.

Religions of Japan. In ancient Japan offering occupied a particularly important place in religion because the relationship of the people to their gods seems frequently to have had the character of a bargain rather than of adoration. It is probable that the offerings were originally individual, but they gradually became collective, especially as all powers, including religious, were concentrated in the hands of the emperor, who officiated in the name of all his people. Human sacrifice to natural deities and at burials



The Eucharist as sacrifice: "La Ultima Cena," oil painting by Juan de Juanes (c. 1523–79). In the Prado, Madrid.
Archivo Mas, Barcelona

was once common but seems generally to have been abandoned in the early Middle Ages. Besides human sacrifices and their more modern substitutes, the Japanese offered to the gods all the things that man regards as necessary (*e.g.,* food, clothing, shelter) or merely useful and pleasing (*e.g.,* means of transportation, tools, weapons, objects of entertainment) for life. These practices, which were found in the traditional religion known as Shintō, were modified when Confucianism and Buddhism were introduced into Japan during the 5th and 6th centuries AD.

Thysia and *sphagia* rites

Ancient Greece. The Homeric poems contain the most complete descriptions of sacrificial rites in ancient Greece. These rites, which were maintained almost without change for more than 10 centuries, were of two types: rites (*thysia*) addressed to the Olympian deities, which included burning part of a victim and then participating in a joyful meal offered to the gods during the daytime primarily to serve and establish communion with the gods; and rites (*sphagia*) addressed to the infernal or chthonic deities, which involved the total burning or burying of a victim in a sombre nocturnal ceremony to placate or avert the malevolent chthonic powers. Besides the official or quasi-official rites, the popular religion, already in Homer, comprised sacrifices of all kinds of animals and of vegetables, fruits, cheese, and honey offered as expiation, supplication, or thanksgiving by worshippers belonging to all classes of society. Furthermore, the secret worship of what are known as the mysteries—cults normally promising immortality or some form of personal relationship with a god—became widespread. This practice became especially prominent during the Hellenistic period.

Judaism. The destruction of the Second Temple in AD 70 marked a profound change in the worship of the Jewish people. Before that event, sacrifice was the central act of Israelite worship; and there were many categories of sacrificial rites that had evolved through the history of the Jews into a minutely detailed system found in that part of the Torah (Law; the first five books of the Hebrew Bible) that is ascribed by biblical scholars to the Priestly Code, which became established following the Babylonian Exile (586–538). The sacrificial system ceased, however, with the destruction of the Temple, and prayer took the place of sacrifices. In modern Judaism the Orthodox prayer books still contain prayers for the reinstitution of the sacrificial cult in the rebuilt Temple. Reform Judaism, however, has abolished or modified these prayers in keeping with the conception of sacrifice as a once adequate but now outmoded form of worship, and some Conservative congregations have also rephrased references to sacrifices so that they indicate solely past events without implying any hope for the future restoration of the rite.

Christianity. The notion of sacrifice emerged in the early Christian communities in several different contexts. The death of Christ upon the cross preceded by the Last Supper was narrated in the Gospels in sacrificial terms; the life of Christ, culminating in his Passion and death, was seen as the perfect sacrifice, and his Resurrection and glorification were seen as God the Father's seal of approval on that life. The notion that members of the church are vitally linked to Christ and that their lives must be sacrificial was also elaborated, especially in the letters of St. Paul. Moreover, from the first decades of the church's existence, the celebration of the eucharistic meal was connected with the sacrifice of Jesus; it was a "memorial" (*anamnēsis*)—a term denoting some sort of identity between the thing so described and that to which it referred—of that sacrifice.

The Eucharist as sacrifice

The interpretation of sacrifice and particularly of the Eucharist as sacrifice has varied greatly within the different Christian traditions, partly because the sacrificial terminology in which the Eucharist was originally described became foreign to Christian thinkers. In short, during the Middle Ages, the Eastern Church viewed the Eucharist principally as a life-giving encounter with Christ the Resurrected; the Roman Church, however, saw it primarily as a bloodless repetition of the bloody sacrifice of Christ on the cross. For the Protestant Reformers in the 16th century, the sacrifice of Christ was unique and all sufficing, so that the idea of repeating it in cult became unnecessary. Sacrifice was separated from liturgy and was associated, especially in Calvinist Protestantism, with the personal ethical acts that should be made by a Christian believer. The ecumenical movement of the 20th century, bolstered by modern biblical scholarship, has led some of the Christian churches—*e.g.,* the Roman Catholic and Lutheran churches—to realize that they are not so far apart in their understanding of the Eucharist as sacrifice as was formerly thought and that they hold many elements of belief in common.

Islām. Sacrifice has little place in orthodox Islām. Faint shadows of sacrifice as it was practiced by the pre-Islāmic Arabs have influenced Muslims, so that they consider every slaughter of an animal an act of religion. They also celebrate feasts in fulfillment of a vow or in thanksgiving for good fortune, but there is no sacrificial ritual connected with these festive meals. On the last day of the annual pilgrimage to Mecca, animals are sacrificed; nevertheless, it is not the sacrificial rite that is important to the Muslims, but rather their visit to the sacred city.

Conclusion. The organization of sacrificial rites in the different cultures and religions has undoubtedly been influenced by a number of factors. Economic considerations, for example, certainly have had some impact upon primitive peoples in the selection of the victim and the time of

sacrifice and in the determination of whether the victim is consumed or totally destroyed and whether the sacrificer is an individual or a collective group. The importance of such factors is an aspect of sacrifice that deserves increased investigation. Nevertheless, sacrifice is not a phenomenon that can be reduced to rational terms; it is fundamentally a religious act that has been of profound significance to individuals and social groups throughout history, a symbolic act that establishes a relationship between man and the sacred order. For many peoples of the world, throughout time, sacrifice has been the very heart of their religious life.

(R.F.)

Rites of passage

Rites of passage are ceremonial events, existing in all historically known societies, that mark the passage from one social or religious status to another. This section describes these rites among various societies throughout the world, giving greatest attention to the most common types of rites, and discusses their purposes from the viewpoints of the people observing the rites, and their social, cultural, and psychological significance as seen by scholars seeking to gain an understanding of human behaviour.

NATURE AND SIGNIFICANCE

Many of the most important and common rites of passage are connected with the biological crises of life—birth, maturity, reproduction, and death—all of which bring changes in social status and, therefore, in the social relations of the people concerned. Other rites of passage celebrate changes that are wholly cultural, such as initiation into societies composed of people with special interests—for example, fraternities. Rites of passage are universal, and presumptive evidence from archaeology in the form of burial finds strongly suggests that they go back to very early times. The worldwide distribution of these rites long ago attracted the attention of scholars, but the first substantial interpretation of them as a class of phenomena was presented in 1909 by the French anthropologist and folklorist Arnold van Gennep (1873–1957), who coined the name rites of passage. Van Gennep saw the rites as means by which individuals are eased, without social disruption, through the difficulties of transition from one social role to another. On the basis of an extensive survey of preliterate and literate societies, van Gennep held that the rites consist of three distinguishable, consecutive elements, called in French *séparation, marge,* and *agrégation,* which may be translated as separation, transition, and reincorporation, or as preliminal, liminal, and postliminal stages (before, at, and past the threshold). The person (or persons) on whom the rites centre is first symbolically severed from his old status, then undergoes adjustment to the new status during the period of transition, and is finally reincorporated in society in his new social status. Although the most commonly observed rites relate to crises in the life cycle, van Gennep saw the significance of the ceremonies as being social or cultural, celebrating important events that are primarily sociocultural or man-made rather than biological. The British anthropologist A.M. Hocart (1884–1939) held that the passage from one status to another was the result rather than the cause of these ceremonies; thus the rites both induced and allayed personal and social stress rather than merely allaying it. Basing his views on circumstances in a few ancient civilizations, Hocart thought that all rites of passage were based on the model of ritual of investiture of kings, in which symbolic killing and rebirth of the new ruler, and sometimes actual killing of the old, were required. Later scholarship has shown that symbolic death and rebirth into the new status are common forms of symbolism in rites of passage of various kinds and that the symbolic killing and rebirth of rulers is therefore not appropriately viewed as the prototype of all rites.

Arnold van Gennep's theory

FUNCTIONS

Modern scholars in the social sciences characteristically accept the views of van Gennep about the social and psychological significance of rites of passage; that is, passage rites are seen to have positive value for the individual in relieving stress at times when great rearrangements in his life occur, such as are brought by coming of age, entering marriage, becoming a parent, or at the death of a close relative, and in providing instruction in and approval of his new roles. The rites are seen also to be socially supporting in various ways. Such support includes roles of the rites in preventing social disruption by relieving the psychological stress of the individuals concerned; providing clear instruction to all members of societies to continue life in normal fashion with new social alignments; the affirmation they provide of social and moral values expressed and thus sanctioned as part of the ceremonies; and the social unity they foster by joint acts and joint expression of social values. During most of man's history, rites of passage have generally been religious events; that is, they have been conducted in a religious framework and regarded as religious acts and hence possessed special authority. From the viewpoint of modern social science, however, their nature is generally seen as being fundamentally secular. Mankind gives social attention to all events regarded as being socially important. Until recent times, religion was intimately connected with most aspects of life, and events of such social importance as the changes in society that the rites celebrate were most frequently incorporated in the system of religious belief and act. The tendency of recent decades toward secularization of rites of passage strongly suggests that the primary significance of most rites is social or secular rather than religious. In the modern, scientifically minded nations of the world many rites of passage, such as rites of initiation into fraternal and honorary societies are wholly secular; others have only small elements of religion, and even marriage may be a wholly secular rite.

Rites as entertainment

One of the primary functions of rites of passage that is often overlooked by interpreters, perhaps because it appears obvious, is the role of the rites in providing entertainment. Passage rites and other religious events have in the past been the primary socially approved means of participating in pleasurable activities, and religion has been a primary vehicle for art, music, song, dance, and other forms of aesthetics.

From its beginning, the study of the significance of rites as a class of phenomena has attempted to account for similarities and differences in the rites among societies of the world. The similarities are striking and doubtless reflect the close similarity in ways of human thought. Modern attempts to account for similarities and differences have generally given little attention to and reached no consensus concerning the nature of the innate psychological factors involved in the genesis of the rites. Attempts to understand rites of passage have instead generally been sociocultural interpretations that view the rites as part of an integrated sociocultural system, the man-made part of human life. Religion and rites of passage are thus seen as elements in this system that affect and are affected by other elements in the system such as means of gaining a livelihood and the manner in which society is aligned in groups. Most modern analysts accordingly have interpreted both differences and similarities in rites of passage on the basis of their sociocultural context. The inventive and symbolic capabilities of mankind are treated as a constant factor, and analytic attention is given to differences and similarities in the sociocultural contexts in which rites are found. In attempting to understand why marriage is an extremely elaborate rite in one society and very simple in another society, for example, scholars have looked to the social order and to the manner of gaining a livelihood to judge the relative importance of the enduring unions of spouses. Following the view that culture, including the social order, composes a coherent, inclusive system, modern scholarship has, in short, most commonly interpreted rites of passage in terms of their functional significance in the social system.

The significance or stated goals of rites of passage as these exist in the minds of the actors are regarded as quite inadequate for gaining an understanding of the functional significance of the rites. Very often, rites of passage are said to have goals such as dispatching the spirit of a dead

person to another world, protecting the newborn, the new adult, and the newlywed from evil influences. Often the explicit goals of the rites have been forgotten and their continuation is a matter of following tradition, so that means have become goals. Although scholars have noted the explicit goals of these rites, they have characteristically given greatest emphasis to inferring functional significances that are not obvious to the actors in the rites. In so doing they have broadened their investigations from observations of the symbolism of rites to include prominently all of the behaviour during the rites and their social contexts, learning the social identities of the performers, and their relationships to other performers and the entire society.

CLASSIFICATIONS OF RITES

No scheme of classification of passage rites has met with general acceptance, although many names have been given to distinguishable types of rites and to elements of rites. The name purification ceremonies, for example, refers to an element of ritual that is very common in rites of passage and also in other kinds of religious events. In most instances, the manifest goal of purification is to prepare the individual for communication with the supernatural, but purification in rites of passage may also be seen to have the symbolic significance of erasing an old status in preparation for a new one (see also *Purification rites and customs*, below).

Other names that have been given to passage rites often overlap. Life-cycle ceremonies and crisis rites are usually synonymous terms referring to rites connected with the biological crises of life, but some modern scholars have included among crisis rites ritual observances aimed at curing serious illnesses. Ceremonies of social transformation and of religious transformation (for both see this section, below) overlap and similarly overlap crisis rites. Religious transformations, such as baptism and rites of ordination, always involve social transformations; social transformations such as at coming-of-age and induction into office may also bring new religious statuses, and life-cycle ceremonies similarly may or may not involve changes in religious statuses. It is nevertheless sometimes useful to distinguish the various rites by these names.

Life-cycle ceremonies. Life-cycle ceremonies are found in all societies, although their relative importance varies. The ritual counterparts of the biological crises of the life cycle include numerous kinds of rites celebrating childbirth, ranging from baby "showers" to pregnancy rites to rites observed at the actual time of childbirth and, as exemplified by Baptism and the fading Christian rite of Churching of Women, a ceremony of thanksgiving for mothers soon after childbirth. These rites involve the parents as well as the child and in some societies include the couvade, which in its so-called classic form centres ritual attention at childbirth upon the father rather than the mother. At this time the father follows elaborate rules of ritual procedure that may include taking to bed, simulating labour pains, and symbolically enacting the successful birth of a child. In all societies some ritual observances surround childbirth, marriage, and death, although the degree of elaboration of the rites varies greatly even among societies of comparable levels of cultural development. Rites at coming-of-age are the most variable in time in the life span and may be present or absent. In some societies such rites are observed for only one sex, are elaborate for one sex and simple for the other, or are not observed for either sex. Characteristically, rites at coming-of-age are not generally observed in the modern industrial civilizations or, as in the Jewish Bar Mitzva and the Protestant confirmation of the United States, exist today more or less as vestiges of formerly important religious rites. Among the elaborate civilizations of Asia, rites at coming-of-age have similarly waned in recent times. The elaborate rites observed a century ago in Japan when young men and young women reached social maturity are only rarely observed today and are virtually unknown to the general population. Death is given social attention in all societies, and the observances are generally religious in intent and import. In societies that fear dead bodies the deceased may be abandoned, but they are nevertheless the focus of ritual attention. Most commonly, rites at death are elaborate, and they include clearly all of the stages of separation, transition, and reincorporation first noted by van Gennep. See below under *Death rites and customs*.

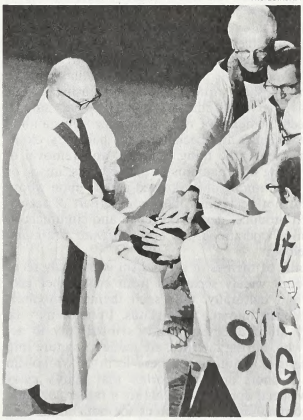
The couvade

Ceremonies of social transformation. Ceremonies of social transformation include all of the life-cycle ceremonies, since these involve social transitions for the subjects of the ritual and also for other persons. When a man or woman dies, for example, he assumes a new social role as a spirit that may be socially important to the living; the bereaved spouse becomes a widow or widower; and the children have an unnamed but changed status as lacking one parent. A vast number of rites of social transformation, such as rites of initiation into common-interest societies, have no direct or primary connection with biological changes. These are abundant in the United States and European nations, usually as secular ceremonies. In primitive societies, rites of this kind mark induction into age-graded societies, principally limited to males, and a variety of common-interest societies such as warrior societies, curing societies (special groups whose purpose is to cure illnesses), and graded men's societies that are hierarchically ranked in prestige. Whether hereditary or achieved by appointment or election, assumption of important office in various kinds of societies is often observed by elaborate ritual. Any other events involving changes in social status tend to become the subjects of institutionalized ritual, which is then a prerequisite for the new status. Common examples are initiation ceremonies of college fraternities, sororities, and honorary societies; adult fraternal societies, and social groups of other kinds centred on common interests. Other social changes of importance that apply to a substantial number of people but do not involve initiation into organized social groups are also given ritual attention. Common among these are graduation exercises, festivities marking retirement from work, and various kinds of award ceremonies.

Ceremonies of religious transformation. Religious-transformation ceremonies signal changes in religious statuses, which may be matters of the greatest importance to the people. Performing ritual such as making sacrifices and offerings may be required in the normal course of life, and these acts may be regarded as conferring a new religious status or state of grace. Sacrifices are a frequent feature of rites of passage, and for important ceremonies such as coronations and funerals of rulers, have sometimes required the sacrifice of many human beings (see also above *Sacrifice*). Among the laity, entry into a religious society or the assumption of any other new religious role is customarily an event celebrated by rites such as those

Sacrifices

The Lutheran



Ordination of a Lutheran pastor by the laying on of hands.

of baptism and confirmation. Among professional religious personnel, the achievement of any distinct status of specialization is ordinarily observed by rites corresponding to the Christian rites of ordination—the rites through which religious functionaries become entitled to exercise their respective functions. As with other rites of passage, these rites may be simple or complex, and their degree of complexity may generally be easily seen as reflecting the religious and social importance of the newly acquired status. A single element of an elaborate rite in one society, such as circumcision or the dressing of the hair in a distinctive way, may in another society be the central or sole event of rites of either social or religious transformation. These ceremonies may, accordingly, be called rites of circumcision or be identified by the name of the style of hairdress.

Other ceremonies. The term rites of passage is occasionally applied to institutionalized rites for curing serious illness and, rarely, to cyclic ceremonies such as harvest festivals. No new social or religious status is ordinarily gained by recovery from illness or participation in harvest rites, however, and these ceremonies have probably been included among the rites of passage because of similarities in their ritual procedures. In some societies, recovery from a very critical illness is regarded as a divine sign that the erstwhile invalid should assume the role of a religious specialist, but rites of ordination are quite separate. Some elements of ceremonies pertaining to changes in the seasons may be seen as incorporating acts of separation and incorporation, symbolically saying goodbye to the old season and welcoming the new, but these are not customarily called rites of passage. Although clearly denoting a change in social status, divorce has rarely been regarded as a rite of passage. Festive observances at this time are perhaps common in some societies, but they are often informal practices of the individual or simple acts of local custom, such as discarding wedding rings, that are not institutionalized in the entire society. The absence of divorce from the conventional roster of rites of passage illustrates an outstanding characteristic of this class of rites: all celebrate events that are either socially approved or, like death and illness, unavoidable. Rites of passage that signal the assumption of social statuses disapproved by society are both out of keeping with the prevailing interpretation of the rites as being socially supportive and would broaden them to cover such events as trials by jury and commitment to prison for serious crimes.

Symbolic aspects of ceremonies. Whatever their subclassification, elaborate rites of passage are commonly rich in symbolism that prominently includes representations of the states of separation and transition and, especially, insignia of the new status. Most common among these markers of new status are alterations and embellishments of visible or invisible parts of the body, distinctive garments and bodily decorations, and insignias corresponding to symbols of office. All parts of the body that may be altered or embellished without ordinarily causing serious disability have served as the symbols of social statuses and have been elements of rites of passage. Outstanding among these insignias are special styles of hairdress, clothing, and ornaments; the filing, staining, and removal of teeth; the wearing of ornaments in pierced ears, noses, or lips; tattoos and, among dark-skinned people upon whom tattoos would not be visible, their counterpart of scarification, which produces designs in relief; and circumcision or other genital operations (see also *Religious dress and vestments* below).

Insignia and body decorations

Several motifs or themes of symbolism commonly recur among societies widely separated from each other geographically and culturally. One such theme symbolizes death and rebirth into the new status. Initiates may be ceremonially killed and then made symbolically to act like infants who, during the rites, are made to mature into their new statuses. Another common form of symbolism makes use of doors or other portals that signify entry into the new social domain. Ordeals are a rather common feature of coming-of-age ceremonies for both males and females, and they are also used in rites of initiation into men's societies of various kinds. Success in passing the ordeals is customary and signifies mastery of the roles that are to be assumed.

A universal feature of rites of passage is the proscription of certain kinds of ordinary behaviour. Sexual continence is a common rule, as is the prohibition of ordinary work such as farming, hunting, and fishing. Many rites prohibit certain behaviour or prescribe the reverse of ordinary behaviour. Among Indians of the western United States, for example, a taboo against scratching the body with the fingers was common during ritual periods. In other societies, ritual behaviour required that the subjects of ritual sit in a remarkable fashion, wear articles of clothing inside out or backward, or wear the clothing of the opposite sex. These acts all may be seen as dramatizations, by contrast, of the events that they celebrate, thereby making them memorable.

A representative example. Rites of passage marking very important events customarily include all of the three stages described by van Gennep. A representative example is afforded by the traditional rites surrounding childbirth as these were commonly observed in Japan until recent years. Observances began when a woman learned she was pregnant. Partly for stated reasons of promoting health and partly for supernaturalistic reasons, she thenceforth abstained from certain foods and ate others. During the fifth month of pregnancy she donned a special girdle, ordinarily procured from a Buddhist temple and supernaturally blessed. Relatives offered prayers for the well-being of the woman and her child. When birth seemed imminent, she was isolated from all other persons except the women who attended her and remained in isolation for a fixed number of days after parturition. This period was most commonly 33 days, which was divided into stages preceeding from severe restriction of her acts to final complete resumption of all normal activities. She had at first to follow a number of special rules of diet and could not perform normal household tasks. During the period of isolation, the mother was regarded as polluted from the flow of blood during childbirth and therefore dangerous to other people and dangerous or offensive to supernatural beings of the Shintō religious pantheon. She could not make the usual offerings or say prayers before the household shrines to Shintō gods or have any other kind of contact with them. To avoid offending the sun goddess, her clothing and that of her child when laundered could never be hung in direct sunlight to dry but instead were placed in the shadows of the eaves of the house. For the same reason, she covered her head with a cloth when she stepped outside the house near the end of the period of isolation. Water and cloths used in washing the mother after parturition were considered to be polluted and were buried in the ground beneath the floor of the room of confinement. After a fixed number of days passed, the mother was permitted to resume bathing and again perform some but not all of her ordinary work in the house. Other restrictions on behaviour were removed at fixed times, and when the full period passed, the mother and her female aides performed a ceremony of purification by sprinkling salt on the mother and on the floors of the dwelling. The beginning of a new, normal period free from pollution also was symbolized by kindling a new fire in the household cooking stove. Now ready to return to normal life, the mother ate a ceremonial meal with other members of the family and resumed ordinary relationships with supernatural beings and other human members of the community.

Japanese passage rites

RITES OF PASSAGE IN THE CONTEXT OF THE SOCIAL SYSTEM

Most of the scholarly interpretations of rites of passage of the 20th century have considered their relation to the social system and have seen the functional significance of the rites as a contribution to the maintenance of society as a system of congruent parts. Explicit or implicit in this line of reasoning is the idea of equilibrium found in any scientific theory concerned with systems. For the system to operate effectively, its elements must be mutually supportive or congruous, and the system is then described as being in a state of equilibrium. Social systems

Changes of status and social equilibrium

embrace a fixed number of people and a fixed number of roles. Changes in either the number of the people or the proportions of statuses disturb social equilibrium. When a child is born, a new member is added to society; the social behaviour and statuses of its parents change, and these changes also affect other members of society. Other social changes that are the subjects of passage rites similarly disrupt the state of social equilibrium. Rites of passage are seen to foster the development of a new state of equilibrium in adjustment to the social changes upon which the rites focus. By means of the rites, members of society are informed of the new social circumstances and at the same time give social approval to them. Individuals upon whom the rites focus are assured of success in their new roles by the ritual observances and are given psychological reassurance in a number of other ways. They and all other members of society are instructed by the ritual enactment of their new social relations to return to normal behaviour incorporating the added or lost personnel and the added, lost, or changed social statuses. The same general kind of reasoning is applied to various other religious ceremonies. The anthropologists Eliot D. Chapple and Carleton S. Coon interpret all rites of passage and other group rites as "rites of intensification." Calling special attention to the ritual depiction of habitual relationships for the statuses involved, Chapple and Coon state that this behaviour "has the effect of reinforcing or intensifying their habitual relations, and thus serves to maintain their conditioned response . . . In the technical (physiological) sense, the performance of these rites prevents the extinction of habits . . . to which the individual has been trained."

Closely related to the function of passage rites in restoring social equilibrium, in the anthropologists' interpretation, are a group of additional effects or functions, some of which apply first to the individuals whose statuses change and, through their behaviour, to the entire social group. Other functional effects apply directly to the entire society. By allaying the anxiety of individuals who are undergoing change, social disruption is avoided. Rites of passage characteristically give assurance of mastery of the new roles and often include instruction in the new roles. In the many societies in which statuses and roles are clearly distinguished by sex, the rites symbolically emphasize these differences, thereby instructing the initiates and aiding them in sexual identification. The anxiety and potential social disruption caused by death and the grief of the bereaved are similarly held in check. Funeral rites customarily point up grief and then firmly instruct the bereaved to resume normal behaviour that is not disruptive to others. The joint performance of rites and the joint expression of moral and other social values that are included among ritual acts may be seen as directly promoting group solidarity through communion with one's fellows and affirmation or reaffirmation of rules and ideals that foster social harmony.

Rites of passage and all other group rites are seen to be socially supporting in still another implicit way. The joint rites are customarily a rehearsal or dramatization, with supernatural sanction, of a part or all of the social order of the society. Relatives have special roles that are congruent with, or enactments of, their positions in normal social life, and the entire social hierarchy may be on display during the rites through the assignment of ritual roles. Thus statuses of kinship, caste, social equality, and hierarchy are all seen to be reinforced by dramatic presentation of them.

Accepting this group of interpretations of the social significance of rites of passage, anthropologists have also attempted to understand variations in the degree of elaboration of rites of passage among societies of the world. A fundamental assumption is the commonplace idea that the greater the importance of a social change the greater the ritual attention will be. The birth, marriage, and death of a ruler obviously are more important to the entire society than these events in the life of a commoner. The importance of such events is not always obvious, however, and their relative importance is often difficult to see when different societies are compared. Rites of marriage, for example, may be very simple or very elaborate in different societies of the same economic base and comparable levels of cultural development. Recourse to consideration of features of the social order has allowed a reasonable explanation of the differences. Marriage rites in matrilineal societies, for example, which are organized into subgroups primarily upon a principle of descent through female lines only, tend to be simple, and divorce in these societies is also simple. Marriage rites in patrilineal societies (in which descent is through male lines), however, tend to be elaborate, and divorce initiated by females is difficult.

Marriage rites in matrilineal and patrilineal societies

In matrilineal societies, the social core is composed of groups of male and female relatives united by female lines, which are economically distinct from other groups and self-sufficient. Where the matrilineal principle of organization is strong, the role of the husband and father, who belongs to a matrilineal group different from that of his wife, is not that of economic provider for his wife and children. Instead, he is the economic mainstay for his sister and her children, and his contact with his wife may be limited to spending nights with her. The brothers or other male relatives of a mother not only provide economically for her children but also assume what is elsewhere the role of the father in socializing children. Enduring unions of marriage are not vital to such matrilineal societies. If marriages end in divorce, the matrilineal ordering of society assures approved social identification, economic support, and affective ties for the children and their mother and also assures continuance of the society as long as males are available as procreators. In patrilineal societies, however, the role of the mother, who is the outsider in the group, is vital for the birth and rearing of the children, and she and her children are dependent upon her husband for economic support. Strong sanctions are placed upon marriages in these societies to help ensure lasting unions. Marriage ceremonies are correspondingly elaborate, often involving the transfer of property, which among some African societies is called marriage insurance for the reason that it must be returned if the marriage falls asunder.

In societies such as those of the United States and European nations, where the important unit of kinship is ordinarily limited to the nuclear family of parents and children and where important social affiliation does not depend upon descent through one sex of progenitors, enduring unions of marriage are also vitally important. Rites of passage at marriage traditionally have been required by law as well as by the church, and many other sanctions on lasting marriages are imposed by laws concerning divorce, communal property, and the care of children. The bride and groom who have undergone the whole series of traditional rites of passage from engagement parties to the religious ceremony may reasonably be seen as more firmly married than couples united by a simple civil ceremony (see also FAMILY AND KINSHIP).

PSYCHOLOGICAL ASPECTS OF RITES OF PASSAGE

Less scholarly attention has been given to psychological than to social or cultural aspects of rites of passage, in large part because the scholars concerned with such rites in world societies have been principally anthropologists, who lean toward sociocultural interpretations. As the foregoing discussion of passage rites in social context illustrates, psychological aspects of rites nevertheless enter strongly if often implicitly into anthropological interpretations as fundamental matters in social solidarity and social disorder. Emotional ties to kin and other members of society, personal identification with social groups and religious statuses, and commitment to religious ideology and other values are reinforced and sometimes created by rites of passage. In a realistic sense, the rites serve as blueprints for social relations and religious behaviour, making clear the acceptable ways to act and at the same time pointing up and reinforcing affective relations with other people and with the supernatural. Familial rites of ancestor worship, for example, are not only reinforcements of familial solidarity but also have psychological value in reinforcing emotional ties among relatives.

Psychological interpretations of passage rites have given greatest emphasis to their value in allaying personal anxiety. A recurrent feature of the rites are acts of magic that

assure that the outcome of the endeavour will be successful. In the words of the anthropologist Bronislaw Malinowski these acts serve symbolically and psychologically "to bridge over the dangerous gaps in every important pursuit or critical situation" that exist because of man's lack of control of the universe. By such magical means as miniature boats floated in streams or carried away by the tide, the dead are shown symbolically to go successfully to the other world, and childbirth and successful maturation are similarly depicted magically. The subjects of rites of passage frequently act out their future roles to the approval of all others. Numerous acts of magic that are not essential to changes in social status may be incorporated in rites of passage and may be seen to give psychological assurance relating to the future life of the individual. Traditional Japanese practices at childbirth, for example, required that when a girl was born, the placenta be buried in the ground outside the entrance to the dwelling to insure that the girl, when mature, marry in normal fashion and leave the family. When a boy was born, the placenta was buried inside the house to ensure that he remain at home when mature. The ordeal that a young man or young woman must often undergo during rites of coming-of-age may similarly be seen to provide psychological assurance of success in the new status. Ordeals of this kind are characteristically uncomfortable or frightening, but they are events that any human being ordinarily can endure.

Psychotherapeutic value of rites

The psychotherapeutic value of passage rites surrounding events in which stress may be acute, such as childbirth, death, and serious illness, is clearly apparent and essentially follows the principles of modern secular psychotherapy. The subject is made the centre of concentrated attention by many people, is given reassuring evidence of their regard for him, and, by means of magic and the intervention of supernatural beings, is assured of a successful outcome. These events are carried out on a high emotional pitch, which gives them added force. When anxiety is induced by religious beliefs themselves, such as by ideas that if ritual acts are not performed calamitous results will follow, the rites of passage may be said both to create and to allay anxiety.

Where particular social statuses have special honour and prestige, the mere existence of these statuses offers opportunities for gaining psychological satisfaction, and the requirements for gaining these statuses serve to guide behaviour in socially approved channels that offer psychological satisfaction.

Other interpretations of psychological aspects of passage rites have relied upon ideas derived from or inspired by the psychoanalyst Sigmund Freud (1856–1939). These have sometimes concerned the symbolism involved in the rites and, in anthropological interpretations, have dealt with both Freudian ideas of symbols and the social order. The psychologist Bruno Bettelheim has interpreted cicatrization (inducement of scars) of males in rites at coming-of-age as symbolic wounds indicating subconscious male envy of the vagina, the counterpart of Freud's idea of penis envy. A psychologically oriented anthropologist J.M. Whiting, and others have combined sociological and psychoanalytic theories in attempting to explain why male initiation ceremonies are conducted in some societies and not in others. Harsh rites, sometimes including genital operations, are held to be correlated with societies in which infant males have long and intimate contact with their mothers, and husbands are prohibited from sexual intercourse with their wives for a period of two years or more. The long and exclusive relationship between mother and son is assumed to lead to strong emotional dependence upon the mother by the son, which becomes potentially disruptive at the time the son reaches puberty. The harsh rites are seen to break the bond of dependency and avoid potential social disruption that might otherwise result from discord between son and father at this time.

RITES OF PASSAGE IN THE CONTEXT OF THE RELIGIOUS SYSTEM

Certain passage rites represent first and foremost transformations in the religious statuses or circumstances of the people concerned. As already noted, rites of passage are customary upon the assumption of a new status as a religious professional. During most of man's history, however, rites of passage have carried among their implications a change to a new religious state for the ordinary members of society as well as for the professional religious person. Among the culturally advanced societies of the world with orders of priests, ideas of the significance of symbolism in passage rites may be elaborate and sophisticated, representing the rites as different states of grace or, as in Hinduism, cyclic states involving death and rebirth. In many societies, one is not fully or properly a human being until he has undergone the rites of passage appropriate for his age and sex. In some societies, fully human status is not reached until the rite of Baptism has been performed, and children who die before that time may be interred with special rites in places separate from those of the dead who have been baptized. When passage rites are religious ceremonies, as has generally been the circumstance until modern times, some state of sacrament or divine blessing, vaguely or clearly defined, is entailed. At the time of death, rites of passage placing the deceased in the realm of the supernatural customarily have been required. Symbolism in many rites of passage denotes communion with the supernatural. In common with many other kinds of religious events, then, passage rites relate the individual and the society to the sacred world, conferring benefit upon him thereby (see also above *Sacred or holy*).

Rites of passage frequently have ethical import of value for the maintenance of social equilibrium. Where ethical or moral codes and religious beliefs are intimately connected or identified as one and the same, as in Christianity, Judaism, and Islām, the role of religious beliefs and acts may be seen to have strong value as social sanctions since the moral injunctions apply to human relations as well as to man's relations with the supernatural world. All societies have moral or ethical codes, rules of what is appropriate and inappropriate in human relations, and these are enforced by various means. Rites of passage, as noted above, commonly incorporate statements or dramatizations of moral values, and rites at coming-of-age often give moral instruction in highly explicit terms. No necessary or inherent connection exists, however, between morality and religious beliefs. Any serious breach of proper moral conduct results in the imposition of a network of sanctions, many of them secular. In some societies, religious beliefs have little bearing on morality in relations with one's fellow men, although violations of rules applying to relations with supernatural beings and supernatural forces may be regarded as bringing inevitable punishment or misfortune through the supernatural agency. Whenever morality is a part of religious precepts, the direct sanctioning force of passage rites stressing moral rules may be powerful and important to the maintenance of society. In other societies, the ethical import of passage rites and other features of religion may operate less directly. An example is provided by societies that revere but do not deify ancestors. Any breach of morality reflects unfavourably upon the ancestors, who may undertake no action of censure but nevertheless serve as a sanctioning force that is reinforced by death rites.

Morality and religious beliefs

PRIMARY RITES OF PASSAGE

In simple, primitive societies dependent for subsistence upon hunting and gathering, in which social groups are small and specialization in labour is limited to distinctions by sex and age, no social statuses may exist except those of child, adult, male, female, and disembodied spirit. Among primitive societies somewhat more advanced technologically and culturally, however, specialized groups based upon common interests appear, and these customarily require rites of induction or initiation. In culturally sophisticated societies, with elaborate divisions of labour, social statuses of leadership and specialized occupation are multiple. If all societies of the world, preliterate and literate, are considered, the most commonly recurrent rites of passage are those connected with the normal but critical events in the human life span—birth, attainment of physical maturity, mating and reproduction, and death.

Birth rites. Rites surrounding the birth of a child are

often a complex of distinct rituals that prescribe different behaviour on the part of the mother, the father, other relatives, nonfamilial members of the society, and with respect to the newborn. Observances may begin when pregnancy is first noted and may continue until the time of delivery, when the full rite of passage is observed, and for a variable period of time afterward. In many simple societies and in European societies of the past, the expectant mother is isolated from other members of society at this time for the stated reason that the blood that flows during childbirth has inherently harmful qualities. Where this belief is strong, the classic couvade may be practiced. Regions of the world in which this practice was formerly common include the Amazon Basin of aboriginal South America, Corsica, Spain, among the Basques of France and Spain, and among various societies of Asia. Old ethnological writings have created the impression that ritual attention is limited entirely to the father. Later investigations have made it appear doubtful that the mother in any society is free from ritual requirements. In many societies, rites that have been called the couvade are observed by both parents. The anthropologist Alfred L. Kroeber (1876–1960) reported that among most of the many tribes of aboriginal California, rites at childbirth were much alike for both mother and father. To prevent harm to their child and to other people during the ritual period, the parents observed food taboos; ate in seclusion; avoided contact with other people; did as little work as possible; and refrained from various other acts of ordinary behaviour that included cooking, touching tools, and eating salt, meat, and fish. Women often were under injunctions to scratch themselves only with a stick or a bone for fear that their nails at this time would leave permanent scars on their bodies.

Isolation of expectant mothers

Practices of sympathetic and contagious magic relating to birth and the later well-being of both child and mother are abundant and diverse. Among Indians of aboriginal British Columbia, the mother inserted a smooth beachstone, an eel, or other slippery object under her garment at the neckline, permitting it to slide to the ground to symbolize and insure quick and successful childbirth. In societies of Southeast Asia and Indonesia, religious specialists dressed as women simulated successful delivery. Rites directed toward the newborn similarly symbolize or ensure health and well-being and, after some days, weeks, or months have passed, often include Baptism or other ritual acts that introduce the child to supernatural beings. Both child and mother are often regarded as being defenseless at this time, and many ritual acts have the purpose of protecting them from harmful supernatural beings and forces. In Southeast Asia and Indonesia, a practice called mother roasting, which requires that the mother be placed for some days over or near a fire, appears once to have had the goal of protecting the mother from such evil influences. This practice survives today in altered form in the rural Philippines, where it is regarded as having therapeutic value.

Mother roasting

Native explanations of the ritual procedures at childbirth reflect beliefs of a mystic affinity between parent and child, and many of the prescriptions have the manifest goals of preventing harm to the infant until it is able to fend for itself. Among South American Indians practicing the classic couvade, this belief of affinity between father and child relates to the soul, which is not fully transmitted to the child until the end of the ritual period.

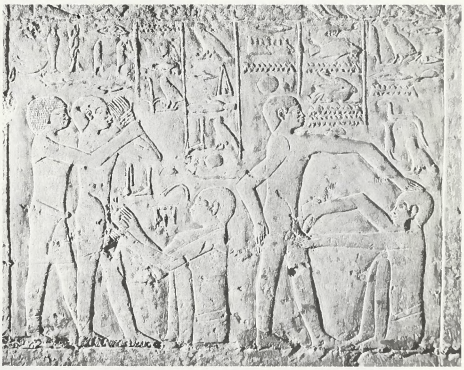
In addition to the social (communal) and psychological significances of birth rites already noted, scholars have offered interpretations of these ceremonies as reinforcing familial ties. The classic couvade has been seen by Malinowski as a sympathetic symbolic stressing of the relationship between the husband and the wife and her kin, which is instituted when the child is born. In addition to serving as a means of allaying husbandly anxiety over the welfare of the wife, the practices of the couvade establish social paternity, which, in turn, promotes familial and societal solidarity.

Initiation rites. The most prevalent of rites of initiation among societies of the world are those observed at coming-of-age. These have frequently been called puberty rites, but, as van Gennep argued long ago, this name is inappropriate. Puberty among females is often defined as the time of onset of menses (the menstrual flow), but no such clearly identifiable point exists in the sexual maturation of males. Moreover, the age at which rites of attaining maturity are observed vary greatly from society to society, going far beyond the normal range of years at which sexual maturity is attained. The definition of maturation is thus seen to be largely social or cultural rather than solely biological.

The full range of stages of passage rites is often followed in rituals at coming-of-age. Ordeals or other tests of manhood and womanhood are also common. Some of these practices in preliterate societies seem incomprehensible or absurd until their nature as evidence of qualification for the new social statuses is understood. Among the Bemba tribe of Africa, for example, girls were required to catch water insects with their mouths and to kill a tethered chicken by sitting on its head. Circumcision or other genital operations are also a fairly common feature of rites celebrating the attainment of maturity. Although most commonly applying to males, genital operations are performed on females in a few societies. It seems quite clear that circumcision and other alterations of the sexual organs have not until modern times been regarded as therapeutic surgery. These operations may have psychological significance following Freudian lines of interpretation, but it seems clear that they are also significant as insignia of social status. Where circumcision is the practice for male initiates, the uncircumcised male is not a full-fledged adult. It may be remembered that at this time other parts of the body are also modified, by incision, piercing, filing, tattooing, and by other kinds of practices that are not painful. Circumcision may in fact have no direct relation to the attainment of sexual maturity. In native Samoa, boys were circumcised at any age from three to 20.

Ordeals and other practices

Henri Stierlin—Ziolo



Egyptian circumcision relief, tomb of Ankhmahor, Saqqārah, 6th dynasty (2345–2181 BC).

An outstanding feature of rites at coming-of-age, which is generally less prominent or absent from other rites of passage, is their emphasis upon instruction in behaviour appropriate to the status of adults. Instruction in dress, speech, deportment, and morality may be given over a period of months. Very commonly, instruction is first given at this time in matters of religion that have heretofore been kept secret, and initiates may at this time be expected or required to commune with the supernatural, sometimes by means of revelatory trances induced by fasting, violent physical exertion, or the consumption of plant substances that produce hallucinations or otherwise alter the sensibilities.

Separation of male initiates from their mothers and all other females is also common, and ritual events may dramatize the transition from a world of women and children to one that is ideally male. Symbolism of these rites dra-

matizes the separation in such ways as by requiring the young men temporarily to wear the clothing of women and by rigid exclusion of all females from participation in the rites.

Among the technologically and scientifically advanced societies of the world, initiation rites have become increasingly secular. The great religions of the world all included rites at coming-of-age, but for much of the modern population of these nations the rites either are not observed or are simply vestiges of the old religious ceremonies. The most common rites of initiation are predominantly or wholly secular ceremonies conducted to celebrate such events as entry into a common-interest association or graduation from school. Rites of initiation such as into age-graded groups or common-interest societies follow essentially the same pattern as those at coming-of-age, and their simplicity or elaboration may be correlated with the importance of the new statuses.

As seen by social analysts, the significance of initiation rites of all kinds is the same as that of other rites of passage. Some emphasis is given to their didactic value and to their significance in sex-role identification. One question that has not been answered is why rites at coming-of-age are so poorly developed today among the technologically advanced societies of the world. Many factors, including changed views of the nature of the universe and changed social conditions, appear to have contributed to the decline of rites of passage. The supernaturalism traditionally present in the rites is no longer acceptable to many people, and in the United States and parts of Europe the association of adult status with sexual maturity as expressed in the term puberty rites has been unwelcome, a matter that should be excluded from notice rather than celebrated. Probably far more important in discouraging the rites has been the extreme variation in these nations in the age of social maturity. In the United States the ages differ at which one may legally drive a car, enter marriage, own and control property, buy alcoholic drinks and tobacco, enter military service, and vote; and in some of these matters the ages differ from state to state. The demands of modern civilization have, moreover, lengthened the age of social maturity, the time at which one is an economically productive member of society, and, dependent upon the number of years of formal education, have produced great variation in the age of full social maturity. The social and psychological value of rites of coming-of-age in making the transition to adulthood appears to be substantial, but modern cultural circumstances seem incompatible with the conduct of such rites.

Rites at coming-of-age in modern times

Marriage rites. It is assumed by anthropologists that marriage is one of the earliest social institutions invented by man, and, as already noted, rites of marriage are observed in every historically known society. These rites vary from extremes of elaboration to utmost simplicity, and they may be secular events or religious ceremonies. Subclasses of rites of marriage, named and unnamed, exist in many societies, beginning with ceremonies of betrothal that require complex formalities of transfer and exchange of goods, which are often regarded as compensation to the bride's kin group for their loss of the bride. Ceremonies of dramatic, sham "capture" of the bride by the groom and his relatives and friends have been common in both preliterate and literate societies. Marriage in these societies is seen by social analysts as a cooperative liaison between two different groups of kin, between which some feelings of hostility exist. Ceremonies of token capture are conducted even when betrothal and all other arrangements for marriage have long been completed to the expressed satisfaction of both sides, and the sham captures are interpreted as socially sanctioned channels for the expression and relief of feelings of hostility between the two kin groups. In some historically known societies of Africa, such sham battles between kin of brides and grooms may occur, with full societal approval, for years after a marriage during any kind of religious rite.

Sham capture of brides

Like rites at coming-of-age, ceremonies at marriage have often included clearly visible insignia of the new social status, in such forms as wedding rings, distinctive hair dress and garments, and tattoos, ornaments, or other embellishments that are regarded also as being decorative. Traditionally, preliminary rites have often provided instruction in the wifely role. Such instruction might be informal or conducted as a part of ritual. Rites of marriage proper also often give instruction through mimicry, dancing, and other symbolic acts that dramatically depict the woman's role in society, expressing her economic and social obligations and privileges with reference to her children, husband, other relatives, and still other members of society. Tests of maturity and rites with the purpose of promoting fertility have also commonly been included.

Bridal instruction



South Asian wedding custom; the bride regards her reflection in a mirror.
Frank Horvat—Black Star

In addition to sharing the functional significances of other passage rites, marriage ceremonies may be seen especially to stress social bonds between husband and wife and their kin groups. In most societies and during most of human history, romantic love has not been the means by which spouses are selected. Convention, often strongly sanctioned, has limited marriage to only certain classes of people. Mutual attraction between the spouses has been a matter of little or no importance. The importance of marriage with respect to spouses, children, other kin, and the orderly maintenance of society is readily inferrable. Rites of marriage place a sanction on unions of marriage that may be very powerful and thus serve as both a means of conducting an orderly and satisfying human life and also as sanctions for the orderly maintenance of society. A general correlation may be seen between the degree of elaboration of marriage rites and the social importance of enduring marriages in the society in question. Where, as in some of the large, industrial nations of the world, marriage rites are simple and sometimes secular, a host of other sanctions operate similarly to foster lasting unions.

Death rites. All human societies have beliefs in souls or spirits and an afterlife, and all conduct rituals when people die. See below under *Death rites and customs.* (E.N.)

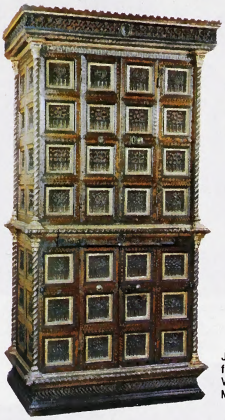
Death rites and customs

Throughout history and in every human society, the disposal of the dead has been given special significance. The practice was originally motivated not by hygienic considerations but by ideas entertained by primitive peoples concerning human nature and destiny. This conclusion is clearly evident from the fact that the disposal of the

Early practices



Multiple gateways and enclosures delimit the sacred area of the temple of Śrīrangam, India.



Means of designating sacred or ceremonial space

Jewish ark of the Law, functioning as sacred furniture, encloses and protects the Torah. Wooden ark from Modena, Italy, 1505. In the Musée de Cluny, Paris.



Muslim prayer rug, protective object associated with prayer, symbolizes the sacred area of the mosque. Silk and wool rug from Turkey, 17th century. In the Staatsbibliothek, Berlin.



Torii marks the entry to the sacred space around a Shintō shrine, Hakone-yama, Japan.

Plate 1: By courtesy of (top) UNESCO, (right) Staatsbibliothek, Berlin; photographs, (top) G. Kruysen, (centre left) Eddy Van der Veen, (bottom left) Ray Manley from Shostal—EB Inc.



Chimú ceremonial knife from Peru, repoussé gold. In the Art Institute of Chicago.



Tibetan skull cup, 18th century. In the Museum of Fine Arts, Boston.



Byzantine chalice, agate. In St. Mark's, Venice.

Objects used in the enactment of rituals





Divine beings worshipping a throne altar beneath the banyan tree. Sandstone relief from the great Stupa at Sānchi, Madhya Pradesh, India, 1st century BC.

Ryoo being performed at Itsukushima-jinja, Japan. The dancer's mask becomes a representation of holy forces and, as such, is an object of worship.

Plate 2: By courtesy of (top left) The Art Institute of Chicago; (top right) the Museum of Fine Arts, Boston, John Gardner Coolidge Collection; photographs, (centre) SCALA—Art Resource, (bottom left) Otsuka Kogei-sha Co., Tokyo, (bottom right) Pramod Chandra

Plate 3: By courtesy of (top centre) the Historical Museum, Stockholm © A.T.A. Stockholm; photographs, (top left) Holle Bildarchiv, Baden-Baden, (top right) FISA (centre, bottom) Shostal—EB Inc., (centre) W. Kaufman, (bottom) C. Reyes

Ritualistic and symbolic uses of the natural elements



Tree element transformed into the parasol shaft or *chattrāvali* of the Buddhist religion. Stone shaft in the Mohrā Morādu *stupa*, Taxila near Rāwalpindi, Pakistan.



Water element brought into the Christian totality in the sacrament of Baptism, performed at baptismal fonts. Stone font from Tingstäde Church, Ostergötland, Sweden, 12th century. In the Historical Museum, Stockholm.



Tree element symbolized by Christians as the cross. Crucifix, polychromed wood, 12th century. In the Museo de Bellas Artes de Cataluña, Barcelona.



Stone element transformed into an altar by the Minoan culture, palace at Mália, Crete.

Mountain element represented by the Mayan culture in pyramidal stone temples. Temple of Inscriptions, Palenque, Mexico.



Devices for the ritual use of sound, fragrance, and light



Rdo-rje and bell, bronze. Tibetan, 19th century. In the Newark Museum, New Jersey.



Tripod incense burner, bronze. Chinese, late Chou dynasty. In the Metropolitan Museum of Art, New York City.



Bull-roarer from the Gulf of Papua, New Guinea. early 20th century. In the Field Museum of Natural History, Chicago.



Hanukka *menora*, silver with enamel medallions, by Johann Adam Boller, early 18th century, Frankfurt am Main, West Germany. In the Jewish Museum, New York City.

Shaman's drum, Lapland, wood and painted hide. In the Nationalmuseet, Copenhagen.

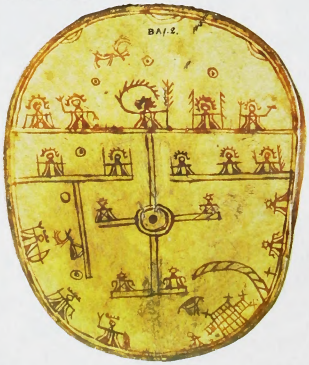


Plate 4. By courtesy of (top left) The Newark Museum, New Jersey, Shelton Collection; (top right) the Metropolitan Museum of Art, New York, Rogers Fund, 1947; (centre left) the Field Museum of Natural History, Chicago, Captain Fuller Collection; (bottom left) Ethnographical Department, Nationalmuseet, Copenhagen; (bottom right) The Jewish Museum, New York City

dead from the earliest times was of a ritual kind. Paleolithic peoples, such as the Neanderthals and later groups, not only buried their dead but provided them with food, weapons, and other equipment, thereby implying a belief that the dead still needed such things in the grave. This very significant practice can be traced back to great antiquity, possibly to about 50,000 BC.

Non-acceptance of death as the definitive end of life

The ritual burial of the dead, which is thus attested from the very dawn of human culture and which has been practiced in most parts of the world, stems from an instinctive inability or refusal on the part of man to accept death as the definitive end of human life. Despite the horrifying evidence of the physical decomposition caused by death, the belief has persisted that something of the individual person survives the experience of dying. In contrast, the idea of personal extinction through death is a sophisticated concept that was unknown until the 6th century BC, when it appeared in the metaphysical thought of Indian Buddhism; it did not find expression in the ancient Mediterranean world before its exposition by the Greek philosopher Epicurus (341–270 BC).

The belief that human beings survive death in some form has profoundly influenced the thoughts, emotions, and actions of mankind. The belief occurs in all religions, past and present, and decisively conditions their evaluations of man and his place in the universe. Mortuary rituals and funerary customs reflect these evaluations; they represent also the practical measures taken to assist the dead to achieve their destiny and sometimes to save the living from the dreaded molestation of those whom death had transformed into a different state of being.

RELEVANT CONCEPTS AND DOCTRINES

Life and death. The evidence of Paleolithic burials shows that already, in that remote age, various ideas were held about death and the state of the dead. The provision of food, ornaments, and tools in the graves implies a general belief that the dead continued to exist, with the same needs as in this life. Other customs, however, indicate the currency of a variety of notions about postmortem existence, particularly about the potentialities and destiny of the dead. Thus, the presence of red ochre in some burials suggests the practice of contagious magic: the corpse had possibly been stained with the colour of blood in order to revitalize it. The fact that in Paleolithic burials the skeleton has often been found lying on its side in a crouched position has been interpreted by some prehistorians as evidence of belief in rebirth, in that the posture of the corpse imitated the position of the child in the womb. In some crouched burials, however, there is reason for suspecting a more sinister motive; for the limbs are sometimes so tightly flexed that the bodies must have been bound in that position before rigor mortis set in. Such treatment of the corpse was doubtless prompted by fear of the dead, for similar customs have been found among later peoples. Preventive action of this kind has a further significance, for it implies a belief that the dead might be malevolent and had power to harm the living.

Fear of the dead

That death was sometimes regarded as transforming those who experienced it into a state of being balefully different from that of those living in this world is evident in later mortuary rites and customs. Indeed, the proper performance of funerary rites was deemed essential by many peoples, to enable the dead to depart to the place and condition to which they properly belonged. Failure to expedite their departure could have dangerous consequences. Many ancient Mesopotamian divinatory texts reveal a belief that disease and other misfortunes could be caused by dead persons deprived of proper burial. The fate of the unburied dead finds expression in Greek and Roman literature. The idea that the dead had to cross some barrier that divided the land of the living from that of the dead also occurs in many religions: the Greeks and Romans believed that the dead were ferried across an infernal river, the Acheron or Styx, by a demonic boatman called Charon, for whose payment a coin was placed in the mouth of the deceased; in Zoroastrianism the dead cross the Bridge of the Requiter (Činvato Paratu); bridges figure also in Muslim and Scandinavian eschatologies (speculations concerning the end of the world and the afterlife)—the Ṣirāṭ bridge and the bridge over the Gjöll River (Gjallarbrú)—and Christian folklore knew of a Brig o' Dread, or Brig o' Death.

It is significant that in few religions has death been regarded as a natural event. Instead, it has generally been viewed as resulting from the attack of some demonic power or death god: in Etruscan sepulchral art a fearsome being called Charun strikes the deathblow, and medieval Christian art depicted the skeletal figure of Death with a dart. In many mythologies death is represented as resulting from some primordial mischance. According to Christian theology, death entered the world through the original sin committed by Adam and Eve, the progenitors of mankind.

Human substance and nature. The conception of death in most religions is closely related to the particular view held about the constitution of human nature. Two major traditions of interpretation have provided the basic assumptions of religious eschatologies and have often found expression in mortuary rituals and funerary practice. The more primitive of these interpretations has been based on an integralistic evaluation of human nature. Thus, the individual person has been conceived as a psychophysical organism, of which both the material and the nonmaterial constituents are essential in order to maintain a properly integrated personal existence. From such an evaluation it has followed that death is the fatal shattering of personal existence. Although some constituent element of the living person has been deemed to survive this disintegration, it has not been regarded as conserving the essential self or personality. The consequences of this estimate of human nature can be seen in the eschatologies of many religions. The ancient Mesopotamians, Hebrews, and Greeks, for example, thought that after death only a shadowy wraith descended to the realm of the dead, where it existed miserably in dust and darkness. Such a conception of man, in turn, has meant that, where the possibility of an effective afterlife has been envisaged, as in ancient Egyptian religion, Judaism, Zoroastrianism, Christianity, and Islām, the idea of a reconstitution or resurrection of the body has also been involved; for it has been deemed essential to restore the psychophysical complex of personality. In Egypt, most notably, provision was made for the eventual reconstitution in an elaborate mortuary ritual which included the mummification of the corpse to preserve it from disintegration.

Two dominant traditions concerning human nature

The alternative view of human nature may be termed dualistic. It conceives of the individual person as comprising an inner essential self or soul, which is nonmaterial, and a physical body. In many religions based on this view of human nature, the soul is regarded as being essentially immortal and as existing before the body was formed. Its incarnation in the body is interpreted as a penalty incurred for some primordial sin or error. At death the soul leaves the body, and its subsequent fate is determined by the manner in which it has fulfilled what the particular religion concerned has prescribed for the achievement of salvation. This view of human nature and destiny finds most notable expression in Hinduism and, in a subtly qualified sense, in Buddhism; it was also taught in such mystical cults and philosophies of the Greco-Roman world as Orphism (an ancient Greek mystical movement with a significant emphasis on death), Gnosticism (an early system of thought that viewed spirit as good and matter as evil), Hermeticism (a Hellenistic esoteric, occultic movement), and Manichaeism (a system of thought founded by Mani in ancient Iran).

The soul

Forms of survival. The conception of human nature held in any religion has, accordingly, determined the manner or mode in which postmortem survival has been envisaged. Where the body has been regarded as an essential constituent of personal existence, belief in a significant afterlife has inevitably entailed the idea of the reconstitution of the decomposed corpse and its resurrection to life. In turn, a dualistic conception of human nature, which regards the soul as intrinsically nonmaterial and immortal, envisages postmortem life in terms of the disembodied existence of the soul. This dualistic conception,

in many religions, has also involved the idea of rebirth or reincarnation. In Hinduism, Buddhism, and Orphism this idea has inspired a cyclical view of the time process and produced esoteric explanations of how the soul becomes reborn into a physical body, whether human or animal.

The ultimate destiny of the dead. Belief in postmortem survival has been productive also of a variety of images concerning the destiny of the dead. This imagery is closely related to the conception of man that is held in each religion. Thus, the magical resuscitation of the dead in ancient Egypt was designed to enable them to live forever in their well-furnished tombs; according to Christian and Islāmic belief, God will ultimately raise the dead with their physical bodies and assess their merits for eternal bliss in heaven or everlasting torment in hell; the Buddhist concept of Nirvāṇa (Enlightenment) is achieved only when the individual has eradicated all desire for existence in the empirical world.

PATTERNS OF MYTH AND SYMBOL

Geography of the afterlife. Inhumation naturally prompted the idea that the dead lived beneath the ground. The mortuary cults of many peoples indicate that the dead were imagined as actually residing in their tombs and able to receive the offerings of food and drink made to them; *e.g.,* some graves in ancient Crete and Ugarit (Ras Shamra) were equipped with pottery conduits, from the surface, for libations. Often, however, the grave has been thought of as an entrance to a vast, subterranean abode of the dead. In some religions this underworld has been conceived as an immense pit or cavern, dark and grim (*e.g.,* the Mesopotamian *kur-nu-gi-a* ["land of no return"], the Hebrew Sheol, the Greek Hades, and the Scandinavian Hel). Sometimes it is ruled by an awful monarch, such as the Mesopotamian god Nergal or the Greek god Hades, or Pluto, or the Yama of Hindu and Buddhist eschatology. According to the view of man's nature and destiny held in a particular religion, this underworld may be a gloomy, joyless place where the shades of all the dead merely survive, or it may be pictured as a place of awful torments where the damned suffer for their misdeeds. In those religions in which the underworld has been conceived as a place of postmortem retribution, the idea of a separate abode of the blessed dead became necessary. Such an abode has various locations. In most religions it is imagined as being in the sky or in a divine realm beyond the sky (*e.g.,* in Christianity, Gnosticism, Hinduism, and Buddhism); sometimes it has been conceived as the "Isles of the Blessed" (*e.g.,* in Greek and Celtic mythology) or as a beautiful garden or paradise, such as the *al-firdaws* of Islām. Christian eschatology, which came to conceive of both an immediate judgment and a final judgment, developed the idea of a purgatory, where the dead expiated their venial sins in readiness for the final judgment. Although the dead suffered there in a disembodied state, because their bodies would not be resurrected until the last day, the purifying flames of purgatory were usually regarded as burning in a physical sense, as Dante's *Purgatorio* vividly shows. The idea of a postmortem purgatory had been adumbrated in the 1st and 2nd centuries BC in Jewish apocalyptic literature (*1 Enoch* 22:9–13). The ten hells of Chinese Buddhist eschatology may be considered as purgatories, for in them the dead expiated their sins before being incarnated once more in this world.

The underworld and heaven

Means of approach to the afterworld. The idea that the dead had to make a journey to the otherworld, to which they belonged, finds expression in many religions. The oldest evidence occurs in the Egyptian Pyramid Texts (*c.* 2375–*c.* 2200 BC). The journey is conceived under various images. The dead pharaoh flies up to heaven to join the sun-god Re, in his solar boat, on his unceasing voyage across the sky, or he joins the circumpolar stars, known as the "Imperishable Ones," or he ascends a ladder to join the gods in heaven. Later Egyptian funerary texts depict the way to the next world as beset by awful perils: fearsome monsters, lakes of fire, gates that cannot be passed except by the use of magical formulas, and a sinister ferryman whose evil intent must be thwarted by magic. The idea of crossing water en route to the otherworld, which first appears in Egyptian eschatology, occurs in the eschatological topography of other religions, as was noted above. Many mythologies describe journeys to the underworld; they invariably reflect the fear felt for the grim experience that was believed to await the dead. Ancient Mesopotamian literature records the visit of the goddess Ishtar to the realm of the dead, the way to which was barred by gates. At each gate the goddess was deprived of some article of her attire, so that she was naked when she finally came before Ereshkigal, the queen of the underworld. It is possible that this successive stripping of the celestial goddess was meant to symbolize the stripping away of the attributes of life that the dead experienced as they descended into the "land of no return." An 8th-century Japanese text, the *Koji-ki,* tells of the first contact with death experienced by the primordial pair, Izanagi and Izanami. When his wife died, Izanagi descended to Yomi, the underworld of darkness, to bring her back. His request was granted by the gods of Yomi, on condition that he did not look at her in the underworld. Impatiently he struck a light and was horrified to see her as a decomposed corpse. He fled in terror and disgust. Blocking the entrance to Yomi with a great rock, he then sought desperately to purify himself from the contagion of death.

The journey after death

Such myths doubtless reflect an instinctive feeling that death works an awful change in those who experience it. The dead cease to belong to the world of the living and become uncanny and dangerous: hence, their departure to the world of the dead must be expedited. To assist that grim journey, various aids have been provided. Thus, on some Egyptian coffins of the 11th dynasty, a plan of the "Two Ways" to the underworld was painted, and from the New Kingdom period (*c.* 1567–1085 BC), copies of the Book of the Dead, containing spells for dealing with perils encountered en route, were placed in the tombs. Orphic communities in southern Italy and Crete provided their dead with directions about the next world by inscribing them on gold laminae deposited in the graves. Advice about dying was given to medieval Christians in a book entitled *Ars moriendi* ("The Art of Dying") and to Tibetan Buddhists in the *Bardo Thödol* ("Book of the Dead"). Chinese Buddhists were informed in popular prints of what to expect as they passed after death through the ten hells to their next incarnation. More practical equipment for the journey to the next world was provided for the Greek and Roman dead: in addition to the money to pay Charon for their passage across the Styx, they were provided with honey cakes for Cerberus, the fearsome dog that guarded the entrance to Hades.

Forms of final determination. Those religions that have taught the possibility of a happy afterlife have also devised forms of postmortem testing of merit for eternal bliss. Ancient Egypt has the distinction of conceiving of a judgment of the dead of an essentially moral kind. This conception finds graphic expression in the vignettes that illustrate the Book of the Dead. The heart of the deceased is represented as being weighed against the symbol of Maat (Truth) in the presence of Osiris, the god of the dead. A monster named Am-mut (Eater of the Dead) awaits an adverse verdict. The judgment of the dead as conceived in other religions (*e.g.,* Christianity, Islām, Zoroastrianism, Orphism) is basically a test of orthodoxy or ritual status, although moral qualities were included to varying degrees. The Last Judgment, as presented in Jewish apocalyptic literature, was essentially a vindication of Israel against its Gentile oppressors. Religions that held no promise of a significant afterlife (*e.g.,* those of ancient Mesopotamia and classical Greece) had no place for a judgment of the dead.

Judgment of the dead

DEATH AND FUNERARY RITES AND CUSTOMS

Before and at death. The process of dying and the moment of death have been regarded as occasions of the gravest crisis in many religions. The dying must be especially prepared for the awful experience. In China, for example, the head of a dying person was shaved, his body was washed and his nails pared, and he was placed in a sitting position to facilitate the exit of the soul. After the death, relatives and friends called the soul to return, possibly to make certain whether its departure from the body

was definitive. Muslim custom decrees that the dying be placed facing the holy city of Mecca. In Catholic Christianity, great care is devoted to preparing for a "good death." The dying person makes his last confession to a priest and receives absolution; then he is anointed with consecrated oil: the rite is known as "anointing of the sick" (formerly called extreme unction). According to medieval Christian belief, the last moments of life were the most critical, for demons lurked about the deathbed ready to seize the unprepared soul as it emerged with the last breath.

By courtesy of the Bibliotheque Nationale, Paris



Muslim gentleman's funeral. Relatives, wearing mourning bands, look on as the body, wrapped in a seamless shroud, is entombed on its side facing Mecca. Illustration from *Maqāmāt* of al-Ḥarīrī, painted by Yahya ibn Mahmūd al-Wāsiṭī, Baghdad, 1237. In the Bibliothèque Nationale, Paris (MS. Arabe 5847).

Modes of preparation of the corpse and attendant rites. After death, it has been the universal custom to prepare the corpse for final disposal. Generally, this preparation has included its washing and dressing in special garments and sometimes its public exposure. In some religions this preparation is accompanied by rites designed to protect the deceased from demonic attack; sometimes the purpose of the rites has been to guard the living from the contagion of death or the malice of the dead; for it has often been believed that the soul continues to remain about the body until burial or cremation. The most elaborate known preparation of the dead took place in ancient Egypt. Because the Egyptians believed that the body was essential for a proper afterlife, a complex process of ritual embalmment was established. This process was intended not only to preserve the corpse from physical disintegration but also to reanimate it. The rites were based upon the belief that, because the dead body of the god Osiris had been preserved from decomposition and raised to life again by the gods, the magical assimilation of a dead person to Osiris and the ritual enacting of what the gods had done would achieve a similar miracle of resurrection. One of the most significant of these ritual transactions was the "opening of the mouth," which was designed to restore to the mummified body its ability to see, breathe, and take nourishment.

Mummification in cruder forms has been practiced elsewhere (notably in Peru), but not with the same complex motives as in Egypt. The preparation of the corpse has also frequently included the placing on or in it of magical amulets; these were variously intended to protect or vitalize the corpse. Evidence found in tombs of the Shang dynasty (*c.* 1766–*c.* 1122 BC) suggests that the Chinese placed life-prolonging substances, such as jade, in the orifices of the corpse. Crosses or crucifixes are frequently placed upon the Christian dead, and sometimes in the Middle Ages the consecrated bread of the Eucharist (the Lord's Supper) was buried with the body. It has also been a Christian custom to furnish a dead priest with a chalice and paten, the instruments of his sacerdotal office.

Modes of disposal of the corpse and attendant rites. The form of the disposal of the dead most generally used throughout the world in both the past and present has been burial in the ground. The practice of inhumation (burial) started in the Paleolithic era, doubtless as the most natural and simplest way of disposal. Whether it was then prompted by any esoteric motive, such as the return to the womb of Mother Earth, as has been suggested, cannot be proved. Among some later peoples, who have believed that primordial man was formed out of earth, it may have been deemed appropriate that the dead should be buried—the idea found classical expression in the divine pronouncement to Adam, recorded in Genesis 3:19: "You are dust, and to dust you shall return." There is evidence that in ancient Crete the dead were believed to serve a great goddess, who was the source of fertility and life in the world above and who nourished and protected the dead in the earth beneath. Inhumation

The mode of burial has varied greatly. Sometimes the body has been laid directly in the earth, with or without clothes and funerary equipment. It may be placed in either an extended or crouched position: the latter posture seems to have been more usual in prehistoric burials. Sometimes evidence of a traditional orientation of the corpse in the grave can be distinguished, which may relate to the direction in which the land of the dead was thought to lie. The use of coffins of various substances dates from the early 3rd millennium BC in Sumer and Egypt. Intended probably at first to protect and add dignity to the corpse, coffins became important adjuncts in the mortuary rituals of many religions. Their ritual use is most notable in ancient Egypt, where the mummies of important persons were often enclosed in several human-shaped coffins and then deposited in large, rectangular wooden coffins or stone sarcophagi. The interiors and exteriors of these coffins were used for the inscription of magical texts and symbols. Sarcophagi, elaborately carved with mythological scenes of mortuary significance, became fashionable among the wealthier classes of Greco-Roman society. Similar sarcophagi, carved with Christian scenes, came into use among Christians in the 4th and 5th centuries and afford rich iconographic evidence of the contemporary Christian attitude to death.

In the ancient Near East, the construction of stone tombs began in the 3rd millennium BC and inaugurated a tradition of funerary architecture that has produced such diverse monuments as the pyramids of Egypt, the Tāj Mahal, and the mausoleum of Lenin in Red Square, Moscow. The tomb was originally intended to house and protect the dead. In Egypt it was furnished to meet the needs of its magically resuscitated inmate, sometimes even to the provision of toilet facilities. Among many peoples, the belief that the dead actually dwelt in their tombs has caused the tombs of certain holy persons to become shrines, which thousands visit to seek for miracles of healing or to earn religious merit; notable examples of such centres of pilgrimage are the tombs of St. Peter in Rome, of Muḥammad at Medinah, and, in ancient times, the tomb of Imhotep at Ṣaqqārah, in Egypt.

The disposal of the corpse has been, universally, a ritual occasion of varying degrees of complexity and religious concern. Basically, the funeral consists of conveying the deceased from his home to the place of burial or cremation. This act of transportation has generally been made into a procession of mourners who lament the deceased, and it has often afforded an opportunity of advertising his wealth, status, or achievements. Many depictions of ancient Egyptian funerary processions graphically portray the basic pattern: the embalmed body of the deceased is borne on an ornate sledge, on which sit two mourning women. A priest precedes the bier, pouring libations and burning incense. In the cortege are groups of male mourners and lamenting women, and servants carry the funerary furniture, which indicates the wealth of the dead man. Ancient Roman funerary processions were notable for the parade of ancestors' death masks. In Islāmic countries, The funeral



Hindu-animist cremation in Bali, Indonesia. Bodies are hidden inside gilded papier-mâché cattle to confuse evil spirits.
Ewing Krainin—Stockpile

friends carry the corpse on an open bier, generally followed by women relatives, lamenting with disheveled hair, and hired mourners. After a service in the mosque, the body is interred with its right side toward Mecca. In Hinduism the funeral procession is made to the place of cremation. It is preceded by a man carrying a firebrand kindled at the domestic hearth; a goat is sometimes sacrificed en route, and the mourners circumambulate the corpse, which is carried on a bier. Cremation is a ritual act, governed by careful prescriptions. The widow crouches by the pyre, on which in ancient times she sometimes died. After cremation, the remains are gathered and often deposited in sacred rivers.

Christian ritual

Christian funerary ritual reached its fullest development in medieval Catholicism and was closely related to doctrinal belief, especially that concerning purgatory. Hence, the funerary ceremonies were invested with a sombre character that found visible expression in the use of black vestments and candles of unbleached wax and the solemn tolling of the church bell. The rites consisted of five distinctive episodes. The corpse was carried (in a coffin if one could be afforded) to the church in a doleful cortege of clergy and mourners, with the intoning of psalms and the purificatory use of incense. The coffin was deposited in the church and covered with a black pall, and the Office of the Dead was recited or sung, with the constant repetition of the petition: "Eternal rest grant unto him, O Lord, and let perpetual light shine upon him." Next, requiem mass was said or sung, with the sacrifice offered for the repose of the soul of the deceased. After the mass followed the "Absolution" of the dead person, in which the coffin was solemnly perfumed with incense and sprinkled with holy water. The corpse was then carried to consecrated ground and buried, while appropriate prayers were recited by the officiating priest. Changes in these rites, including the use of white vestments and the recitation of prayers emphasizing the notions of hope and joy, were introduced into the Catholic liturgy only following the second Vatican Council (1962–65).

Human sacrifice

In some societies the burial of the dead has been accompanied by human sacrifice, with the intention either to propitiate the spirit of the deceased or to provide him with companions or servants in the next world. A classic instance of such propitiatory sacrifice occurs in Homer's *Iliad* (xxiii:175–177): 12 young Trojans were slaughtered and burned on the funeral pyre of the Greek hero Patroclus. The royal graves excavated at the Sumerian city of Ur, dating *c.* 2700 BC, revealed that retinues of servants and soldiers had been buried with their royal masters. Evidence of a similar Chinese practice has been found in Shang-dynasty graves (12th to 11th centuries BC) at An-yang. In ancient Egypt models of servants, placed in tombs, were designed to be magically animated to serve their masters in the afterlife. A particular type of these models, known as an *ushabti* ("answerer"), was inscribed with chapter VI of the Book of the Dead, commanding it to answer for the deceased owner if he were required to do service in the next world.

Dismemberment of the body

The custom has also existed among some peoples of dismembering the body for burial or subsequently disinterring the bones for storage in some form. There is Paleolithic evidence of a cult of skulls, which suggests that the rest of the body was not ritually buried. The Egyptians removed the viscera, which were preserved separately in four canopic jars. The Romans observed the curious rite of the *os resectum:* after cremation a severed finger joint was buried, probably as a symbol of an earlier custom of inhumation. In medieval Europe the heart and sometimes the intestines of important persons were buried in separate places: *e.g.,* the body of William the Conqueror was buried in St. Étienne at Caen, but his heart was left to Rouen Cathedral and his entrails for interment in the church of Chalus. To be noted also is the Zoroastrian and Parsi custom of exposing corpses on dakhmas ("towers of silence") to be devoured by birds of prey, thus to avoid polluting earth or air by burial or cremation.

The alternative use of inhumation or cremation for the disposal of the corpse cannot be interpreted as generally denoting a difference of view about the fate of the dead. In India, cremation was indeed connected with the fire god Agni, but cremation does not necessarily indicate that the soul was thus freed to ascend to the sky. Burial has been the more general practice, whether the abode of the dead be located under the earth or in the heavens.

Post-funerary rites and customs. Funerary rites do not usually terminate with the disposal of the corpse either by burial or cremation. Post-funerary ceremonies and customs may continue for varying periods; they have generally had two not necessarily mutually exclusive motives: to mourn the dead and to purify the mourners. The mourning of the dead, especially by near relatives, has taken many forms. The wearing of old or colourless dress, either black or white, the shaving of the hair or letting it grow long and unkempt, and abstention from amusements have all been common practice. The meaning of such action seems evident: grief felt for the loss of a dear relative or friend naturally expresses itself in forms of self-denial. But the purpose may sometimes have been intended to divert the ill humour of the dead from those who still enjoyed life in this world.

Purification of the mourners

The purification of mourners has been the other powerful motive in much post-funerary action. Death being regarded as baleful, all who came in contact with it were contaminated thereby. Consequently, among many peoples, various forms of purification have been prescribed, chiefly bathing and fumigation. Parsis are especially intent also on cleansing the room in which the death occurred and all articles that had contact with the dead body.

In some post-funerary rituals, dancing and athletic contests have had a place. The dancing seems to have been inspired by various but generally obscure motives. There is some evidence that Egyptian mortuary dances were intended to generate a vitalizing potency that would benefit the dead. Dances among other peoples suggest the purpose of warding off the (evil) spirits of the dead. Funeral games would seem to have been, in essence, prophylactic assertions of vitalizing energy in the presence of death. It has been suggested that the funeral games of the Etruscans, which involved the shedding of blood, had also a sacrificial significance.

Another widespread funerary custom has been the funeral banquet, which might be held in the presence of the corpse before burial or in the tomb-chapel (in ancient Rome) or on the return of the mourners to the home of the deceased. The purpose behind these meals is not clear, but they seem originally to have been of a ritual character. Two curious instances of mortuary eating may be mentioned in this connection. There was an old Welsh

custom of "sin eating": food and drink were handed across the corpse to a man who undertook thereby to ingest the sins of the deceased. In Bavaria, *Leichennudeln*, or "corpse cakes," were placed upon the dead body before baking. By consuming these cakes, the kinsmen were supposed to absorb the virtues and abilities of their deceased relatives.

Chastisement of the Tomb

A remarkable post-funerary custom has been observed in Islām; it is known as the Chastisement of the Tomb. It is believed that, on the night following the burial, two angels, Munkar and Nakir, enter the tomb. They question the deceased about his faith. If his answers are correct, the angels open a door in the side of the tomb for him to pass to repose in paradise. If the deceased fails his grisly interrogation, he is terribly beaten by the angels, and his torment continues until the end of the world and the final judgment. In preparation for this awful examination the roof of the tomb is constructed to enable the deceased to sit up; and, immediately after burial, a man known as a *fiqī* (or *faqih*) is employed to instruct the dead in the right answers.

CULTS AND MEMORIALS OF THE DEAD

Commemorative rites and services. The attitude of the living toward the dead has also been conditioned by the particular belief held about the human nature and destiny. Where death is regarded as the virtual extinction of the personality, the dead should logically have no more importance beyond that which their memory might stir in those who knew them. Even in the negative eschatologies of ancient Mesopotamia and Greece, however, the dead were thought of as still existent and capable of malevolent action if food offerings were not made to them. In those religions that have envisaged a more positive afterlife, the tendance of the dead has been developed in varying ways. In Egypt, it led to the building and endowment of mortuary temples or chapels, in which portrait images preserved the memory of the dead and offerings of food and drink were regularly made. In China, an elaborate ancestor cult flourished. The ancestral shrine contained tablets, inscribed with the names of ancestors, which were revered and before which offerings were made. The number of tablets displayed in the shrine was determined by the social status of the family. When the tablet of a newly deceased member was added to the collection, the oldest tablet was deposited in a chest containing still older ones: offerings to the remoter ancestors were made collectively at longer intervals. In India, three generations of deceased ancestors are venerated at the monthly *śraddha* festival, at which mortuary offerings were made.

The Christian cult of the dead found early expression in the catacombs, where mural paintings and inscriptions record the names of those buried there and the hopes of eternal peace and felicity that inspired them. Special chapels were made where the bodies of martyrs were entombed, and the anniversaries of their martyrdoms were commemorated by the celebration of the Eucharist (the Lord's Supper). The development of cults of martyrs and other saints in the medieval church centred on the veneration of their relics, which were often divided among several churches. The introduction of the doctrine of purgatory profoundly affected the postmortem care devoted to the ordinary dead. It was believed that the offering of the sacrifice of the mass could alleviate the sufferings of departed souls in purgatory. Consequently, the celebration of masses for the dead proliferated, and wealthy Christians endowed monasteries or chantry chapels where masses were said regularly for the repose of their own souls or those of their relatives. Prayers for the well-being of the dead have an important place in Mahāyāna Buddhism, and so-called "masses for the dead" were celebrated by Chinese Buddhists, influenced originally perhaps by the practice of the Nestorian Christians, who entered China in the 7th century AD.

Periodic commemorations

In many religions, in addition to private cults of the dead, periodic commemorations of the dead have been kept. The oldest of the Hindu sacred texts, the Rigveda (Ṛgveda), records the practice of the ancient Aryan invaders of India. The sacred beverage called *soma* was set out on "the sacred grass," and the ancestors were invited to ascend from their subterranean abode to partake of it and to bless their pious descendants. A similar ceremony, called the Anthesteria, was held in ancient Athens. On the day concerned, the souls of the dead (*kēres*) were believed to leave their tombs and revisit their former homes, where food was prepared for them. At sundown they were solemnly dismissed to the underworld with the formula: "out, *kēres*, the Anthesteria is ended." Buddhist China kept a Feast of Wandering Souls each year, designed to help unfortunate souls suffering in the next world. The Christian All Souls Day, on November 2, which follows directly after All Saints Day, commemorates all the ordinary dead: requiem masses are celebrated for their repose, and in many Catholic countries relatives visit the graves and place lighted candles on them. After World War I the public commemoration of the fallen was instituted on November 11, the day of the armistice in 1918, in many of the countries concerned: the memory of the dead was solemnly recalled in a two-minute silence during the ceremony. The body of an unknown soldier, killed in the fighting, was also buried in the capital cities of many countries and has become the accepted focus of national reverence and devotion.

Images of the dead

Cult of the dead. Among many peoples it has been the custom to preserve the memory of the dead by images of them placed upon their graves or tombs, usually with some accompanying inscription recording their names and often their achievements. This sepulchral iconography began in Egypt, the portrait statue of King Djoser (second king of the 3rd dynasty [*c.* 2686–*c.* 2613 BC]), found in the *serdab* (worship chamber; from the Arabic word for cellar) of the Step Pyramid being the oldest known example. The Egyptian images, however, had a magical purpose: they not only recorded the features of the deceased but also provided a locus for his *ka*, the mysterious entity that constituted an essential element of the personality. The sculptured gravestones of classical Athens deserve special notice, for they are among the noblest products of funerary art. They are expressive of a restrained grief for those who had departed to the virtual extinction of Hades. The deceased are often shown performing some familiar act for the last time. The inscriptions are very brief and usually record only the name and parentage; sometimes the word farewell is added. Etruscan mortuary art is characterized by the effigy of the deceased, sometimes with his wife, represented as reclining on the cover of the funerary casket. These images are obviously careful portraits, but whether they had some magical use as substitute bodies or are only commemorative is unknown. Roman funerary images seem to have been essentially commemorative, as were those of Palmyra.

Christianity has provided the richest legacy of funerary monuments. In the catacomb art of the 4th and 5th centuries, the deceased was sometimes depicted on the plaster covering of the niche in which his body was laid. From the early Middle Ages onward, the more affluent dead were represented in sculptured effigy or engraved in outline on stone or brass. In this tomb iconography, they are shown in a variety of postures: lying, kneeling, seated, standing, and sometimes on horseback. They are generally presented in the dress appropriate to their office or social standing: kings wear crowns, knights their armour; bishops are in copes and mitres and ladies in the fashionable attire of the day. This iconography is patently commemorative of the appearance in life, the achievements, and the status of the persons concerned. In the later Middle Ages, however, there was a remarkable innovation in this funerary art, which was designed to emphasize the horror and degradation of death. In what are known as memento mori tombs, below the effigies of the deceased as they were in life, there were placed effigies of their naked decaying corpses or skeletons. Such tomb sculpture reflected a contemporary obsession with the corruption of death.

PSYCHOLOGICAL AND SOCIOLOGICAL ASPECTS OF DEATH

The Paleolithic burials reveal that the pattern of man's reaction to the fact and phenomena of death has been set from the dawn of culture. Unlike the other animals, man has been unable to ignore the mysterious cessation

of activity and lapse of consciousness that cause his body to decay and befall members of his own kind. Death has, accordingly, constituted a problem for man, and he has felt impelled to take special action to cope with it. The pattern of his reaction has been twofold: confronted with the deaths of his companions, he has recognized an obligation to attend to their needs as he has conceived them, believing that they continued to exist in some form, either in the grave or in an underworld to which the grave gave access. But man's concern with death has not been confined to his tendance of the dead; for in the deaths of his fellows he has seen a presage of his own demise. This anticipation on the part of the living of the experience of dying has been a factor of immense psychological and social import. It is essentially a human characteristic; it stems from a consciousness of time, of which the immense cultural significance is only now beginning to be properly evaluated.

Awareness of time and anticipation of death

Awareness of time in its three categories of past, present, and future has decisively contributed to man's success in the struggle for existence. For it has enabled him to draw upon past experience in the present to anticipate future needs. Thus, from the making of the first stone tools to the complex structure of his modern technological civilization, man has sought by planning to render himself economically secure and to improve the standard of his living. But his time consciousness, which has made this immense achievement possible, is an ambivalent endowment. For, although it has enabled man to win economic security, it has also made him acutely aware of his own mortality and the inevitability of his own demise. Hence, his anticipation of death presents him with a profound emotional challenge, unknown to other species. The repercussions of this challenge can be traced in almost every aspect of his social and cultural life; but it is in his religions that man's reaction to death finds its most significant expression. All religion is concerned with postmortem security—with linking mortal man to an eternal realm—whether it be achieved by ritual magic, divine assistance, or mystic enlightenment.

MODERN NOTIONS OF DEATH

Continuation of traditional responses. Religious rites and customs continue to be practiced, because of conservatism, long after the ideas and beliefs that originally inspired them may be forgotten or abandoned. This is particularly true with regard to rites and customs pertaining to death. It is difficult to assess to what extent in the more sophisticated societies of the modern world the traditional eschatologies are still effectively held. Although a general skepticism obviously manifests itself toward the medieval imagery of death and judgment, of purgatory, heaven, and hell, modern modes of thinking have not lessened the mystery of death and its impact on the emotions. Indeed, in modern society, where expectation of life has been prolonged and standards of living raised, the negation of death is probably felt more keenly and also more hopelessly than in any other age.

Avowed secular inattention and unconcern. The reaction to death most apparent today among those having no effective religious faith is that of seeking to treat it as a disagreeable happening that must be dealt with as quickly and unobtrusively as possible. Funerals are no longer elaborately organized, mourning attire is rarely worn, and graveyards are landscaped, thus discreetly removing the earlier memorials of death. The increasing use of cremation facilitates this disposition to reduce the social intrusion of death and banish the traditional grave as a reminder of human mortality.

Rites and customs among secular materialists. It is significant, however, that, even where secularist principles are consciously professed, the dead are rarely disposed of without some semblance of ceremony. A deeply rooted feeling prompts most people to treat a dead human body with a respect that is not felt for a dead animal. It is significant that Communists make pilgrimages to the graves of Lenin and Marx; and, in the modern State of Israel, great effort is being made to record in the shrine of Yad va-Shem the names of those who died in the persecution of the Jews in Germany during the Nazi regime of Adolf Hitler in the 1930s and '40s and, if possible, to bring their ashes there. In America, morticians strive to preserve the features of the dead as did the embalmers of ancient Egypt, though for somewhat different motives. Finally, as further evidence of modern preoccupation with death, it may be noted that, in Western society, Spiritualism witnesses to a widespread desire to have communication with the dead, and recently, in England, there has even been a recrudescence of necromancy. (S.G.F.B.)

Purification rites and customs

Purification rites and customs, based on concepts of purity and pollution, are found in all known cultures and religions, both ancient and modern, preliterate and sophisticated. Assuming a wide variety of types and forms, these rites and customs attempt to re-establish lost purity or to create a higher degree of purity in relation to the Sacred or Holy (the transcendent realm) or the sociocultural realm.

CONCEPTS OF PURITY AND POLLUTION

General concepts. Every culture has an idea, in one form or another, that the inner essence of man can be either pure or defiled. This idea presupposes a general view of man in which his active or vitalizing forces, the energies that stimulate and regulate his optimum individual and social functioning, are distinguished from his body, on the one hand, and his mental or spiritual faculties, on the other. These energies are believed to be disturbed or "polluted" by certain contacts or experiences that have consequences for a person's entire system, including both the physical and the mental aspects. Furthermore, the natural elements, animals and plants, the supernatural, and even certain aspects of technology may be viewed as operating on similar energies of their own; they too may therefore be subject to the disturbing effects of pollution. Because lost purity can be re-established only by ritual and also because purity is often a precondition for the performance of rituals of many kinds, anthropologists refer to this general field of cultural phenomena as "ritual purity" and "ritual pollution."

The rituals for re-establishing lost purity, or for creating a higher degree of purity, take many different forms in the various contemporary and historical cultures for which information is available. Some purification rituals involve one or two simple gestures, such as washing the hands or body, changing the clothes, fumigating the person or object with incense, reciting a prayer or an incantation, anointing the person or object with some ritually pure substance. Some involve ordeals, including blood-letting, vomiting, and beating, which have a purgative effect. Some work on the scapegoat principle, in which the impurities are ritually transferred onto an animal, or even in some cases (as among the ancient Greeks) onto another human being; the animal or human scapegoat is then run out of town and/or killed, or at least killed symbolically. Many purification rites are very complex and incorporate several different types of purifying actions.

Ritual purity and pollution are matters of general social concern because pollution, it is believed, may spread from one individual or object to other members of society. Each culture defines what is pure and impure—and the consequences of purity and pollution—differently from every other culture, although there is considerable cross-cultural overlapping on certain beliefs. Cultures also vary greatly in the extent to which purity and pollution are pervasive concerns: Hinduism, Judaism, and certain tribal groups such as the Lovedu of the African Transvaal or the Yurok of northern California in the United States seem highly pollution-conscious, whereas among other peoples pollution concerns are relatively isolated and occasional. Even within the so-called pollution-conscious cultures, attitudes toward the cultural regulations may vary considerably: the Yurok, on the one hand, are said to consider their purification rituals to be rather a nuisance, albeit necessary for the success of their economic endeavours; but Hindus, on the other hand, seem to incorporate and embrace more

fully the many regulations and rituals concerning purity prescribed in their belief and social systems.

The transmission and symptoms of pollution

Pollution is most commonly transmitted by physical contact or proximity, although it may also spread by means of kinship ties or co-residence in an area in which pollution has occurred. Because purity and pollution are inner states (though there usually are outer or observable symptoms of pollution), the defiled man—or artifact, temple, or natural phenomenon—may at first show no outward features of his inner corruption. Eventually, however, the effects of pollution will make themselves known; the appearance of a symptom or disaster that is culturally defined as a consequence of pollution, for example, may be the first indication that a defiling contact has occurred. Common cross-cultural, human symptoms of pollution include: skin disease, physical deformity, insanity and feebleminded-ness, sterility, and barrenness. Nature also may become barren as a result of pollution; but, on the other hand, the natural elements and magical or supernatural forces may run amok as a result of pollution.

In general, the vital energies of man, nature, or the supernatural, as a consequence of pollution, may become either hypoactive or hyperactive. The vital energies may tend to operate in a manner that leads toward decline, loss of potencies and fertility, and death; they may also, however, tend to operate in an opposite manner that leads toward excess, increase and perversion of potencies, and chaos. Both of these tendencies presumably contrast with the tendencies of a state of purity, although the properties, symptoms, or consequences of purity rarely are explicitly defined in cultural ideologies, in contrast to the wealth of detail elaborated on the consequences of pollution.

On the whole, purity seems to be equated with whatever a culture considers to be the most advantageous mode of being and functioning for achieving the paramount ideals of that culture. Thus, throughout most Asian religions (*e.g.,* Hinduism, Buddhism, Jainism, and Taoism), purity is equated with calmness (physical, mental, and emotional equilibrium) in keeping with the ideal goal—at least for religious adepts—of achieving spiritual transcendence or liberation. In contrast to such Asian religions, groups whose dominant cultural orientation is pragmatic and this-worldly, such as the Yurok, often equate the state of purity with vigour and quickness of mind and body.

Purity and pollution in relation to religious concepts. Concepts of purity and pollution may tend to merge with several concepts of religion: the sacred, sin, and the forces of evil.

Consequences of contact with the sacred or the polluted

Pollution and the sacred. The consequences of contact with both the sacred (the transcendent realm and objects infused with transcendent qualities) and the polluted may be identical, although the reasons for the consequences in the two cases are quite different. The dangers of contact with the sacred may arise from the belief that the gods are offended by pollution; they will punish a person who defiles a sacred precinct or object (for example, in Buddhism and many other religions, a menstruating woman who enters a temple or shrine). The gods may even punish an entire village or tribe for such an offense. To come into contact with the sacred is also viewed as dangerous because the sacred is highly powerful or "charged" with energy; thus, one must be properly strengthened (usually by purification) for the encounter. If one is not thus strengthened, he will be overwhelmed. Although contact with the sacred may have negative consequences for a person, this is not because the sacred is polluting. On the other hand, the dangers of encounter with a polluted person (*e.g.,* an "untouchable" in India) or object (*e.g.,* feces, in most cultures) arise directly from the pollution that passes from that person or object to oneself.

Pollution and sin. Purity and pollution beliefs may become incorporated into a religious morality system in which pollution becomes a type of sin and an offense against God or the moral order, and purity becomes a moral or spiritual virtue. Thus, for example, in the Old Testament, the pollution of birth must not only be cleansed by symbolic or ritual gestures; it must also be atoned for as a cultic sin that offends the sacred precincts of the Lord. In general, the more universalistic religions—Christianity, Buddhism, and Islām—seem to de-emphasize true pollution concerns, and to subsume them within their frameworks of moral and religious beliefs. Both the Qur'ān of Islām and the New Testament of Christianity show a sharp decrease in rules of specific pollution avoidances (*e.g.,* fewer food prohibitions) compared to the Old Testament. Similarly, the sacred texts of Buddhism stress the unimportance of specific avoidances and rituals (in implicit contrast to the multiple and detailed purity regulations of Hinduism) and the necessity for cultivating one's spiritual and moral development instead.

Pollution and the forces of evil. Ideas of pollution are often closely associated with beliefs in demons, sorcerers, and witches. All of the latter may be viewed, in part, as personifications of the powers of pollution. People in polluted states are believed to be dangerous not only to others because they may spread their pollution, but they themselves are often thought to be in danger of attack by demons, who are attracted by the defiled person's impurities (see also DOCTRINES AND DOGMAS; OCCULTISM).

CATEGORIES AND THEORIES OF POLLUTION AND IMPURITY

Categories of pollution and impurity. Four major categories of what various religions and societies have regarded as polluting or inherently impure phenomena may be distinguished. Virtually any type of impure person, object, or state (as defined in various cultures) may be assigned to one of these four categories, or may be shown to have symbolic associations with one (or sometimes with several) of these four sets.

Universal belief that bodily functions involve pollution

Physiological processes. The functions of the human body are, for the most part, universally considered polluting, although all functions are not considered polluting in all cultures. The intensity with which the various processes are abhorred also varies from culture to culture. The list of polluting organic processes and things includes menstruation, sexual intercourse, birth, illness, death, and all bodily excretions and exuviae (urine, feces, saliva, sweat, vomit, blood, menstrual blood, semen, nasal and oral mucous, and hair and nail cuttings). Associated with this category symbolically may be various persons, animals, natural objects, sense-related objects, and professions: women in general (because they menstruate), pregnant women, prostitutes, and widows (the latter because of their additional association with death); pigs, dogs, and other scavengers because they eat or associate with excrement and garbage; carrion-eating animals because of their association with death; leftover food, because it has come in contact with saliva via the fingers or utensils that have touched the mouth, or because it may visually resemble vomit or the undigested contents of the stomach; pungent vegetables or spices (such as garlic, onions, and leeks) and strong-smelling meats or fish because they cause foul breath odours; food in general because of its ultimate state as excrement; certain professions because their members are required to handle corpses or bodily exuviae; and things associated with lowness—the entire body below the navel, the feet, the hem of the garment, the floor or ground—because most bodily excretions derive from the lower part of the body.

Violence and associated processes. A second major category of polluting phenomena involves violence and all associated aspects. This entire category may be reduced to beliefs in the polluting nature of blood and death, but the extensive development of various ideas connected with violence pollution merit its being classified as a separate category. Violence pollution involves a wide variety of activities: murder, hunting, warfare, physical fights, quarrelling, cursing or speech that is considered foul, aggressive language, lying, and various aggressive human passions (*e.g.,* greed, anger, and hatred). Various phenomena considered polluting in one culture or another may be placed in this category because of their symbolic associations with violence: Satan, demons, witches, predatory ghosts, and the practice of black magic; alcohol because it stimulates aggressive impulses; carnivorous, predatory, and aggressive animals; meat because of the act of slaughtering the animal; certain professions because their members manufacture weapons or kill or fight for a living.

Anomalies. The third major category includes strange, unusual, or unclassifiable phenomena: (1) certain events of nature (*e.g.,* comets or lunar or solar eclipses); (2) unusual deaths (*e.g.,* death by lightning); (3) unusual births (*e.g.,* twins or other multiple births, breech deliveries, miscarriages, or stillbirths); (4) physical deformities, especially sexual deformities (*e.g.,* monorchids [men possessed of only one testicle], hermaphrodites, or eunuchs); (5) speech defects and voices appropriate to the opposite sex; (6) unusual developmental sequences (*e.g.,* children who cut their upper teeth before their lower); (7) anomalous animals or types of plants that have features of several species; (8) viscous substances that seem neither solid nor liquid; (9) persons in liminal (threshold or transitional) categories or states (*e.g.,* persons undergoing initiation rites, strangers, or captives); (10) persons not considered fully in control of their faculties (*e.g.,* children, drunken persons, the insane, or the mentally or physically handicapped, such as cretins); and (11) perversions of social relationships, especially sexual, that a culture generally considers to be normal (*e.g.,* adultery, homosexuality, bestiality, incest, births of children to unwed parents or as a result of adulterous relationships, or the breaking of vows of celibacy by monks or nuns). That pollution results from a confusion of classification rules may explain beliefs that certain objects must not be mixed lest pollution result. The Old Testament prohibition (also found in certain African groups) that meat and milk should not be mixed with one another or the prohibition in the Vedas (ancient Hindu scriptures) against carrying water and fire at the same time are examples of attempts to maintain classificatory purification rules (see below under *Dietary laws and food customs*).

Pollution in relation to caste professions and ethnic structures

Social classifications: classes and castes. The belief that the lower castes pollute the upper castes has been explicit in India, where a true caste system has existed. These lower castes, to some extent, are considered polluting because they engage in professions that have been or are associated with the physiological processes or with violence. Many lower caste occupations (*e.g.,* pottery making or basket weaving), however, do not have such associations, and thus the categorization of pollution attached to all lower castes cannot be so explained. Outside true caste systems, there are de facto systems of racial or ethnic hierarchy, in which certain races or ethnic groups are considered to be inherently lower than others. In most such systems, the notion that the lower groups pollute the higher is not stated explicitly in terms of pollution; the language of racial or ethnic prejudices in such systems, however, is often strongly reminiscent of pollution concepts—*e.g.,* that the lower groups are "dirty," have peculiar bodily odours, engage in sexual promiscuity or perversions, are "animals," or are violent and dangerous. Relations between the dominant race or ethnic group and the subordinate one often resemble the relations between upper and lower castes in India. In such social systems, eating together and intermarriage generally are not condoned, and segregated neighbourhoods and public facilities to maintain minimal physical contact are encouraged by law or custom.

Theories of pollution and impurity. Though these four major categories indicate the great diversity of phenomena considered polluting cross-culturally, no one culture considers every item noted in these categories as polluting. Furthermore, within a single culture, not every item considered polluting is necessarily polluting to every member of the society, because the connotation of pollution often is dependent upon the occasion and on the status of a person. The pollution of death, for example, may be confined to those who have actual contact with the corpse, the immediate family of the deceased, certain categories of kinsmen, or all members of the village in which the death has occurred.

The rules dictating avoidance of certain groups or individuals because of the threat of pollution may be seen as means that a society has at its disposal for emphasizing its important social categories. Thus, in the case of death, if relatives on the father's side but not on the mother's side are considered polluted by the death, it may be theorized that this is one of the society's ways of emphasizing the greater social significance of the patrilateral relatives in the kinship system. Sociologists and anthropologists, on the one hand, tend to stress such social implications of pollution rules. On the other hand, some psychologists, philosophers, and theologians are more interested in explaining what there is about polluting events and processes (*e.g.,* death and menstruation) in themselves that would result in their being considered polluting in so many cultures.

Psychoanalytic and social classification theories

Two general theories have been proposed in relation to these emphases or questions. The first theory derives primarily from psychoanalytic theories developed by Sigmund Freud in which the quest for sexual, excretory, and aggressive pleasures are viewed as instinctual drives in man that are repressed or greatly limited in the socialization of the individual. Hence, because many of the phenomena viewed as polluting cross-culturally are related to these concerns, pollution fears are interpreted as projections or symbolizations of these repressed instincts. The second theory that attempts to explain the specific content of pollution-belief systems (as opposed to the social effects of those beliefs) maintains that, in a very broad sense, things are considered polluting by virtue of their relationship to cultural classification. This theory holds that everything considered polluting in any culture either is anomalous in relation to basic cultural categories or is positioned at the extremities—*i.e.,* the margins—of major conditions or situations of individual or social existence. Birth and death, for example, are at the margins of an individual's life, and the lower castes are at the margin of society.

Both of these theories, however, contain certain problems that may be resolved by subsuming them under a more general theory. The theory derived from psychological considerations is regarded by many scholars as being too narrow in scope because it ignores many types of pollution data; the theory based upon cultural classification, because it is capable of such broad interpretation, loses its coherence as a theory. A more general view incorporates these two theories within a single more fundamental one based on denial. Thus, pollution fears might be interpreted as symbolizations of any material that is denied full expression—psychologically, culturally, or socially. The Freudian theory, emphasizing the psychoanalytic notion of repression of instinctual drives, thus becomes significant in interpreting the first two categories—physiological processes and aggression (*i.e.,* violent emotional processes). The classification theory, which emphasizes cultural attempts to ignore or suppress phenomena that do not fit its cognitive-classification schemes, then becomes significant in interpreting the third category of polluting things—anomalies, unusual occurrences or types of persons, and "mixings." To account for the fourth category, involving the fear of lower castes, classes, and ethnic groups as polluting, the sociopolitical notion of oppression may thus be introduced. All these concepts—repression, suppression, and oppression—are related to the notion of something or someone being forcibly prevented from expression; that is, of being under some sort of pressure. This idea suggests why polluting things are viewed as threatening and not simply as interesting peculiarities of the world, because things under pressure are volatile, liable to escape, or capable of erupting at any moment.

TYPES OF PURIFICATION RITES

Occasions and symbolism of purification rites. Purification rites are required whenever there has been some kind of polluting contact. In addition, cultures may institutionalize regular, periodic purification rituals on the general principle that pollution occurs all the time. Important changes of status or quests for special or sacred status may be viewed as progressions from lesser to greater states of purity, and such changes or quests thus entail rites that promote the anticipated progressions. Purification is invariably required before any contact with the sacred. Purification also is generally considered necessary after any kind of traffic with the demonic forces and black magic, because these contacts with the nether realm are viewed as polluting experiences. Purification rites also may be required before undertaking a major endeavour in order

Rites before or after contact with the sacred

to ensure the participant's success and a right relationship with the special powers involved in the project.

Though every culture has rituals to rectify unavoidable pollution, prescriptions of avoidance, abstention, separation, and seclusion are utilized to minimize contacts with polluting persons, objects, or places. Seclusion devices, which confine the very pure or the very polluted within an enclosed area away from other members of society, include menstrual huts, nuptial huts, and birth huts. Initiates are generally confined to special houses or isolated from the community by living for certain required periods of time in the bush or forest. Priests often withdraw to the inner rooms of temples to prepare for or to participate in contacts with the sacred; monks and nuns confine themselves or are confined to monasteries in order to remain undefiled by the world, among other reasons. Seclusion or containment may also be symbolically effected by the use of veils or by the drawing of circles or other enclosures around the object in question. Under the general heading of segregation, groups of different grades of purity may retire to their respective parts of a town when their periods of contact with other members of their community are completed for the day. Men may have special houses for their esoteric activities from which women are excluded. Impure persons may be required to cook over a separate fire; persons of different grades of purity often are not permitted to eat together, to sleep under the same roof or in the same room, and, almost universally, to marry or have sexual relations with one another. Finally, complete abstention, for a fixed period of time, from such polluting activities as sex, eating, and other sensuous indulgences is a significant aspect of purification processes in many societies around the world.

Classification of purification rites. Various kinds of avoidances and abstentions represent the passive aspect of purification. The active aspect consists of the purification rites themselves. Such rites may be classified according to the principle on which they operate.

Rites of cleansing

The removal of pollution. Based on the analogy of cleansing outer dirt or stains by means of bathing or washing in everyday life, purification of man's inner state of being is almost universally believed to be effected by rituals involving various forms of washing. The polluted individual might be required to swim or bathe in the sea, a river, a pond, or special tank. Bathing in swift-flowing streams is often considered especially effective because the rapidly flowing water not only removes the impurities but carries them away. A polluted person might wash his entire body with water or only certain parts of the body that represent the body or person as a whole—rinsing or cleaning the mouth by other means is common. Water may be poured, sprinkled, thrown, or blown upon a polluted person or object. Simply touching water is a purifying gesture in the Vedas; gazing at it is considered purificatory in Sri Lanka (Ceylon). In the absence of water various kinds of moist substances may be used—clay, mud, wet herbs, or plants. The Qur'ān (the Islāmic sacred scriptures) directs desert dwellers and travellers to rub themselves with high clean soil because of the scarcity of water. In cultures in which saliva is not considered polluting, expectorating or breathing on something may be viewed as purificatory gestures.

Other modes of purification based on the analogy of cleansing outer dirt include: the use of wind or aeration to blow or carry away the impurities; sweeping a house or certain area of the ground or brushing the polluted person or object, often with a brush made of fibres from a symbolically pure source; scraping the surface of a polluted object or utensil; shaving and cutting the hair and nails; removing clothing and washing it or destroying it; and putting on clean or new clothes.

The expulsion of pollution. Based on the analogy of expelling internal physical poisoning or corruption, a second category of purification rites involves the actions of expelling, ejecting, purging, or drawing out the pollution from the defiled person or object. The use of purgatives in purification rites to induce vomiting is not uncommon. Sweat baths and steam baths are believed to bring the impurities out of the person as symbolized by the emerging sweat. Some purification rites involve bloodletting in order to drain out impurities. The use of salt in some rites may be based on the fact that salt has drawing or draining properties. In corporate acts of expelling pollution, an entire community may purge itself of a polluted individual in its midst by excommunicating him and forcing him to leave the religious group, caste, tribe, or area.

The transfer of pollution. Closely connected with the practice of drawing pollution from the defiled person or object is the notion that pollution may be transferred from a person or community to another object that is either immune to pollution itself or that can be discarded or destroyed. The most dramatic rites embodying this principle are scapegoat ceremonies in which pollution is transferred to an animal or person by either touching, bathing with, or simply pronouncing the pollution transferred to the scapegoat. The scapegoat is then run out of town or killed, actually or symbolically. The victim may further be made into an offering or sacrifice to the gods on the general ritual principle of keeping the gods satisfied. In the classic scapegoat ceremony of the Old Testament, as noted in Leviticus, chapter 16, the animal—called Azazel (a desert demon)—was simply released to wander the wilderness; in Bali (in Indonesia) birds act as scapegoats and are then released to fly away.

Ceremonies involving scapegoats

Less dramatically, pollution may be transferred to a relatively worthless talisman (charm). Some talismans are regarded as convenient because they are disposable and of little value; after they have served their purposes in specific situations they are thrown away. In Bali a three-month-old child is purified by transferring his impurities to a chicken; this chicken may then become his pet and continue to absorb the pollutions to which the child is exposed. It may never be killed or eaten, and when it dies it is buried with respect.

The destruction of pollution. Pollution is also believed to be eliminated by destroying the polluted object. The killing of the scapegoat belongs to this general category; more dramatically, a severely polluted person may himself be killed rather than being allowed the opportunity to transfer his impurities onto a more dispensable animal or object. The execution of a polluted person or a scapegoat animal often takes the form of drowning, choking, suffocating, or clubbing so that the pollution might not escape with a flow of blood. Polluted metal objects may be melted down; polluted fires are extinguished; polluted clothing, utensils, and other items are torn, broken, and often buried.

Forms and objects of pollution-destroying rites

The most common means of destroying pollution is by burning the polluted objects. Fire is a most efficient destroyer; when the flame no longer exists there is virtually nothing left of the objects. Fire is generally conceived, however, as having more positive purifying properties, not only destroying pollution but creating purity.

The transformation of pollution into purity. Fire is perhaps one of the most symbolically complex phenomena in the history of human culture. It renders raw meats and vegetables into cooked and edible food, base minerals into useful and durable metals, and porous dirt and clay into watertight pottery. It destroys the forests and brushlands, but its ashes make the earth fertile and productive. Fire is thus viewed as a powerful transformer of the negative to the positive. Because of such properties, fire is commonly found in purification rites throughout the world. Polluted persons may be required to walk around, jump over, or jump through fire. Polluted items may be singed, fumigated, or smoked. The widespread use of incense smoke in purification rites is based on the transforming powers of fire, as well as on the additional purificatory powers of sweet smells. Polluted persons or things may be rubbed with ashes or soot, and polluted objects may be boiled, subject to the double purificatory powers of fire and water. Exposure to sun and to intense heat are also regarded as practices falling into this same general category. The extinguishing of old fires in temples and villages and the kindling of new ones are common practices after a death or as part of annual renewal and purification ceremonies. Alchemic experiments, which attempt to purify mineral substances and turn them into gold, involve boiling or melting down the solution or elements over pure and

intense heat and then recrystallizing them in newer and higher forms (see also OCCULTISM; TAOISM).

The introduction of purity. In addition to the cleansing, purging, destruction, and transformation of pollution, most purification rites involve the positive introduction of purity. Many phenomena are considered inherently pure; ingestion of, or contact with, or simply exposure to such phenomena is believed to bring purity to the object of the ritual.

Phenomena viewed as intrinsically pure

Objects, activities, or persons commonly considered to have intrinsic purity cross-culturally include: fire; water; sweet smells created by flowers, fragrant plants and herbs, perfumes, fragrant oils, or incense; milk, ghee, and other dairy products; white objects; earth in its natural form; sacred objects (*e.g.*, relics) and sacred personages (*e.g.*, priests); the recitation of spells, incantations, and names of gods; magical amulets and stones; gold and, in one culture or another, silver, bronze, jade, and crystal; virgins; the right as opposed to the left side of things in many cultures (*e.g.*, the Abaluyia of Kenya); morning, sunshine, and daylight as opposed to darkness; whole or perfect objects, including circles and wheels and perfect numbers—*e.g.*, the number nine (because the digits of any of its square products always add up to nine) or four (because quaternity is viewed as perfection); and physically perfect specimens of their species. In addition, cultures idiosyncratically define certain things as pure because of special cultural associations: cow dung and cow urine are pure in Hinduism because of the sacredness of the cow; dogs are considered to be pure in Zoroastrianism (a religion founded in the 6th century BC by the Iranian prophet Zoroaster) because as scavengers they purify the world for everyone else (most cultures view dogs as impure because of their scavenging habits); and all cool things are considered pure among the Lovedu of the Transvaal because pollution is associated with heat.

Other purification rites. Purification practices in which pollution is introduced qua pollution in order to achieve purity are also found in various religions and cultures. These rather paradoxical practices work on several different principles. The use of garlic, sulfur, or an amulet made of impure materials apparently operates on the principle of like attracting like; the impure amulet draws the impurity encountered in some situation toward itself, thus preventing it from polluting the wearer of the charm. Another set of practices apparently works on an inoculation principle—a baby, a magical implement, or a special work area may be briefly exposed to menstrual blood, for example, to protect it against future pollution from the same kind of item. A third group of such paradoxical practices, found primarily in Asian religions, involves immersing oneself in what is viewed as utter pollution, either by meditating on foul things or by actually keeping oneself permanently unclean, in order to achieve transcendence over pollution. Ordeals, mutilations, and blood rituals in general may also be regarded as fitting the transcendence pattern.

Use of pollution to achieve purity

In highly developed and elaborated systems of thought, purity and pollution meet and merge. Buddhist monks are considered to be extremely pure, yet they are directed to make their robes from cemetery cloths, and beds or litters used in funerals may be donated to their monasteries. Buddhist relics with great purifying power are often composed of bits of hair, nails, and bones (albeit of the Buddha or other great saints); in Sri Lanka the word (*dhātu*) for such relics is the same as the word for semen. Monks and nuns of Jainism (an Indian religion founded by Mahavira in the 6th century BC) are ordered not to bathe and under no circumstances to clean their teeth. In Hinduism, if a Brahmin (a member of the highest caste) enters a street of the untouchables (outcastes), he is polluted, but the whole street also falls prey to disease, famine, and sterility. In a Burmese folktale, an alchemist became discouraged with his experiments and threw his alchemic stone into a latrine pit; on contact with the excrement, the stone achieved purity—thus indicating that contacts with pollution may bring about purity.

Many rituals considered to effect purification do not utilize any of the specific purifying techniques outlined above. They simply make use of techniques believed to have generalized ritual efficacy, no matter what the disorder. Thus, some purification rites involve reversals, especially reversals of roles between men and women, on the general principle that they represent a return to chaos and then a change back to order. Another widely practiced ritual principle involving the symbolism of reversal is that of death and rebirth; man and the world, with all their disorders, are symbolically put to death and then symbolically renewed in a purer and better state. Because blood is associated with both life and death, the use of blood in purification rites is often central to the symbolic renewal process. Nearly all rituals involve the reading or reciting of spells, texts, or prayers that have a generalized efficacy over negative forces, and in many cases purification may be accomplished by these means without any further symbolization of cleansing or a recreation of purity. When pollution becomes one of many possible offenses against the gods, purification may be accomplished simply by making sacrifices or offerings to the gods. Pollution often becomes identified with immoral or sinful behaviour and in such instances purification may be effected by punishment of the offender, by the offender's spiritual atonement, or by acts of penance and virtue, such as giving alms. Purity also may become identified with the struggle against the demonic forces, and in this transcendent dimension purification is effected in rites of exorcism or in rites that placate the demons. The use of weapons in purification rites is often based on a symbolic battle with the forces of evil; the use of firecrackers in some purification rites is viewed as a means of frightening away the demons; the use of curses, abuse, ridicule, and ribaldry in purification rites among the ancient Greeks, for example, was regarded as forms of protection against the demons. Some purification rites involving blood are structured in terms of giving demons what they want in order to turn away their polluting presences (see also *Sacrifice*, above).

Ritual renewal and sacrificial rites

EXAMPLES OF PURIFICATION RITES

Most full-scale purification rites combine several of the principles outlined above. A few of the immense number of complex purification rites in the religions and cultures of the world follow.

Rite for purifying a cured leper in ancient Judaism. In the Old Testament purification rites for a person who has been cured of leprosy, as described in Leviticus, the leper and the priest meet outside the camp, and the priest examines the man to ascertain that he is cured. The priest then calls for two live, clean birds, cedar wood, a scarlet item, and hyssop (an aromatic herb). One of the birds is killed in an earthen vessel over running water. The live bird and the other ingredients are then dipped in the blood of the dead bird and used to sprinkle blood seven times upon the leper while the priest pronounces him clean. The live bird is then allowed to fly away. The leper washes his clothes, shaves off all his hair, and washes himself, after which he is allowed to enter the camp, although he must remain outdoors for seven days. On the seventh day he once again shaves off his hair, including his eyebrows, and washes his clothes and body. On the eighth day he goes to the temple to make various offerings to the Lord. The priest then takes some of the blood of one of the offerings and places it on the man's right ear, thumb, and large toe of the right foot, after which he does the same with some oil that is being offered, also pouring some oil on the man's head. The sacrifices are then offered to the Lord upon the altar, thus completing the required ritual: "the priest shall make atonement for him, and he shall be clean."

The Navajo sweat-emetic rite. The Navajo sweat-emetic rite is part of most major Navajo ceremonials for curing illness or rectifying other ritual disturbances. It is specifically viewed as a rite of purification.

A ritual hut is prepared with sand paintings, and a fire is then built. A procession of patients, led by the chanter, enters the hut and circumambulates the fire, pausing at each of the four directions to sing an appropriate chant. In some cases there is fire jumping; the men are required to jump over the fire, and the women to walk as close to it as possible. The audience then enters, with men and

women sitting in segregated groups. The chanter heats wooden pokers in the fire and applies them to himself, mainly on the legs, and then to all the patients. Basins in front of each patient are filled with the emetic formula, the fire procession is repeated, and the emetic is then drunk. Everyone is expected to vomit; if they do not, it is regarded as inauspicious. Vomiting is done into receptacles containing sand, and the contents of these receptacles may then be sprinkled with ashes from the pokers. A bullroarer (a heavy stone on a string that produces a deep roaring sound when whirled) is sounded outside six times and then brought in and applied to the patients. The audience leaves the hogan (hut) in procession, this time led by assistants who carry out the basins with their contents. The contents of the basins are deposited neatly in a row outside the hogan and allowed to be dispersed by the natural elements. The patients, however, remain inside the hogan, perspiring in the heat. Later, the audience re-enters; the fire is broken up and extinguished, and all remnants of it are removed to a place near the basin area. The chanter sprinkles all present with a medicinal lotion and then fumigates everyone with incense. All then leave in procession and dress outside.

The Zoroastrian "Great Purification" rite. The "Great Purification" rite (*baresnum*) of Zoroastrianism originally was intended for purification from serious polluting contacts, especially for corpse bearers after contact with death. The rite was later pre-empted for initiation into the priesthood, or for attaining higher statuses within it.

In preparation for this rite a priest seeks a piece of ground regarded as clean (*i.e.,* dry and unfrequented by men or animals). He then cuts down any trees located on the area selected. Nine pits are dug in a certain arrangement; furrows are drawn around three, then around six, and finally around all nine pits. Thereafter, the whole area is covered with sand. After these activities have been completed, the priest stands outside the outer furrow, and the subject requiring purification advances to the first pit and is told to recite praises to the "Purity of Thought." The priest, holding a stick with nine knots and with a spoon fastened to the end, uses the spoon to pour consecrated cow's urine (*gomez*) upon the hands of the subject, who washes his hands with the urine three times. He then washes his entire body with *gomez,* progressing from the head down to the feet. The pollution is said to leave the toes in the form of a foul-smelling fly. After the one seeking purification has washed himself with *gomez,* the priest recites purifying formulas. This process is repeated at each of the first six pits; at a prescribed distance from the seventh pit, the subject sits down and rubs himself 15 times with sand, making sure that he is completely dry. At the seventh pit he washes his body once, from head to toe, with water; at the eighth pit he does this twice and at the ninth pit three times. His body is then fumigated with the smoke of fragrant wood, after which he dresses in clean clothes. In certain versions of the ceremony, a dog is presented to the candidate, who, after each washing at each pit, must touch the left ear of the dog with his left hand. At the end of the ceremony the candidate is required to recite the following formula: "The Evil Spirit of pollution is put down. The head and the body have become purified. The soul has been purified. The dog is holy, the priest is holy."

The candidate then retires to a house and is required to have no contact with fire, water, cultivated land, trees, cattle, men, or women. On the fourth, seventh, and tenth days he again bathes with *gomez* and then with water. After the final bath he is considered "perfectly purified."

POLLUTION BELIEFS IN MODERN SOCIETY

Pollution beliefs and fears occur in modern society as well as in any other, although they are not systematized and usually not understood as such. Racism and other forms of prejudice apparently play upon pollution fears. Of less serious consequence are such notions that warts result from masturbation (traditionally considered a polluting or impure practice in conventional Western societies), that there is something dangerous or polluting in intercourse with menstruating women, and that (as in a New York state law) men and women should not have their hair cut or beauty services performed in the same room. Physiological processes (*e.g.,* urination and other forms of elimination) are often viewed with disgust, and as a result many modern notions of sanitation are based on not entirely rational principles. The highly developed mortuary profession (especially in Western countries) protects persons in contact with death not only from grief but probably from pollution fears as well. On the whole, however, there are fewer pollution beliefs in modern society than in traditional societies. This trend may be attributed in part to the assimilation of these beliefs into moral and religious concepts. (S.B.O.)

Dietary laws and food customs

Like all other biologically and physically necessary things and acts, food and eating are always surrounded by social regulations that prescribe what may or may not be ingested under particular social conditions. These prescriptions and proscriptions are sometimes religious; often they are secular; frequently, they are both. This section surveys the variety of laws and customs pertaining to food materials and the art of eating in human societies from earliest times to the present. It will be seen that behaviour in respect to food—whether religious, secular, or both—is institutionalized behaviour and is not separate or apart from organizations of social relations.

By an institution is meant here a stable grouping of persons whose activities are designed to meet specific challenges or problems, whose behaviour is governed by implicit or explicit rules and expectations of each other and who regularly use special paraphernalia and symbols in these activities. Social institutions are the frames within which man spends every living moment. This survey explores the institutional contexts in which dietary laws and food customs are cast in different societies; the attempt will also be made to show that customs surrounding food are among the principal means by which human groups maintain their distinctiveness and help provide their members with a sense of identity.

Theoretical approaches to food customs

Other points of view about food customs cover a wide range. What may be labelled an ecological approach suggests that food taboos among a group's members prevent over-utilization of particular foods to maintain a stable equilibrium in the habitat. Recently, investigators of such customs have been exploring the hypothesis that they provide an adaptive distribution of protein and other nutrients so that these may be evenly distributed in a group over a long period instead of being consumed at one time of the year. The ecological approach also suggests that many food taboos are directed against women to maintain a low population level; this seems to be an adaptive necessity in groups at the lowest technological levels, in which there is a precarious balance between population and available resources.

There are also psychological approaches to food customs. Psychoanalytic writers speculate that food symbolizes sexuality or identity because it is the first mode of contact between an infant and its mother. This point of view is most clearly exemplified in ideas that attitudes toward food, established early in life, tend to shape attitudes toward money and other forms of wealth and retentiveness or generosity. According to Claude Lévi-Strauss, a French anthropologist, the categories represented in food taboos enable people to order their perceptions of the world in accordance with the principle of polarities that govern the structure of the mind. Thus, they aid in maintaining such dichotomies as those between nature and culture or between man and animal.

NATURE AND SIGNIFICANCE

There are no universal food customs or dietary laws. Nor are food customs and dietary laws confined to either preliterate ("primitive") or advanced cultures; such regulations are found at all stages of development. Nevertheless, different types of regulations in respect to food are characteristic of groups at different levels of cultural or socio-technological development.

Each society has attached symbolic value to different

foods. These symbolizations define what may or may not be eaten and what is desirable to eat at different times and in different places. In most cases, such cultural values bear little relationship to nutritive factors. As a result, they often seem difficult to explain. Moreover, dietary customs and laws are resistant to rational argument and change. For example, experts from health and nutritional agencies find it difficult to persuade mothers to give cow's milk to children in societies in which it is looked upon as undesirable. Such customs and laws also prevent people from adopting alternative foods during periods of shortage. During and after World War II, some Indians refused to eat Western wheat and rioted and died rather than accept it.

Food as a material expression of social relationships. Cutting across dietary laws and customs is the more general association of food and drink with those social interactions that are considered important by the group. In many societies the phrase "We eat together" is used by a man to describe his friendly relationship with another from a distant village, suggesting that even though they are not neighbours or kinsmen they trust one another and refrain from practicing sorcery against each other. Among the Nyakyusa of Tanzania, "for conversation to flow merrily and discussion to be profound, there must be . . . 'the wherewithal for good fellowship,' that is, food and drink—and very great stress is laid on sharing these." In Old Testament times, almost every pact, or covenant, was sealed with a common meal; eating together made the parties as though members of the same family or clan. Conversely, refusal to eat with someone was a mark of anger and a symbol of ruptured fellowship. Eating salt with one's companions meant that one was bound to them in loyalty; references to this are found in the New Testament.

The common meal

Such sentiments, however, are not confined to tribal or ancient cultures. In Israeli kibbutzim (communal settlements), the communal dining room is a keystone institution, and commensality is one of the hallmarks of kibbutz life. The decline of communal eating and the increasing frequency of refrigerators, cooking paraphernalia, and private dining in kibbutz homes is regarded by some observers as a sign of the imminent demise of kibbutzim. In many U.S. communes there is a single facility for cooking and dining. Dinners must be taken communally; private dining is taken as a signal that one is ready to leave the group.

The provision of food and drink, if not actual feasting, is characteristic of rites of passage—*i.e.*, rites marking events such as birth, initiation ceremonies, marriage, and death—in almost all traditional cultures and in some modern nontraditional groups as well. The reason for this is that these events are regarded as being of importance not only to the individual and his family but also to the group as a whole because each event bears in one way or other on the group's continuity.

Furthermore, food and drink are almost universally associated with hospitality. In most cultures, there are explicit or implicit rules that food or drink be offered to guests, and there are usually standards prescribing which foods and drinks are appropriate. Reciprocally, these sets of rules also assert that guests are obligated to accept proffered food and drink and that failure to do so is insulting. In many societies, there are prescribed ritual exchanges of food when friends meet. Food is thus one of the most widespread material expressions of social relationships in human society.

Regulations about the quantity of food and drink consumed. It is extraordinarily rare for cultures to condone gluttony, the conventional exaggerations of the eating behaviour of the ancient Roman elite notwithstanding. Most people cannot afford to be gluttons. There are more examples of the other extreme, asceticism, though these too are infrequent.

A clear-cut example of gastronomic asceticism is provided by Indians of the U.S. Northeast, such as the Micmac, Montagnais, and Ojibwa. It was an ideal among them to eat sparingly. Preparation for this attitude began in early childhood with short fasts of a day or two, culminating in the puberty fast; the latter lasted about 10 days, during which time the child was isolated in a tiny wigwam without food or water. The puberty fast also had important religious significance. During the fast, the child had to supplicate the deities for a vision (easily induced under such conditions), which came in the form of a supernatural figure, usually in animal shape; this was to become his guardian spirit.

Rules pertaining to drink are even more varied. Tribal groups throughout the world (except in Oceania and most of North America) knew alcohol; in each case, this led to the adoption of rules concerning its use.

Alcoholic beverages

Although a high intake of alcohol always has physiological effects, people's comportment is determined more by what their society tells them is the way to behave when consuming alcohol than by its toxic effects. In many societies, drinking is an established part of the total round of social activities. Robert McC. Netting, a U.S. anthropologist, observed that the Kofyar of northern Nigeria "make, drink, talk, and think about beer." All social relations among them are accompanied by its consumption, and fines are levied in beer payments. Ostracism takes the form of exclusion from beer drinking; they "certainly believe that man's way to god is with beer in hand." Their beer, however, is weak in alcoholic content and is quite nutritious, and they rarely consume European beer and never distilled liquor. Among Central and South American peasants, men are allowed or required to drink themselves into a state of stupefaction during religious celebrations (fiestas); though this drinking is frequent and heavy, it does not appear to result in addiction. Representative of the other extreme are the Hopi and other Indian tribes of the U.S. Southwest who have banned all alcoholic beverages (and almost all narcotics), asserting that these substances threaten their way of life.

Most cultures, however, prescribe moderation in drinking. In ancient Mesopotamia, beer played an important role in temple services and in the economy; but the code of Hammurabi—the monument of law named after the king of Babylon—strictly regulated tavern keepers and servants (these places were supposed to be avoided by the social elite). Similar patterns obtained in ancient Egypt. The ancient Greeks sought to attribute their intellectual and material culture to the introduction of vine and olive growing. The use of wine was quite general in biblical times; it belonged to the category of indispensable provisions listed in the Old Testament in the Book of Judges (chapter 13) and the First Book of Samuel (chapters 16 and 25). Wine was no less important in New Testament times; in Revelation to John (chapter 6) it is said that only wine and oil are to be protected from the apocalyptic famine. Wine is also frequently used in biblical imagery. In both Testaments, however, wine is both praised and condemned.

Use of food in religion. The most widespread symbolic use of food is in connection with religious behaviour. In fact, eating and drinking are minimal elements in most religious behaviour and experience, whether in eating, sacrifice, or communion. According to many anthropologists, there are essentially two reasons for this. First, religion is one of the systems of thought and action by which the members of a group express their cohesiveness and identity. Implicitly or explicitly, the members of every cultural group assert that its unity and distinctiveness derive from the deity or deities associated with it. Religion is a tie that binds. But no symbolic activity in human society stands alone and without material representation. Like all other symbolizations of institutional relationships, those of religion must also have substantial form. Food and drink—and their ingestion—are among the most important substances of religion.

Reasons for symbolic use of food in religious behaviour

The second reason, closely related to the foregoing, is that one element of dogma in every religion is the definition of polluting, or supernaturally dangerous, objects or personal states. Just as there is no objective or scientific connection between the nutritive qualities of different foods and the symbolic values attached to them, there is no objective relationship between an object or a personal state and its definition as polluting. Cultures vary in the objects and states that are defined as defiling, such as saliva, sneezing, menstruation, killing an enemy in warfare, a corpse,

parturition, but cutting across these is the belief held in every religion that there are foods and drinks that are polluting or defiling.

As Mary Douglas, a British anthropologist, has suggested in her analysis of the religiously sanctioned food taboos in Leviticus (chapter 11) and Deuteronomy (chapter 14), *Purity and Danger,* concepts of pollution and defilement are among the means used by preliterate or tribal societies to maintain their separateness, boundedness, and exclusivity; thus, these concepts and rules contribute strongly to the sense of identity—the social badges—that people derive from participation in the institutions of their firmly bounded or encapsulated groups. More concretely, when a person proclaims his affiliation with and allegiance to a particular group that he regards as his self-contained universe and beyond whose margins he sees danger, threat, and alienation, he simultaneously invokes—explicitly or implicitly—the many badges of his social identity; these include the totem (*i.e.,* the emblem of a family or clan) that he may not eat, the foods that are regarded as defiling, the drinks that he must avoid, the sacred meals in which he participates, and the other rituals associated with his exclusive group. He thereby asserts his separateness from people in all other groups—usually referred to in pejorative terms—and his identification with the members of his own group. Food customs are not always formalized, however; they are sometimes cast in terms of preference. Americans, for example, unless they are members of ethnic or religious groups that have their own dietary laws, often shun the "exotic" foods of alien cultures; but these avoidances are not phrased in religious or other institutional terms.

LAWS AND CUSTOMS AT DIFFERENT STAGES OF SOCIAL DEVELOPMENT

Although there are dietary laws and customs in all societies, groups differ in this regard in two important ways: in the range or extent of foods that are defined as polluting or tabooed and in conceptualizations of the consequences resulting from violations of these laws and customs. In comparing societies, however, it must be remembered that the range of variability among them is so great that it would be necessary to list hundreds of societies and their customs to get a complete and detailed picture of their food customs and laws. For purposes of both economy and conceptual coherence it is necessary to group societies into levels, or stages, of social and technological development and to compare these; in this approach, individual societies are regarded as special or particular exemplary cases of the general class of the level of development in which the groups are found or classed.

Hunter-gatherers. The earliest cultural level that anthropologists know about is generally referred to as hunting-gathering. Hunter-gatherers are always nomadic, and they live in a variety of environments. Some, as in sub-Saharan Africa and India, are beneficent environments; others, such as those of the Arctic or North American deserts, are harsh and dangerous. Encampments of hunter-gatherers are usually small (generally fewer than 60 persons) and are constantly splitting up and recombining. An important rule among almost all hunter-gatherers is that every person physically present in a camp is automatically entitled to an equal share of meat brought into the group whether or not he has participated in the hunt; this rule does not usually extend to vegetables or fruits and nuts.

It may be thought that hunter-gatherers who live in habitats of scarcity and in which hunting is dangerous would try to make maximum use of all potentially available food; they are, however, also characterized by customs and beliefs that proscribe certain foods or at least limit their consumption. Many Alaskan Eskimo groups, for instance, make a sharp distinction between land and sea products; the Eskimo believe that products of the two spheres should be kept separate, maintaining that land and sea animals are repulsive to each other and should not be brought together. Thus, for example, before hunting caribou (a spring activity), a man must clean his body of all the seal grease that has accumulated during the winter; similarly, before whaling in April, the individual's body must be washed to get rid of the scent of caribou. Weapons used for hunting caribou should not be used at sea; implements used at sea, however, may be used to hunt caribou. If these rules are violated, the hunter or whaler will be unsuccessful in his food quest; the consequences of this, of course, can be dire.

Alaskan Eskimo groups

In addition, the Eskimo observe food taboos in connection with critical periods of the individual's life and development. Among the most outstanding of these are the food taboos that a woman is subject to for four or five days after giving birth. She may not eat raw meat or blood and is restricted to those foods that are believed to have beneficial effects on the child. For example, it is felt that she should eat ducks' wings to make her child a good runner or paddler. Because the Eskimo are often beset by food shortages, they sometimes have to eat forbidden foods. In such cases, there are several things that a person can do to neutralize the taboo. He first rubs the forbidden food over his body and then hangs the meat outside and allows it to drain. Another act that is regarded as particularly efficacious is to stuff a mitten into the collar of his parka with the hand side facing outward; it is believed that the harmful effects of the taboo food go into the mitten and travel away from him.

There are, of course, other food avoidances observed by the Eskimo, but these examples will suffice to illustrate the basic principles of dietary customs and laws among hunter-gatherers. First, the taboos are always thought to have magical consequences for the individual; observing them will assure health and strength, violating them will result in illness and weakness for the person or, in the case of a parturient mother, for her child. Second, food taboos are generally associated with critical periods during the life cycle, as in pregnancy, menses, illness, or dangerous hunts. Third—and this is true of almost all societies, not only those of hunter-gatherers—in every group's system of thought there are categories or types of foods that are regarded as dangerous, defiling, or undesirable. At first glance, these rules and customs seem arbitrary and capricious, but evidence is accumulating that there are rational elements in them. Although it would be difficult in the present stage of knowledge to apply this principle to every dietary taboo or custom in every society, it seems that prohibitions are placed on those foods that are the most difficult and dangerous to procure. Sometimes, however, these foods are also highly prized.

Corporate kin groups. With the development of corporate kin groups in social history, largely (but not exclusively) as an accompaniment of horticultural cultivation, a significant change occurred in the role of food in institutional life. Underlying the development of corporate kin groups was the development of the notion of exclusive rights to territory claimed by a group of kinsmen. This exclusive territoriality was probably designed, in large measure, to protect investments of time and effort in particular plots. The solidarity and sense of kin-group exclusiveness implicit in a corporate kin group grew out of kin-group ownership of the land and the individual's reliance on interhousehold cooperation in his productive activities. Such groups quickly evolve insignia, rules, and symbols that represent their ideals of exclusivity and inalienability of social relations; food plays an important role in this. Hence, taboos are thought to have consequences for the group as a whole rather than for the individual alone.

Territorial rights and initiatory rites

Another significant accompaniment of the development of corporate kin groups is the elaboration of initiatory rites, which mark an individual's transition from childhood to full membership in his community or kin group; they confer citizenship in the fullest sense of the term. Such events are celebrated by feasts, reciprocal exchanges of food, and food taboos, in addition to the ceremonial rituals themselves. Preparations for these feasts sometimes occupy the group for several months, especially when it is necessary to acquire from relatives and friends the animals that will be slaughtered and eaten, because it is rare for one family, or even one village, to own enough animals for a proper feast. They lay the groundwork for one of the basic rules of the group into which the individual is being initiated, namely, that the distribution of food (and

interhousehold cooperation in its acquisition) is one of the most significant ways in which he and the members of the group are knit together.

Feasting is also an integral element of religious assemblages and ritual in these societies, as are offerings to deities, whether spirits or ancestors. Because one of the main purposes of religious activity is to symbolize the solidarity of the group, food is used as a material representation of this cohesiveness. Additionally, it is believed in almost all tribal societies, whether or not they are characterized by corporate groupings, that all plant and animal foodstuffs are made available to man through the beneficence of the gods. Man's relationship with the deities in tribal societies is always, in part, an economic one involving the deities' provision of food. A gift from the gods must be balanced by a reciprocal gift to them from their adherents. In prayer, men thank their deities for these gifts; in sacrifice and offerings, they offer gifts to their deities.

Chiefdoms. The next major social and political developments in human history are the appearance of institutions in which political and economic power is exercised by a single person (or group) over many communities. Often referred to as chiefdoms by anthropologists, this development signalled a process evident today throughout the world, namely, the steady growth of centralized power and authority at the expense of local and autonomous groupings.

Economic powers of the chief

Political authority in chiefdoms is inseparable from economic power, including the right by rulers to exact tribute and taxation. One of the principal economic activities of the heads of chiefdoms is to stimulate the production of economic surpluses, which they then redistribute among their subjects on different types of occasions, such as feasts in the celebration of religious ceremonies and rites of passage of members of chiefly families, and during periods of famine. The accumulation of these surpluses requires conservation policies. Because techniques of food preservation were poorly developed in preliterate chiefdoms, the heads of chiefdoms often adopted the policy of placing taboos—often phrased in religious terms—on different crops or areas where food could be gathered or hunted, forbidding the consumption of such foods until the prohibitions were lifted. These taboos, however, were not exclusively for the purpose of conservation; they were also occasionally designed to underwrite higher standards of living for the chiefs themselves. For instance, in some Polynesian societies, as in Samoa, fishermen were required to obey a taboo that a portion of their catch must be given to the chief. The penalties for violating such taboos were supernaturally produced illness or other misfortunes.

Complex societies. As societies became increasingly complex, heterogeneous, and divided along lines of caste, class, and ethnic affiliation, their dietary customs became correspondingly less uniform because they mirrored these divisions and inequalities. Although these distinctive customs are almost always placed in the context of religious belief and practice, according to many anthropologists, the dietary observances in everyday behaviour are primarily shaped by economic and social considerations; moreover, observances at the village level rarely correspond directly to formal prescriptions and proscriptions.

Social stratification as basis for food customs

The dietary laws and customs of complex nations and of the world's major religions, which developed as institutional parts of complex nations, are always based on the prior assumption of social stratification, traditional privilege, and social, familial, and moral lines that cannot be crossed. Taboos and other regulations in connection with food are incompatible with the idea of an open society. Nevertheless, complex nations were characterized by caste organizations that, in almost all cases, religion helped to legitimate. Caste systems, in addition to their other characteristics, are supported by deeply felt fears of pollution or contamination as a result of unguarded contact of the more "pure" with those who are less "pure."

Although there is no doubt that the development of caste is linked to some form of occupational separation in a society, which, in turn, leads to the development of ideas concerning the separation of unclean persons from the ordinary or of the ordinary from the superpure, there is considerable controversy over the origins of caste systems. Regardless of the origins, however, the separation of castes is always mirrored in rules for eating that, when breached, represent a threat to the social order and to the individual's sense of identity. There is also a question among scholars whether or not caste is unique to India. Nevertheless, in Japan as well as India, eating together implies social and ritual equality, as it does in the United States, where, unlike Japan and India, food-related caste behaviour has not been institutionalized in religion (largely because of the U.S. history of religious freedom, which has promoted religious diversity). In India and Japan a person who cooks for another and serves his food must be equal or superior in rank to the recipient of the food; only in this way can the latter avoid pollution. By contrast, in the caste system of the United States before the civil-rights movement, a black might cook and serve food to, but not eat with, whites. Violation of these eating taboos constitutes defiance of caste, and observance of the etiquette is evidence of the acceptance of caste.

RULES AND CUSTOMS IN WORLD RELIGIONS

Judaism. Perhaps the best known illustration of the idea that the dietary laws and customs of a complex nation and its religion are based on the prior assumption of social stratification or, at least, of a sense of separateness, is provided by Judaism as spelled out in the Mosaic Law in the Old Testament books of Leviticus (chapter 11) and Deuteronomy (chapter 14). Prohibited foods may not be consumed in any form: all animals—and the products of animals—that do not chew the cud and do not have cloven hoofs (*e.g.*, pigs, horses); fish without fins and scales; the blood of any animal; shellfish (*e.g.*, clams, oysters, shrimp, crabs) and all other living creatures that creep; and those fowl enumerated in the Bible (*e.g.*, vultures, hawks, owls, herons). All foods outside these categories may be eaten.

Prohibited foods

Interpretation of Jewish laws. Mary Douglas has offered probably the most cogent and widely accepted interpretation of these laws in her book *Purity and Danger.* She suggests that these notions of defilement are rules of separation; they symbolize and help maintain the biblical notion of the separateness of the Hebrews from other societies. A central element in her interpretation is that each of the injunctions is prefaced by the command to be holy and that it is the distinction between holiness and abomination that enables these restrictions to make sense. "Holiness means keeping distinct the categories of creation. It therefore involves correct definition, discrimination, and order." The Mosaic dietary laws exemplify holiness in this sense. The ancient Hebrews were pastoralists, and cloven-hoofed and cud-chewing hoofed animals are proper food for such people; hence, Douglas maintains, they became part of the social order and were domesticated as slaves. Pigs and camels do not meet the criteria of animals that are fit for pastoralists. As a result, they are excluded from the realm of propriety. Douglas notes that there is remarkable consistency in Mosaic dietary laws. The Bible "allots to each element its proper kind of animal life. In the firmament two-legged fowls fly with wings. In the water scaly fish swim with fins. On the earth four-legged animals hop, jump, or walk. Any class of creatures which is not equipped for the right kind of locomotion in its element is contrary to holiness." People who eat food that is "out of place," as it were, such as four-footed creatures that fly, are themselves unclean and are prohibited from approaching the Temple.

There is, however, another dimension to Old Testament food customs. In addition to expressing their separateness as a nation—membership in which was ascribed by birthright—Israelite food customs also mirrored their internal divisions, which were castelike and were inherited. Though the rules of separation referred primarily to the priests, they also affected the rest of the population. The priest's inherent separateness from ordinary men was symbolized by the prescription that he must avoid uncleanness more than anyone else. He must not drink wine or strong drink, and he must wash his hands and feet before the Temple service. Explicit in Old Testament prescriptions is

that an offering sanctifies anyone who touches it; therefore, often the priests alone were permitted to consume it.

These rules symbolizing the priestly group's castelike separateness also validated a system of taxation benefitting them, couched in terms of offerings, sacrifice, first-fruit ceremonies, and tithes. The religious rationalization of taxation is illustrated in the Old Testament by the first-fruits ceremony. Fruit trees were said to live their own life, and they were to remain untrimmed for three years after they were planted. But their fruits could not be enjoyed immediately: God must be given his share in the first-fruit ceremonies. These first fruits represent the whole, and the entire power of the harvest—which is God's—is concentrated in them. Sacrifice is centred around the idea of the first-fruits offering. Its rationalization was that everything belonged to God; the central point in the sacrifice is the sanctification of the offering, surrendering it to God. Its most immediate purpose was to serve as a form of taxation to the priests; only they were considered holy enough to take possession of it.

Elaboration of the Jewish laws. After the exile of the Jews from Palestine following the conquest by Rome in the 1st century AD, a remarkable elaboration in their dietary laws occurred, probably as a result of the Jews' attempts to maintain their separateness from nations into whose midst they were thrust. Many customs evolved that have taken on the force of law for those Jews who have sought to maintain a traditional way of life. For example, the Bible does not prescribe ritual slaughter of animals, yet this practice has taken on the same compulsion as the taboo on pigs and camels; a permitted food (*e.g.*, cattle, chicken) that has not been ritually slaughtered is now regarded to be as defiling as pork. Similarly, one of the hallmarks of the Passover holiday in Judaism is the eschewal of all foods containing leaven, the consumption only of foods that have been designated as "kosher for Passover," and the use of special sets of utensils that have not been used during the rest of the year. But these, too, are postbiblical customs that have been given the force of law; the Bible prescribes nothing more than eating unleavened bread during the Passover season.

Shtetl communities

Further elaborations on the Mosaic Law in regard to food can be observed in the dietary customs of certain groups of modern Jews in their daily lives. In the pre-World War II eastern European Jewish community (or *shtetl*), behaviour in regard to food not only included the biblical prescriptions and proscriptions but, in many ways, resembled the behaviour of people in the corporate communities of tribal societies. The major life crises were celebrated by feasts or other uses of food. Wine and other foods were integral parts of circumcision ceremonies and of a boys' attainment of ritual majority (Bar Mitzwa). Weddings were also celebrated with huge feasts that required weeks, if not months, of preparation, and guests were seated at the wedding feast according to their social rank. Following the wedding celebration, grain was sprinkled on the couple's heads, apparently to promote fertility. Those who visited mourners were to eat hardboiled eggs or other circular food because roundness symbolizes mourning.

Aside from the daily requirements of following the Mosaic dietary laws, which apply to everyone, the heaviest burden for maintaining these observances falls on the women; their ritual and secular statuses are always inferior to the men's. It is the task of the housewife to be sure that meat and dairy foods are not mixed, that ritually slaughtered meat is not blemished, and that cooking equipment and dishes and utensils for meat and dairy are rigidly separated. The only personal states of ritual pollution relating to food in *shtetl* culture also refer only to women. For instance, a woman who has not been ritually cleansed after her menses must not make or touch pickles, wine, or beet soup. If she violates this customary rule, it is believed that these foods will spoil.

"Ḥasidic" Jews

A further illustration of the idea that dietary rules and customs are inextricably associated with the maintenance of group separateness is provided by one sect of Jews in the United States, those who refer to themselves as Ḥasidim (Pious Ones). These people live in self-contained enclaves; most of them are immigrants from the *shtetl*. In addition to preserving their distinctiveness from surrounding non-Jewish communities, they are equally devoted to preserving their distinctiveness vis-à-vis other Jews; no matter what their degrees of piety, the latter are regarded by Ḥasidim as nonreligious.

This is clearly reflected in their behaviour in regard to food. The Ḥasidim assert that the larger Jewish community (and its rabbis) do not meet Ḥasidic standards and qualifications in the manufacture, preparation, handling, and sale of food; even non-Ḥasidic ritual slaughterers are classed with assimilated Jews who do not observe dietary laws at all. Hence, their food products are regarded as forbidden, and Ḥasidim consider only their own products as permissible for consumption. Even neutral foods, such as vegetables, are defined as nonkosher if handled by a non-Ḥasid since there is always the suspicion that it may come into contact with nonkosher—and thus contaminating—matter. Thus, for instance, only milk that they designate as "Jewish" can be used; only noodles prepared by someone from the Ḥasidic community may be consumed because there is the suspicion that eggs with a drop of blood (which are forbidden) may have been used in the noodles' preparation; only approved sugar may be used; and even paper bags that hold food come under these restrictions because only a member of the community is above the suspicion that forbidden matter has been included in the glue that is used in manufacturing the bags.

The extremity of Ḥasidic strictures with regard to food has to be viewed in the context of their setting in the United States and not only in the light of their Jewish sources. The Ḥasidim regard the growing secularization of U.S. life as the greatest threat to the perpetuation of the ancient tradition of Judaism; their extremism is the wall they have erected to stave off this danger of threatened assimilation.

Black Muslim movement. Until relatively recently, the separatism of U.S. Negroes was underwritten by an intricate combination of law and custom. The attempt of the United States government to achieve an integration of blacks and whites in daily social, economic, and political life was viewed by some Negroes as a threat to their social identity. Ideologies designed to legitimate the maintenance of their social identity began to develop, especially after the desegregation decision of the Supreme Court in 1954, the most notable of which is known as the Nation of Islam (the Black Muslims). In their attempt to separate themselves from the larger aggregate of U.S. Negroes, as well as from the rest of U.S. society, the Black Muslims sought to develop a separate social identity by adopting a set of symbols to which they attached particular meanings. A person's membership in the group not only depended on assuming a Muslim name but also on eating certain foods and avoiding others, including alcohol and tobacco. Forbidden foods include meats and fish proscribed by the Bible and Qur'ān and also more than a dozen vegetables that were staples in the slave diet.

Islām. Islāmic dietary laws—as spelled out in the Qur'ān—also illustrate their relationship to the establishment of a sense of social identity and separateness. Muḥammad, the founder of Islām, was among other things a political leader who welded a nation out of the mutually warring tribes of Arabia. His religious ideology legitimated the unification of these autonomous tribes and his own paramount rule over them. The main religious tenets of Islām were derived from Judaism and early Christianity, and it is clear from the Qur'ān that Islām was intended to encompass all aspects of life.

Relationship with Judaism and Christianity

Muḥammad apparently knew more about Judaism than about Christianity, and many of his strictures in the Qur'ān were explicit in establishing distinctions between Arabs and Jews. This is evident in his dietary regulations, which borrow heavily from Mosaic Law. Specifically, Muḥammad proscribed for Muslims the flesh of animals that are found dead, blood, swine's flesh, and food that had been offered or sacrificed to idols. The most radical departure of Qur'ānic from Mosaic dietary laws was in connection with intoxicating beverages. Though Jews frown upon alcoholic beverages, they do not forbid them, and wine is

an important element in many rituals and feasts; Muḥammad, however, absolutely forbade any such beverages.

Specific departures from Mosaic and Christian dietary rules notwithstanding, Islām represents a more fundamental removal from all other major religions: what is polluting, forbidden, and enjoined for one person in Islām applies equally to all. Islām's sharpest contrast in this regard is to the religions of India. This difference is highlighted by the fact that Muslims of all social statuses in an Indian village eat freely with each other, worship in the same mosques, and participate in ceremonies together.

Christianity. Christianity did not develop elaborate dietary rules and customs. This probably grew out of the controversy between the Judaizing and Hellenizing branches of the church during the earliest years of Christianity over whether or not to observe Mosaic food laws. The Council of Jerusalem settled on the formula that meat offered to idols, blood, and things strangled must be abstained from, thus freeing the Gentiles in all other respects from the law. The apostle Paul's position on the matter, however, was that "nothing is unclean in itself"; and it was thus that the New Testament repudiated the entire body of laws of purity, especially those pertaining to food. Jesus is said to have declared that defilement could not be caused by any external agent. The apostle Peter's vision of the sheet lowered from heaven and containing all types of animals that the divine voice pronounced clean and fit for food provided the church with a mandate to abandon the Old Testament food laws.

Last Supper and Eucharist

Food, however, in terms of the Last Supper and the Eucharist, plays an important role in Christianity. As told by the early Christians, Jesus foresaw his death and performed a simple ceremony during a last meal to bring home the significance of his death to the Twelve: he broke a loaf into pieces and gave it to them saying, "Take this, it is my body." After they had eaten, he took the cup of wine and said, "This is my blood."

During the 1st century AD, Christian communities developed into self-contained units with an organized life of their own. When they were beginning to see themselves as a church, they held two separate kinds of services: (1) meetings on the model of the synagogue that were open to inquirers and believers and consisted of readings from the Jewish scriptures and (2) *agapē*, or "love feasts," for believers only. The latter was an evening meal in which the participants shared and during which a brief ceremony, recalling the Last Supper, commemorated the Crucifixion. This was also a thanksgiving ceremony; the Greek name for it was *eucharist*, meaning "the giving of thanks." This common meal gradually became impracticable as the Christian communities grew larger, and the Lord's Supper was thereafter observed at the conclusion of the public portion of the scripture service; the unbaptized withdrew so that the baptized could celebrate together.

Thus, from the very inception of Christianity, food and beverage has symbolized that religious experience is not purely personal but also communal. Moreover, differences in interpretation of the Lord's Supper have provided some of the contrasts among the major Christian churches. The opposing views of Roman Catholics and Protestants over whether the Eucharist bread is changed in substance or is a symbol of the flesh of Christ is an example of the role of food as a representation of religious differences within Christianity.

The rituals of the Eucharist provide the clearest examples in the Christian churches or confessions of the relationship between social stratification and food behaviour. Christianity, unlike Judaism or Hinduism and other Asian religions, was never tied to a caste system; correspondingly, it repudiated the entire body of purity–pollution laws of the Old Testament. Christianity was, however, part of the early European social system that was based on clear-cut separations of social classes. Religious food customs in Christianity, most notably in the Eucharist, reflect this.

The first Christian churches developed alongside the most rigid social stratification in European history, with elaborate notions of class authority and superiority and subordination. The separation of those in authority from the masses of ordinary people is mirrored in the Roman eucharistic ritual in which the sacrament's celebrant—the officiating priest—partook of the bread and wine first and then served only the bread to those of the faithful who wished it.

With the Reformation during the 16th century, which was (among other things) an overthrow of the traditional social order, a slight but important change in the eucharistic ritual was introduced, reflecting the weakening—but not the abandonment—of stratification and its attendant hierarchies of authority. In many Protestant confessions the officiating minister also partook of the bread and wine first, then served it to the congregation. In the Presbyterian ritual, the minister partook first and then served it to the elders who then served the people. Although this continued to reflect a system of stratification, it was a radical departure from the Roman rule that only the officiating priest could serve everyone. These rules for both Roman Catholics and Protestants are gradually changing in the 20th century.

Fast and abstinence

Until relatively recently, the most notable dietary law in Christianity was the Roman Catholic prescription to abstain from eating meat on Friday. This ban was lifted as part of the modernization of Roman Catholicism that was begun during the reign of Pope John XXIII. In Roman Catholic abstinence meat is forbidden, but there is no restriction on the amount of food eaten; fasting means that the quantity of food is also restricted. Historically, there have been several categories of fasts. The 40 days of Lent have traditionally been a period of mortification, including practices of fast and abstinence; the rules, however, have been greatly modified in recent years. Ember Days—a Wednesday, Friday, and Saturday at each of the four seasons—seem to be survivals of full weekly fasts formerly practiced four times a year. Vigils are single fast days that have been observed before certain feast days and other festivals. Rogation Days are the three days before Ascension Day and are marked by a fast preparatory to that festival; they seem to have been introduced after an earthquake about 470 as penitential rogations, or processions, for supplication.

Also important in the Christian complex of fasting is that associated with monastic life. Mortification is seen as essential to the practice of asceticism, and, in many rules of monastic life, fasting is regarded as one of the most efficient exercises of mortification.

Religions of India. It is in the religions of India that one can most clearly observe the principles outlined above concerning the relationship between dietary laws and customs and the existence of social stratification, traditional privilege, and social, familial, and moral lines that cannot be crossed. Hinduism provides the best example, although the same principles also obtain in the religions of Jainism and Sikhism.

Relationship between dietary customs and caste systems

Food observances help to define caste ranking: Brahmins are the highest caste because they eat only those foods prepared in the finest manner (*pakkā*); everyone else takes inferior (*kaccā*) food. *Pakkā* food is the only kind that can be offered in feasts to gods, to guests of high status, and to persons who provide honorific services. Food is regarded as *pakkā* if it contains ghee (clarified butter), which is a very costly fat and which is believed to promote health and virility. *Kaccā* is defined as inferior because it contains no ghee; it is used as ordinary family fare or as daily payment for servants and artisans. When food serves as payment for services (*e.g.*, barbering), the quality of the food depends on the relative ranks of the parties to the transaction; the person making the payment gives inferior food, such as coarser bread, to a lower ranking person performing the service. Performance of a service denotes that a person is ready to accept some kind of food, and giving food denotes an expectation that a service will be performed. Members of subordinate castes pick up the dirty plates of members of superior castes, as at village feasts. Food left on plates after eating is defined as garbage (*jūṭhā*); it is felt to have been polluted by the eater's saliva. This garbage may be handled in the family by a person whose status is lower than the eater's, such as a wife. Such food may be fed to domestic animals; among humans outside the family it can only be given to members of the

lowest castes, such as sweepers. The highest Brahmins do not accept any cooked food from members of any other caste, but uncooked food may be received from or handled by members of any caste. Nor will such Brahmins accept water across caste lines. Cow's milk is ritually pure and cannot be defiled, but a Brahmin will not accept milk from an untouchable—a member of the lowest caste groups—lest it has been diluted with water.

Water is easily defiled, but, if it is running in a stream or standing in a reservoir, it is not polluted even by an untouchable in it. Water in a well or container, however, is defiled by direct or indirect contact with a person of low caste. Thus, a ritually observant Brahmin will not allow a low-caste person to draw water from his well, although this rule is lapsing, possibly because of the introduction of plumbing and the removal of water from the list of scarce resources.

In the general Hindu system of purity–pollution, meats are graded as to their relative amount of pollution. Eggs are the least and beef the most defiling; but the highest caste Brahmins avoid all meat products absolutely. Also, certain strong foods (*e.g.,* onions and garlic) are thought to be inappropriate to Brahminical status. Alcohol too is prohibited; it is not considered polluting in itself, but the prohibition seems related to the Brahminical value of self-control. Alcohol's manufacture and trade is confined to members of lower castes.

People who eat at each other's feasts hold equal rank. People who eat at every house in a village occupy a very low status, and refusal to take food from another constitutes a claim to higher caste rank. More generally, givers of food outrank receivers. This, however, is a definition of collective, not of individual, rank. If a member of one caste gives food to a member of a second, all members of the first caste are regarded as higher than a third, even if there is no direct transaction between the first and third castes. Thus, the behaviour of every person in a village has consequences for the entire village.

Local variability of dietary restrictions

In actual practice, however, there is not an automatic enactment of these formal rules in village life; instead, they vary considerably according to local conditions. For instance, one of the formal rules of Hindu religious caste organization is that vegetarians outrank meat eaters, because contact with killed animals is regarded as polluting. Nevertheless, McKim Marriott, a U.S. anthropologist who has investigated village caste relationships, has found instances in which meat eaters outrank vegetarians. He concludes from his observations that it is caste rank—mostly in terms of the kinds of work that people in different castes do—that determines purity and pollution. In daily social relations this sometimes means that a caste of sufficiently high status may not be demeaned by receiving food from a lower caste if the latter is not too far below and if the proper food and vessels are used.

Status is rarely immutable over long stretches of time. In most societies, people who occupy low status try to exploit every opportunity to improve their position, and, Marriott found, Indian villagers are no exception. Because food in this culture is one of the principal indices of rank, it is used as a pawn in manoeuvres for social mobility. Specifically, members of a low caste will try to gain dominance over persons in another by feeding them, although the latter cannot be too far above the upwardly mobile group. There is no direct way of forcing a higher group to accept food; one of the techniques most often used is for the lower caste to threaten to withhold services unless a heretofore slightly higher caste receives food from the former. Such mobility, as noted earlier, affects not only the two castes concerned but also all other groups in the village, and the manoeuvring involves everyone in the community.

Marriott's emphasis on occupation (and, therefore, rank) as the determinant of food customs has not been accepted by all students of Indian society. He continues to leave some aspects of caste behaviour unexplained, such as the extreme statuses of Brahmins and untouchables, to say nothing of the existence of the total caste system itself and the mechanisms by which it is maintained. These problems have yet to be worked out. In any case, there can be no doubt that concepts of pollution and purity in regard to food in India, as everywhere else, are governed by a systematic set of rules analogous to a language's grammar and that applications of the rules are logical and consistent within the grammatical framework. Observations of daily village life do not contradict this concept of the codification of food rules; they only suggest that earlier "grammars" may have been too narrowly conceived.

Buddhism. Buddhism is, perhaps, the most difficult religion to discuss in terms of dietary laws and customs because it does not have any unity; its tradition has a complex history, and individual believers are characterized by varied faiths. Though Buddhism originated in India, it also diffused to—and had a great impact on—Ceylon, Tibet, China, and Japan. In each case, it was reshaped to conform with local conditions, especially those of social stratification. For example, most of the countries of Southeast Asia have caste systems in which there are outcastes or untouchables; Buddhism has been important in supporting such systems. Specifically, untouchability and the occupation of butchering animals tend to go together both in Buddhism and in many of the countries of Southeast Asia. But Burma, where Buddhism is the dominant religion, is an exception; having no caste system, Burmese society has not made butchering a basis of untouchability.

Lack of unity in Buddhism

Buddhism developed its own class distinctions, most notably between the monastic elite and the lay devotees. The social and political ethic of the laity was based on a merit-making ethic that was geared primarily to the urban mercantile and artisan classes. Thus, Buddhism claimed from its inception to be a Middle View (Mādhymika), opposed equally to the extremes of sensuousness and indulgence and of self-mortification. This Middle View was exemplified in the "five precepts": no killing, stealing, lying, adultery, or drinking of alcoholic beverages. These precepts were translated into an ethic of moderation in diet. A person must allay his hunger so that he may practice the religious life. Buddhism holds that man is weak and helpless by himself; thus it sees the purpose of religious action as bringing a return from the deities. Deriving from this is the practice of holding ritual vegetarian feasts for large numbers of monks, a noble patron, or for the benefit of a departed soul to promote health and longevity. Another Buddhist custom is the issuing of a prohibition against killing animals to end a drought or to speed the recovery of a sick emperor. According to the Vedic treatise the *Śatapatha Brāhmaṇa,* food, when enclosed in the body, is linked to the body by means of the vital airs. The essence of food is invisible. Food is the highest of all things that can be swallowed, and food and breath are both gods.

The prohibition of killing animals is more stringent in Buddhism than the injunction against eating them. Buddhism allows pure flesh to be eaten if it has not been procured for eating purposes or if the eater has not supposed it to be. The sin is upon the slayer, not the eater. This notion has been used in India and Japan to justify the outcasting or untouchability of butchers.

Religions of China. China is an example of the proposition offered above that religion alone does not give rise to eating rules; instead, religion serves to legitimate customary patterns of behaviour and social relations that emerge out of economic (especially occupational) and political relationships. Although China was under strong Buddhist influence, the Chinese never developed the institution of untouchability or outcaste. Indeed, Buddhism did not really penetrate China until after the beginning of the 2nd century AD; during the previous century, Buddhism was confined to foreigners in the northern commercial cities.

Chinese culture before the Han dynasty

Before 200 BC (the approximate beginning of the Han dynasty), Chinese culture was based on a rather elaborate system of social stratification in which mobility was rare and difficult. It was, in other words, a relatively closed social system, if not feudal. During this time, there were restrictions on the consumption of food: beef, mutton, and pork were to be eaten by an emperor; beef by feudal lords; mutton by high-ranking state ministers; pork by lower ministers; fish by generals; and vegetables by commoners (who probably could not afford meat or fish anyway). Officials, in fact, were known as "meat eaters,"

and it was generally only the aged commoners who were allowed to eat meat.

During this time, military affairs and sacrifices were considered the two most important things in the state. Sacrifice was inseparable from veneration of the ancestors, and almost no ceremonies were conducted without sacrifice and offerings. These ceremonies were integral features of daily life and, as a result, foodstuffs became associated with the moral code that was based on maintaining fixed social and political relationships. God and ancestors were often referred to as those "who are sacrificed to," and disobedience to them was believed to result automatically in catastrophe. Ceremonies marking important personal transitions (*e.g.*, initiation to adulthood and marriage) were held in the ancestral temple and were accompanied by feasts and sacrifices to the ancestors.

Ancestral veneration and the ethos of religiously validated legitimate authority remained as integral features of Chinese culture until the most recent years. Religious belief and observance notwithstanding, however, Chinese culture underwent a drastic change with the establishment of the Han dynasty. Most notably, the social class system was opened up—at least ideally—by the adoption of the principle of recruitment for public office; in later dynasties, this was expanded into the well-known system of written examinations, of grading officeholders by merit, and other features of the famous Chinese civil service. Correlated with the removal of the barriers to social mobility and establishment of the principle of ideally open recruitment to the civil service, the pre-Han food restrictions disappeared. This was also the time of Buddhism's greatest thrusts into Chinese thought and life.

Food continued to occupy an important social and religious place in villages, at least until the establishment of the People's Republic of China. For instance, marriage ceremonies traditionally last four days; the highlight of each day's celebration is a feast or sacrifice to the ancestors, sometimes both. Feasts and sacrifices are also important features of funerals, some of which are marked by two feasts in one day. These ceremonial occasions often work considerable economic hardship on families, forcing many of them into debt. (Y.A.C.)

Religions of Japan and Korea. Japan and Korea exhibit many of the same characteristics with respect to food customs as India, though with much less elaboration, and thereby the same relationships to Buddhism, though in an opposite direction. These relationships to Buddhism are also highlighted by contrasting Japan and Korea with China. Whereas post-Han China placed emphasis on achieved status and on personal superiority rather than on considerations of race or blood as a basis of social position, Japan and Korea (and also Tibet) established and continued a system of hereditary status and outcasting. As in India, therefore, the Japanese and Koreans considered pollution to be a hereditary taint; Buddhism played a major role in the legitimation of this ideology.

Outcaste groups

Outcastes in Japan traditionally were referred to pejoratively as *eta* (literally, "pollution abundant"). The accepted usage now is *burakumin* (meaning "hamlet people"), although this term has also taken on pejorative connotations. They are discriminated against in employment and intermarriage, live rurally or in slum conditions, have the lowest educational levels in the nation, and often suffer from malnutrition. In the past they were required to wear special clothing, slippers, and hairstyles; to stay away from other households; to remain in their own hovels at night; and to prostrate themselves before higher-caste people.

The history of the Japanese caste system in respect to food customs gives important clues to its origin. Among the ancient Japanese, meat was included in the diet, and the flesh of animals, fishes, and birds was offered to the gods as sacrifice. The flesh of ox, horse, dog, monkey, and fowl was prohibited, but that of deer, rabbit, and pig was not. During the 8th century AD the Japanese began to depend mostly upon plant rather than animal foods. In Japan's limited territory, it is understandable that cattle were raised for plowing and other agricultural work rather than for meat and milk. In 741 a law was passed forbidding the killing of cattle and horses, the latter being necessary for military as well as productive purposes. This provided a conducive atmosphere for Buddhist influences in the 6th and 7th centuries (primarily from China and Korea) that stressed the abhorrence and ritual impurity of blood and death.

Buddhism, however, was only one of several sources of outcasting slaughterers and butchers. During the 8th century, Shintō—the only indigenous religion of Japan—began to stress concepts of uncleanness as things that are displeasing to the gods: wounds, disease, death, menstruation, and childbirth; and this too contributed strongly to the development of *eta* status. It was apparently about this time that the belief developed in Japan that a person's association with blood and death changed his nature; this contamination not only carried over to a man's descendants but was thought to be communicable. It was apparently also at this time that Japanese cuisine began to favour fish (especially raw fish) as a staple source of protein.

Shintō

Important in this connection is that occupational specialization began to flourish in Japan during the 9th and 10th centuries; by this time, Buddhism was widespread in Japan. Traditional occupational roles became spheres of monopoly; in the face of competition from economically specialized groups who forced them out, people dealing with slaughtering, butchering, and tanning began to form guilds. This was rationalized by Buddhist and Shintō ideas that occupations associated with animal slaughter and processing (confined to *eta*) should be separated from the general body of commoner and slave occupations.

During the Heian period (794–1185), communities whose members were engaged in occupations related to death and animal products were forced outside the normal society, and they thus came to form the main body of outcastes in Japan. Increasingly, the latter were outcasted and considered untouchable, a pattern that reached its heights in the Tokugawa period (1603–1867). By the 17th century, the idea developed—supported by Shintō and Buddhism—that eating the flesh of all animals caused pollution for 100 days. After the mid-19th century, though, largely because of the emerging influence of Western cultural habits, meat consumption began to be more widespread in Japan, and among some Japanese the consumption of beef became associated with progress and enlightenment.

Soon after the Meiji Restoration (1868) the caste system and the legal discrimination against the *eta* were abolished. Outcasting, however, dies slowly. Though the egalitarian ideologies of modern industrialization are incompatible with caste, outcasting tends to remain in Japan and, alongside it, some of the food customs associated with the caste system. As in India, eating together (along with marriage and social visiting) between untouchables and members of normal society is disdained. In many parts of Japan, especially in traditional villages, the diet remained largely vegetarian until after World War II, when the consumption of meat and other Western dietary practices rapidly increased. Even the consumption of milk, which had been considered unclean, became common. (Y.A.C./Ed.)

Ceremonial and ritualistic objects

Throughout the history of religions and cultures, objects used in cults, rituals, and sacred ceremonies have almost always been of both utilitarian and symbolic natures. Ceremonial and ritualistic objects have been utilized as a means for establishing or maintaining communication between the sacred (the transcendent, or supernatural, realm) and the profane (the realm of time, space, and cause and effect). On occasion, such objects have been used to compel the sacred (or divine) realm to act or react in a way that is favourable to the participants of the ceremonies or to the persons or activities with which such rituals are concerned, or to prevent the transcendent realm from harming or endangering them. These objects thus can be mediatory devices to contact the divine world, as, for example, the drums of shamans (religious personages with healing and psychic-transformation powers). Conversely, they can be mediatory devices used by a god or other supernatural being to relate to man in the profane realm.

They may also be used to ensure that a chief or sovereign of a tribe or nation achieves, and is recognized to have, the status of divinity in cultic or community ceremonies. Of such a nature may be phallic cult statues bearing the name of a king associated with that of the Hindu god Śiva, in areas under Indian influence (such as in ancient Vietnam, Cambodia, and Indonesia, where the lingam was worshipped under a double name: Indreśvara [Indra, king of the gods, plus Īśvara, Lord, a name of Śiva]), or the Buddhist "body of glory" statues in Cambodia dating from the end of the 7th century. The religious dance masks of many societies, including those used in ancient Tibet and in Buddhist sects of Japan, may, to some extent, also belong to this class.

VARIETIES

Because such objects vary as much in nature as they do in form and material, they are difficult to evaluate. If limited strictly to religious practices, an inventory of ceremonial and ritualistic objects remains incomplete, because these objects have played significant roles on solemn secular occasions, such as consecrations, enthronements, and coronations, which may be closely linked to the divine order, as in Hindu-, Buddhist-, and Christian-influenced cultures.

Icons and symbols. Constituting a most significant category of cult objects are representations of a deity. Though such representations are often depicted in the form of statues and images (icons) of divine or sacred beings, they may also be either figurative or symbolic, the meanings often being equivalent. In Tantrism (a Hindu and Buddhist esoteric, magical, and philosophical belief system centred on devotion to natural energy), for example, the sacred Sanskrit syllable Om—which is a transcendent word charged with cosmological (order-of-the-universe) symbolism—is identified with the feminine counterpart of the god. In its written form, particularly on Tibetan banners (tankas), the word Om (often corresponding with the feminine counterpart—Tārā—of the patron of Tibet) is considered to be eminently sacred, even more so, in some instances, than an anthropomorphic (human-form) divine effigy.

Religious statues and images

Statues and painted images occur most frequently in religious iconography, as noted above. These are often viewed as the permanent embodiments of the deities they represent, whether they are located in sacred places of religious communities, such as temples, shrines, or chapels, or on domestic altars, which contain statues or icons of the divinities of prosperity and fertility, mother goddesses, household gods, saints, relics, the tablet of the ancestors in ancient China, and other similar domestic cult objects. Many household cult objects are made from clay or terracotta and are sometimes multicoloured. The material of which major cult objects are composed is often explicitly defined and assumes a certain importance. If the statue is fashioned in wood, the choice of the wood (acacia, sandal, or any other) is symbolically important because it is considered auspicious. By the same token, the choice of stone is likewise important, depending on the region. If metal is chosen, it is one that is deemed precious (*e.g.*, golden statues bring prosperity). In the case of bronze statues and other cult objects, the composition is carefully defined and often corresponds to alloys to which symbolic values are attached. In addition to a proper and distinct form and material, the technique of fabricating and the procedural patterns of composing such objects are controlled by traditional rules that have become established rituals in many religions—sophisticated, folk, and primitive. In the production of statues in human or animal form, the last procedure is often the "opening the eyes" (*i.e.*, the painting of the eyes of a statue of a deity or inserting gold in them by an officiating priest during the installation of the statue [*pratiṣṭhā*] in the sanctuary, along with the reciting of appropriate prayers that make the statue "living" and "real"), particularly in Brahmanic India and Chinese-influenced areas (see also RELIGIOUS SYMBOLISM AND ICONOGRAPHY).

Religious dress and vestments. The practice of wearing special garments for conducting rites, participating in worship, or even witnessing such ceremonies is very unevenly distributed, and the conceptions associated with this practice are highly varied and complex (see below *Religious dress and vestments*).

Instruments for worship and religious ceremonies. The types and varieties of instruments used in worship and religious ceremonies are almost innumerable. The role they play in ritual occasions may be as containers and sacred furniture, as objects with properties necessary for worship, and as "mediatory" objects through which a magical or mystical connection is believed to be made between the human and divine worlds. There are also the materials used in bloody or nonbloody sacrifices.

Amulets and talismans. Amulets (charms) have been used for protection in all ages and in all types of human societies; they persist even today in industrial societies, in which they are mass-produced by the most modern methods (*e.g.*, mustard seeds encased in plastic to be worn as necklaces reminding the wearer of Jesus' words about the growth of the Kingdom of Heaven). The purpose of most amulets is not so much religious as it is for protection against danger, sickness, and bad luck (*e.g.*, the "mystical eye" of the ancient Egyptians or the "Hand of Fatima" of Muslims). The same is true of talismans, which offer the additional advantage of conferring supernatural power on other people, even on the deity, from a distance. Dancers' masks and jewels, such as earrings, bracelets, necklaces, and belts, may be classified with amulets. Such objects are individually worshipped in order to gain their goodwill among some Hindus in India and among the Pueblo and Navajo Indians of North America (see also OCCULTISM).

TYPES OF SACRED SETTINGS FOR CEREMONIAL AND RITUALISTIC OBJECTS

Places of worship and sacrifice. Throughout history there is evidence of worship at natural sites as well as at sites constructed for ritualistic purposes. In the protohistory and perhaps the prehistory of most ancient civilizations, people venerated trees, stones, bodies of water, and other natural objects, which gradually became the objects of established cults and which often were included, in some form, as aspects of later official ritual. Initially, the objects of this frequently occurring process were sacred trees considered to be the habitats of spirits or gods, such as in Vedic, Brahmanic, and Buddhist India or pre-Islāmic Arabia; sacred stones, such as fragments of meteorites, menhirs (upright stones), and rocks—for example, the Black Stone of Mecca in the Kaʿbah; flowing waters, natural lakes, and sacred and purifying rivers, such as the Ganges; crossroads and junctions, such as the *tīrtha* (river fords and, by extension, sacred spots) in India; and other such objects or places of nature. According to Hesiod, an 8th-century-BC Greek writer, such objects of nature were venerated in the popular piety of the rustic people of Greece in his times.

Importance of the natural elements

The association on the same site of four natural elements (mountain, tree, stone, and water) is supposed to constitute a sacred whole (a quarternity of perfection), a sacred landscape or "geography" similar to the world of the gods. Such sites, in many civilizations, were the initial points of departure for pilgrimages or for the establishment of places of worship. In some instances the natural sacred places were gradually adapted for religious use (*e.g.*, the oracle at Delphi, in Greece), but in others the earlier natural sites were artificially recreated by using man-made symbolic equivalents. An artificial or natural hill, such as a barrow, mound, or acropolis (elevated citadel), often served as a base for the temple, but in many instances the temple itself has been an architectural representation of the mountain, as were the *bamot* ("high places," usually constructed with stones) of the ancient Hebrews, the ziggurats (tower temples) of the ancient Babylonians, and the pyramidal temples of Cambodia, Java, and pre-Columbian Mexico. A branchless tree has often been transformed into a cultic object: a sacrificial post, such as the Vedic *yūpa;* the central pole of a nomadic tent in Siberia and Central Asia, the *yurt,* or initiation hut; or a parasol shaft (*chattrāvali*) in the Buddhist *stūpas* (reliquary buildings) and the Japanese and Chinese pagodas. If represented in stone, the tree evolved into a column gnomon (a perpendicular shaft), such as the Buddhist *lāṭ,* the sacred pillar (*matzeva*)

of the ancient Hebrews, or the obelisk of pre-Hellenistic Egypt (before the 4th century BC, especially from the 3rd millennium to the early 1st millennium BC). Stone, transformed into an altar, has been used either to support or seat the image or symbol of the deity, or to receive sacrifices, burnt offerings, plant offerings, or aromatic perfumes. Water, because to it is generally ascribed a power that is purifying or even curative or miraculous, almost always plays an important role in or near sacred places. The whole assemblage of actual or symbolic mountains, trees, stone, and water is usually arranged architecturally within an enclosed space. An example of this arrangement is the typical Christian church, with its raised chancel (the mountain), the cross or crucifix (the tree), the altar (usually stone, but sometimes wood), and the baptismal font or tank (water).

This widespread scheme is almost everywhere bound up with a cosmology (theory of the universe) that establishes a symbolic identity between the divine world and the temple. This identity holds true in all stages of culture; *e.g.,* the sacred sites of the Algonquin, Sioux, and Blackfoot North American Indian tribes; the *templum* (temple) of the Etruscans in ancient Italy; the temple of Bel at Palmyra (in Syria); the Mithraic crypts centring on devotion to the Iranian god Mithra throughout the Roman Empire; the *kiva* (a circular, partly underground ceremonial room) of the Pueblo villages; the Temple of Heaven at Peking and that at Hue (Vietnam); the Buddhist *stūpa;* and Brahmanic, Buddhist, and Mexican mountain temples. The cosmic character of the Israelite king Solomon's Temple, of the 10th century BC, constructed on Mt. Moriah in Jerusalem, was not given such an interpretation, however, until hellenistic times (3rd century BC–3rd century AD), as in the writings of Philo of Alexandria and Josephus. That of the Muslim mosques is very subdued, although the Ka'bah of Mecca, which contains the black stone, is believed by Muslims to be the centre of the cosmos. The cosmological scheme has been applied to Christian basilicas and churches—with square floor plans, an overarching dome, and symbolic ornamentation—from as far back as the 6th and 7th centuries.

Sacred furniture and related objects. Whatever its size and form, a sacred area is usually delimited by an enclosure, such as a simple fence around sacred trees or Buddhist *stūpa*s or high walls with immense gates around temples. The sacred space may comprise multiple enclosures, such as that of huge sacred structures—such as the temple of Śrīraṅgam in southeastern India, which has seven concentric enclosures. The dominant idea in delimiting the holy place is to protect the sacred element and its mystery. Access to the sanctuary is often hidden by grills or screens: the veil of the Jewish Temple in ancient Jerusalem, which separated the holy area (or *hekhal*) from the Holy of Holies (or *devir*); or the Eastern Orthodox *ikonostasis* (image screen), which hides the chancel from the view of the faithful except on certain ritual occasions when it is opened to them. Hindu sanctuaries also are concealed by hangings. In Roman Catholic, Lutheran, and Anglican churches, the chancel has usually been separated from the nave by a railing, before which the faithful kneel to receive the eucharistic (communion) meal.

Objects hiding access to sanctuaries

In Indo-European civilizations the essential element of the sacred furniture is the altar, the site of which varies according to the cult and period under consideration. Tables for sacrifice, burnt offerings, and offerings of plants or perfume have sometimes been placed outside the temple, as at the Temple of Solomon in Jerusalem and in temples of ancient Egypt. In early Christian cults, a single altar was placed in the chancel. Later, about the 6th century, the number of altars was increased, with one in each chapel of the larger church building.

The most sacred furnishings of temples are those most closely related to altars, such as the Jewish ark of the Law, or *aron ha-qodesh,* in the synagogues, which is made in the image of Moses' ark of the Covenant, and the tabernacle (the receptacle containing the consecrated bread and wine) of Roman Catholic and Eastern Orthodox churches. The ark, which is portable, is a kind of chest (*aron*) with a cover (*kapporet*), and the tabernacle, made of wood, metal, or stone, is a locked chest. On the fire altars of Zoroastrianism (a religion founded by the Iranian prophet Zoroaster in the 7th century BC) is a sacred metal urn (*ātash-dān*), containing the eternal fire, ashes, and aromatic substances.

When temples or other major sanctuaries are also places for assembly and common prayer, as, for example, Muslim mosques and Catholic and Protestant churches, pulpits are provided. They may be integral parts of the masonry, of the anterior screen of the chancel—as are ambos (raised platforms), or wooden furnishings fixed to the walls, like the formerly mobile *minbar* (domed boxes in mosques). In Manichaeism (a dualistic religion founded by the Persian prophet Mani in the 3rd century AD), the Bēma Feast was centred on the exaltation of a reconstructed pulpit (Bēma), which symbolically represented the rostrum from which Mani spread his teachings. Another important element of sacred furniture is the lectern, on which is placed one or more sacred books (from which one of the officiants reads aloud) or a collection of hymns and religious chants intoned by a cantor in monasteries or other religious structures.

Use of pulpits and lecterns

Permanent lighting is also required in certain cults. This has encouraged the creation of supports or vessels for inflammable materials, the most characteristic of which are the seven-branched candelabrum of the Jerusalem Temple, the Easter candle holder of Roman Catholicism, the sanctuary lights of Roman Catholicism that signal the presence of the Eucharist in the tabernacle, lights suspended before icons in Orthodox rituals, glass or perforated-metal lamps in mosques, and spherical lanterns adorned with an eye, which represents the universal monad (one), of Vietnamese Cao Daism (a syncretistic religion combining Confucionism, Taoism, Roman Catholicism, and Buddhism).

Protective devices and markers of sacredness. Other objects, such as fans, flyswatters, parasols, and standards—analogous to the symbols of royalty—often complete the permanent furnishings of sacred places. In addition to their utilitarian role, they are endowed with a sacred character; fans used in Brahmanic and Buddhist cults may be compared to the *flabella* ("fans") in the Roman Catholic and Orthodox churches. They are waved before the iconostasis during the Eucharist in the divine liturgy of the Eastern Orthodox Church, and they also are placed on either side of the papal chair in solemn processions. The parasol, or umbrella, is generally a symbol of the vault of heaven, as in India and China; the domes of *stūpa*s are often surmounted by parasols (*chattras*). In its symbolic and protective role the umbrella can be compared to the baldachin (canopy) in many of its forms. Whether it covers the altar, the statue or symbol of a deity, or even the imperial throne—as in Zoroastrian Iran during the Sāsānian period (3rd–7th centuries) and Orthodox Byzantium (during the 4th–15th centuries)—the baldachin's celestial symbolic ornamentation is generally explicit, and its cosmic character is apparent. The standard (*dhvaja*), in the Brahmanic cults, takes on the appearance of a high column (*dhvaja-stambha*) erected in front of temples; it is surmounted by a divine effigy, most often that of the sacred steed, or *vāhana,* of the god. Simultaneously a signal (because of its height) and a protective device, it first receives the homage of pilgrims. The poles adorned with flags erected before the pylons of the temples of ancient Egypt may also have had such a double character.

The symbolic character of parasols, canopies, and standards

FORMS OF CEREMONIAL AND RITUALISTIC OBJECTS ACCORDING TO THEIR FUNCTIONS

Summoning, mediating, and expelling devices. In the form of magic or sacred words, singing, and music, sound plays or has played an important role in the worship of most religions. The same is true of light and of aromatic substances, such as oils, perfumes, and incense. The importance of these elements has brought about the creation or adoption of specific objects with functions that often serve converging purposes in worship. In most cases they are used to draw the attention of the deity, to establish a connection with it, and to exorcise forces that are evil or harmful to the god and to men. Because of the need to attract the deity's attention, the sound-producing

instruments are usually percussive or shrill, rather than melodic, and drums, gongs, cymbals, bells, conchs, and sistrums (timbrels, or rattles) are the most common forms.

Sound devices. Summoning devices are played either alone, as objects to accompany prayers or litanies, as in Tantric Buddhism, or as instruments in a temple orchestra. Their size and form and the materials used to make them vary according to locale. Generally viewed as sacred, they are often worshipped, as in West Africa, Malaysia, and Burma, and partake of divine attributes, as in Brahmanism, Mahāyāna (Greater Vehicle, or northern) Buddhism, and Tantrism. Drums vary greatly in both size and form. The two-skinned *ḍamaru* (drum) of Śaivism (devotion to the Hindu god Śiva) and Tantrism, believed to be effective in communicating with the divine world, is shaped like an hourglass and fitted with two pellets that hang from cords and that strike the skins when the drum is twirled. Gongs usually are suspended metallic disks, with or without a central protuberance. The gongs of ancient and contemporary China, however, are of varied form, with cutout designs, and may be made of resonant stone or of jade. Cymbals are very widespread and were used in the Hellenistic mystery (salvatory) religions, such as those of Dionysus (a god of wine) and the Eleusinian mysteries (centred on devotion to Demeter, a seasonal-renewal goddess). They were the only instruments played in the Temple of Jerusalem, where they were known as *metziltayim* or *tzeltzelim.* The sistrum, used in pre-Hellenistic Egypt in the worship of the goddesses Isis and Hathor and in Rome and Phoenicia, as well as among the Hebrews, is composed of a handle and frame with transverse metal rods and mobile disks. Producing a sharp ringing sound, it was regarded as particularly sacred and was carried to the temple by women of high rank. There are countless types of bells; the Indian *ghaṇṭā,* or Tibetan *dril-bu,* a metal handbell with a handle shaken during prayers in order to attract beneficent spirits and to frighten away evil ones, is used particularly during Brahmanic and Mahāyāna Buddhist ceremonies.

In this category of objects, the shaman's drum of the Buryats, Yakuts, Altaic Turks, and Eskimos is composed of a skin stretched over a circular or oval frame provided with a handle; it is struck with a curved beater. It plays the same magical role as the *ghaṇṭā,* but it also serves as a mode of ascending to the realm of the sacred for the shaman. The bull-roarer—a flat, elongated piece of wood, ivory, reindeer antler, or other material—used in primitive religions of Australia, equatorial Africa, western North America, Colombia, Brazil, and Sumatra, and the similar *rhombos* of the Hellenistic mystery religions, was propelled and whirled by a thin strap. Its humming sound and trajectory gave it the dual character of a summons to the divine world and a link with the celestial regions.

Lighting devices. In comparison with sound, which in worship usually presents a coercive character, lighting and fire, whether permanent or occasional, generally signify a sacred or spiritual presence, an offering, prayer, intercession, or purification. They are often viewed as sacred or even of divine origin, if not directly identified with the deity, as in the Zoroastrian fire altars. Their supports and containers can be made of either durable or perishable materials, depending on the ritual or ceremonial requirements. Torches have been used throughout history: in ancient Assyria and Babylonia they were used to carry a newly consecrated fire from torch to torch throughout the city three times a month; in ancient Rome they were sometimes placed in a hollow clay or metal shaft; and in the ancient Hebraic religion a lamp (*ner*) filled with sacred oil was used in the worship of the god Yahweh. In the Roman Catholic Church, from about the 10th century on, wax candles have been used, with bronze or copper candle holders—the forms of which changed according to style. Two of them were placed on the altar for the mass, and two others were carried by acolytes (light bearers). The Easter (Paschal) candle, made of beeswax around a wood core, had a candle holder appropriate to its size. At Westminster, in England, during the 14th century, a *candela rotunda* ("round candle") was the centre of a "festival of lights" during the feast of the purification of the Virgin Mary (February 2), also called Candlemas Day.

The use of torches

Festivals of lights have been and still are common throughout the world, especially among the Jews, who celebrate Ḥanukka, the Feast of Dedication of the Temple. In India and in Indian-influenced countries (particularly Thailand), the festival of lights (Dīpāvālī or Divālī) is celebrated by the Vaiṣṇava Hindus (devotees of the god Vishnu [Viṣṇu]) in October–November, at the end of the monsoon season. It is practiced on other religious occasions by the Jaina (followers of the Indian reformer Mahāvīra, of the 6th century BC), Thais, and Tibetans, who celebrate it in December. The lamps, which are lit everywhere (*e.g.,* in temples, in houses, and at crossroads), are also set afloat on streams, rivers, and lakes. Some lamps are made of glass—like the votive lights of Roman Catholicism—with a wick dipped in a vegetable oil, usually coconut; some are made of clay; and others are made of rice paste with a central hollow filled with ritual clarified butter, or ghee (*ghī*), or are cut out of a plant stalk in the shape of a bark or raft. The Jaina use earthen saucers containing either wicks immersed in coconut oil or pieces of lighted camphor. Another form of this festival was known in Thailand, where three earthen pots, containing rice, seeds, beans, and an oil-soaked wick, were placed at the top of three poles opposite the temple entrance, and the fire was kept burning for three days.

The "cordons of light" placed around the sacred places of Buddhism during great festivals, such as at Bodh Gayā, in India, for the Buddhajayantī (the commemoration of the Buddha's 2,500th birthday) in 1958, are composed of thousands of small brass lamps in the form of footed cups filled with ghee, in which a cotton wick is soaked.

Incense and other smoke devices. The use of incense or the fumes of aromatic substances is especially widespread in the great religions of the world and has many symbolic meanings. It may signify purification, symbolize prayer (as among the Hebrews), or be an offering that rises to the celestial or sacred realm. Bronze incense burners were cast very early, as exemplified by those from the Chou period (*c.* 1111–255 BC). Their forms were often inspired by cosmological themes. In early Taoist ritual the fumes and odours of incense burners produced a mystic exaltation and contributed to well-being. Under the T'ang dynasty (AD 618–907), perforated golden vessels with handles were carried in the hand to accompany a votive offering. In Japan the censer (*kōdan*)—a vessel with a perforated cover and carried by chains—was used in Buddhist and Shintō rituals. In pre-Hellenistic Egypt and among ancient Jews, incense was burned in golden bowls, which sometimes had handles, and in cauldrons placed on or beside the altar or outside the temple. In pre-Columbian Mexico and Peru, incense burners were made of terra-cotta and sometimes of gold. Censers of precious metal provided with chains for hanging have been used since the 4th century in Christian churches, and the rite of swinging the censer is practiced in many rituals, both Christian and others.

Symbolism of incense and other aromatic substances

Expelling and other protective devices. Several of the objects already described serve as protection against evil or demonic spirits. Of such a nature are the *ghaṇṭā* and *dril-bu,* the shaman's drum, the lamps of the Indian Dīpāvālī, and the burning of incense, which has also been practiced in ancient Greece, pre-Columbian America, Morocco, and many other regions. The possession of a large number of the same form of a protective object often is believed to be effective; this is the reason for the large number of bells (*ghaṇṭāmālā*) suspended on lattices on the handrail of the balustrade (*vedikā*) around the *stūpas* of ancient India; even today, small bells are hung from the roofs of Buddhist pagodas in Sino-Japanese regions. Like the small bells seen on the roofs of Romanian country dwellings until the beginning of the 20th century, these bells have a clapper provided with a feather or plaquette that enables the wind to ring them continually. Perhaps the most effective protective object, however, is the "diamond thunderbolt" (Sanskrit *vajra;* Tibetan *rdo-rje*) of Mahāyāna Buddhism, Tantrism, and Lamaism (a Tibetan form of Buddhism and folk religion). Well-known in early Buddhism as an instrument held in the hand, the *vajra* is handled in the middle and has, at one or both ends, four curved points

that meet at the tips. Of varying size, they are usually made of gilded or ungilded bronze. The Tantric *vajra* is also associated with the *ghaṇṭā* (*vajra-ghaṇṭā*), for which it forms a handle. A symbol of the indestructible force of religion, it is believed to be able to drive away all manifestations of evil. Although they are perishable, gunshots and firecrackers are viewed as protective and expelling devices, as in China and Cambodia (where soldiers, in the early 1970s, fired ammunition at a lunar eclipse to drive away the dragon they believed was devouring the moon).

Representational objects. In many religions, the god or divine order is represented among men by objects, which may be regarded simply as the god's material form on earth or may be totally identified with the god and endowed with his powers. In pre-Hellenistic Egypt the god was believed to be present in his statue, and elsewhere the statue frequently was believed to contain the god.

Statues of divine beings and sacred masks

Figures. Statues of human or animal figures are the most explicit of the objects representing the divine order. In most iconic (image-using) religions the gods are generally anthropomorphic, half man, half animal (as in Egypt and India) or often entirely animal. In most cases the statues conform to an ideal physical type that is symbolic and conventional. The formulation of the ideal is governed by precise aesthetic and iconometric (ritual image proportion) rules, as well as by iconographic (image-representation) requirements, as in Egypt, Greece, and India. All such standards and requirements guarantee conformity to the divine model and, therefore, the effective presence of the god in his statue. Typical in this regard are the sculptured animals of the Hindu pantheon, such as elephants, lions, horses, bulls, and birds, which—erected at sacred places in India and other Hindu-influenced countries—serve as ever-ready sacred mounts (*vāhana*) for the journeys of the corresponding gods.

The masks representing beneficent and maleficent sacred or holy forces in religious dances—particularly in Buddhist monasteries of Nepal, Tibet, and Japan and in the majority of primitive societies—constitute another category of sacred representational objects. They are usually worshipped just as statues are worshipped.

Certain customs incorporating representational figures have been widespread since prehistoric times and appear to be more related to magic than to religion. One example of this type of practice is a custom observed in primitive or prehistoric societies—the incorporation of a skull in an anthropomorphic statue in order to emphasize its divine, sacred, or magical character. To some extent, a similar use of a skull, human bones, a mummified corpse, or a skeleton appears in Christian churches in the veneration of relics.

Plants and plant representations. In all civilizations, plants and trees have been viewed as sacred. Generally, the tree is either a god's habitat or the god himself and is worshipped. Such was the case, for example, in early Indian Buddhism. Trees may also be associated with the divine order because of some incident and subsequently venerated, as was the bodhi tree, under which the Buddha received his Enlightenment. Fences or even open-air temples, the form adopted for the early Bodh Gayā Buddhist temples, are built around such trees. Innumerable cases of sacred or divine trees and their painted or sculptured representations are found throughout written religious tradition and in the ethnological data. The branches of trees such as the palm, olive, and laurel are often associated with the gods; such branches may crown the god or be included among his attributes. Many are used in worship, as are the branches of the *bilva* (wood-apple tree) among the adepts of Śiva, and the *tulasī* (basil), symbol of Lakṣmī (Hindu goddess of prosperity and Vishnu's wife) and sacred plant of the Vaiṣṇavites.

Plants as symbols of life and immortality

As symbols of life and immortality, plants such as the vine of the Greco-Roman and the Christian world and the *haoma* (a trance-inducing or intoxicating plant) of pre-Islāmic Iran are planted near tombs or represented on funerary steles, tombstones, and sarcophagi. Two similar and related rites involving plants, the *haoma*, noted in the Avesta (ancient Zoroastrian scriptures), and the *soma*, noted in the Vedas (ancient Hindu scriptures), pertain to the ritual production of exalted beverages presumed to confer immortality. The ritualistic objects for this ceremony included a stone-slab altar, a basin for water, a small pot and a larger one for pouring the water, a mortar and pestle for grinding the plants, a cup into which the juice drips and a filter or strainer for decanting it, and cups for consuming the beverage obtained. In many sacrifices, branches or leaves of sacred plants, such as the *kuśa* plant (a sacred grass used as fodder) of the Vedic sacrifice and the Brahmanic *pūjā* (ritual), are used in rituals such as the Zoroastrian sprinkling (*bareshnum*), or Great Purification, rite, in which the notion of fertility and prosperity is combined with their sacred characters (see above under *Purification rites and customs*).

Use of banners and shields

Other representational objects. The staves of martial banners or standards are often surmounted by the figure of a god, which is frequently in its animal form. Such effigies, used by the Indo-Iranians, the Romans, the Germanic tribes, the Celts, and other ancient peoples, were probably meant to ensure the presence of the god among the armies. From the 4th century on, Byzantine armies placed the *labarum* (a cross bearing the Greek letters XP, signifying Christ) on their standards. Shields, such as the Greek *gorgonōtos* ("gorgon-headed"), were also often decorated with sacred figures, emblems, and symbolic themes, particularly in post-Gupta (4th-century) India, as seen in the 6th-century findings from the frescoes of Ajantā. In the Mycenaean civilization (15th–12th centuries BC) of ancient Greece, shields were worshipped in front of the temple, and at Knossos (in Crete) votive offerings were made of clay and ivory in the form of shields. The famous *ancilia* ("figure of eight" shields) of Rome were kept by the Fratres Arvales (a college of priests) and used by the Salii (Leapers), or warrior-priests, for their semiannual dances (in March and October) honouring the god Mars.

Relics. Relics of saints, founders of religions, and other religious personages, which are often objects of worship or veneration, generally consist of all or part of the skeleton (such as the skull, hand, finger, foot, or tooth), a piece or lock of hair, a fingernail, or garments or fragments of clothing. Such veneration is nearly universal, as is the production of reliquaries, or shrines that contain relics. The size, form, and materials of reliquaries vary greatly and often depend on the nature of the relic being exhibited. They may be fixed but are generally portable so that they can be carried in processions or on pilgrimages. Wood, bone, ivory, quartz, glass, semiprecious stones, and metals such as gold, silver, bronze, and copper are frequently used materials, and chasing (embossing), enamelwork, and precious stones often ornament reliquaries. They vary considerably in form; like the Tibetan reliquaries, or *ga'u*, they may be constructed on a small scale to look like churches, chapels, towers, *stūpas*, or sarcophagi, but sometimes they assume the form of the relic, such as in the form of anthropomorphic statues, busts, hands, feet, and other forms. Occasionally, as in Tantrism and Tibetan Lamaism, the bones of holy persons are used to make ritual musical instruments—flutes, horns (*rkang-gling*), and drums (*damaru*)—or objects such as the ritual scoop made of a skull cup (*thodkhrag*) and a long iron handle encrusted with silver.

In many Asiatic regions, however, human relics are replaced by copies of sacred texts introduced into statues of bronze, as in Tibet and Yunnan (China), or of stucco, as in Afghanistan (Hadda, an archeological site near Jālālābād excavated since 1928) in about the 4th–6th centuries.

Other ritual objects. *Objects used in prayer and meditation.* In many religions the practice of prayer requires the use of certain objects, among which rosaries (strings of beads) and chaplets (circular strings of beads) occupy an important place in the popular piety of various religions. They are widespread in Hinduism, Buddhism, Islām, Roman Catholicism, Eastern Orthodoxy, and Judaism, although they are not found in Shintō. Brahmanic and Buddhist rosaries have 108 beads, made of *tulasī*, or basil (in Vaiṣṇavism), of lotus seeds or small bones (in Śaivism), or of small disks of human bone (in Lamaism). In China, rosaries are composed of coloured beads. Elsewhere, their number varies; the rosary of Japanese Buddhism has 112

Use of rosaries and chaplets

wooden beads, that of Islām has 99 amber beads, and that of the Christian world—and of the well-to-do Jaina—has 150 beads made of various materials, such as wood, pearl, mother-of-pearl, precious or semiprecious stones, gold, and silver. The beads of Brahmanic and Buddhist rosaries are usually strung continuously, except in Japan, where cords—which may or may not have beads on them—are tied to the principal cord in several combinations. The Christian rosary is divided into "decades" (tens) with intercalations, and in many cases the rosary has a "head" composed of a larger bead, several other beads, and a Christian cross.

There are several other objects pertaining to prayer—in addition to the rosary, which is principally a mnemotechnic (memory-technique) device. One example is the Lamaist prayer wheel (*'khor-lo*), which varies widely in size. It is a cylinder, generally of chased metal, rotating on an axis and containing prayers inscribed on strips of paper, fabric, or parchment. Weighted by two balls suspended externally on small cords, the prayer wheels are set in motion when a hand rotates a handle extending from the axis or when the prayer wheels are aligned along a common axis. Some are driven by hydraulic power and others even by electrical power. There is some evidence of the use of prayer wheels among other peoples, such as the Japanese, the ancient Celts and Bretons, the ancient Greeks, and the ancient Egyptians. The idea of permanent prayer through the agency of objects is found in the candles lit in churches, in the perpetually burning lamps (*chōmyōtō*) of Buddhist Japan, and in Tibetan prayer flags, with sacred formulas painted on them, which wave in the wind around temples, houses, and villages. The phylacteries (*tefillin*) worn by traditional Jews during weekday morning prayers consist of two leather cases bound by leather straps to the forehead and left forearm; they contain parchment citations from the Pentateuch enjoining this as a reminder of God's commandments. An amuletic function has been attributed to them, but this is disputed. Among protective objects associated with prayer are Muslim prayer rugs, the rectangular shape of which symbolizes the sacred area of the mosque, and the fringe-trimmed prayer shawl (*tallit*) worn by devout Jews during synagogue services.

Permanent prayer by means of ritualistic objects

Related to prayer and meditation are sacred and magical diagrams. Typical examples are the *yantras* (two- or three-dimensional meditation apparatus, often geometrical or anthropomorphic in form) and *maṇḍalas* (symbols of the cosmos in the form of circles, squares, or rectangles) of Brahmanism, Tantric Buddhism, and Lamaism and found in India, Nepal, Tibet, China, Korea, and Japan. Derived from sacred syllables (*mantras*) or from geometric designs endowed with mystical and cosmological symbolism, they are executed on sand, on the ground with coloured powders, and on durable materials. They may be made on stones, engraved on plates of copper, silver, or some other metal, or drawn and painted on skins, linen, silk, or hempen cloth. Like statues, they are consecrated by the rite of "initiation of breath," *prāṇapratiṣṭhā* (see also *Prayer*, above).

Objects used in purification rites. Large numbers of purification rites are performed universally on widely varying occasions, both in private life, from conception to death, and in religious ceremonies. Such rites employ materials that include water, dust, or dry sand (in Islām); water and henna, a reddish-brown dye (in Islām); oil, incense, balm, and natron, a salt (in ancient Egyptian religion); ale (*ōl*) or wine (in post-15th-century Germanic religion); salt (in Shintō); bread, sugar, spices, and animal blood (in ancient Greek and Scandinavian religions); paper, used in the Shintō *gohei*, a white paper "whip" that is shaken; ashes, among the Brahmans; and other materials. Water, fire, and light play especially important roles in purification rites. Objects used in such rites include water vessels of various shapes and sizes used for ablutions; jugs and vats containing ale or wine; terra-cotta or glass containers used for balms and perfumes; incense burners, cauldrons, and censers for fumigation; containers used in Confucian rituals, which include a basin (*chin-lei*) for pure water, another small basin (*huan-po*), and seven goblets (*chio*) for the sacrificial wine; and ewers and basins of gold, silver, or copper used in purifying the hands and feet, as in pre-Hellenistic Egypt, or for ritual sprinklings.

Forms of objects in purification rites

The wearing of new clothes that have not yet been washed is also a purification rite, practiced, for example, in the spring of the year (October–November) in Brahmanic India, where it is associated with the festival of lights, the Dīpāvāli.

Purification may also be attained through mortification and penance, practices that were especially common in medieval Christianity and in Judaism. Methods included the wearing of hair shirts or sackcloth, wearing haircloth undergarments and belts bristling with spikes next to the skin, and flagellating oneself with a scourge made of leather straps or lashing oneself with a whip, such as the *sraoshō-karana* of Persia (see also *Purification rites and customs*, above).

Objects used in rites of passage. Most of the objects noted above have played or still play a role in rites of passage. Such objects play a secondary role in all such rites, which include rites of initiation, marriage, and death.

Objects of initiation and ordination rites

Circumcision in pre-Hellenistic Egypt and among the Hebrews, Muslims, Ethiopians, and certain primitive peoples was and is performed with a flint-blade knife, with some other kind of sharp knife, perhaps of metal, with a razor, or (as in Africa) with a pair of scissors. Among the Zulus and other African tribes, bull-roarers were launched on such an occasion of initiation. In the Brahmanism, Zoroastrianism, and Parsiism of the Indo-Iranian world, a sacred cord (Pahlavi *kuštīʃ* Sanskrit *yajñopavīta*) is the mark of initiation; in Iran and among the Parsis (Zoroastrians in India), the *kuštī* is wound around the torso, and in India the *yajñopavīta* is passed diagonally from shoulder to waist. Among the Parsis, including the women, the cord is made of strands of lamb's wool or of goat's or camel's hair, and in India the material varies according to caste and may be cotton, hemp, or wool. In addition, the Zoroastrians and Parsis wear a sacred shirt (*sudra*) made of two pieces of white cambric stitched together. For ordination, a shawl, a cotton veil (*padān*) to cover the nose and mouth, and a mace are added; the Brahmanic (Vedic) initiate also receives a tall staff and a black antelope skin. In Sikhism (an Indian religion combining Hindu and Muslim elements, founded by Gurū Nānak in the 16th century), initiations of novices formerly included drinking water into which sugar had been mixed with the blade of a dagger (*khaṇḍā*).

In the initiation of Buddhist monks, the tonsure (cutting the hair of the head) is performed with a razor with a handle, and each initiate receives three red or yellow garments, a belt, a bowl for alms (*pātra*), a filter or ewer (*kuṇḍikā*), an alms-collector's staff (*khakkara*), a needle, a toothpick, and a fan. Japanese Tantric monks are initiated when they are past 50 years of age, at which time they are baptized (*abhiṣeka*) by having water from five *kuṇḍikā* poured on their heads and receive, in addition to the objects listed above, a *vajra* ("thunderbolt"), a wheel (*cakra*), and a conch (*saṅkha*). The principal objects involved in the initiation of Christian priests and monks are the tonsure and sacerdotal vestments. The Buryat shaman receives, in addition to his magical cloak and drum, a four-legged chest (*shiré*) decorated with lunar and solar symbols.

Objects related to rites of marriage

The religious character of marriage is not universal. Objects involved in the ceremonies of betrothal and marriage include jars (*loutrophoroi*) for the water of the prenuptial bath of ancient Greece; metal rings placed on the ring finger of the betrothed or married couple among Hebrews, Zoroastrians and Parsis, and persons in classical Rome and in both Eastern and Western Christianity; the bridal veil, orange (*flammeum*) in Rome and white in the Christian and Slavic worlds; the bride's crown, made first of marjoram and verbena and later of myrtle and orange blossoms in Rome and of various materials in the Christian and Slavic worlds; and the crown held above the heads of the bridal couple in Eastern Orthodox marriage ceremonies. In Roman and Slavic marriage rites a tunic or shirt was used, and in Hindu rites a yellow wool bracelet (*kautukasūtra*) is tied around the wrist of the betrothed girl by her mother.

The marriage ceremony sometimes takes place under a

marriage pavilion or canopy, as among the ancient Etruscans of Italy. The Hebrews first used a closed tent (*huppa*) and later a silk or tapestry canopy to symbolize the nuptial chamber. Hindus and Parsis use a tent or pavilion (*pandāl*), in which the bridal couple are initially separated by a curtain. Among the Sikhs, a paper parasol (*agast*) is rotated continually over the head of the bridegroom.

In some areas, particularly in contemporary Hindu India, a swing (*dolā*) is set up under the *pandāl*, on which the couple seat themselves after the official ceremony. The seesaw here symbolizes prosperity, love, and the union between earth and sky. The *aiōra* ("swing") used in the ancient Athenian Dionysiac festival, the swings of the spring festivals at Puri (Orissa) and in Thailand, also have similar symbolic connotations. During the winter solstice, a Vedic sacrifice (*hotr*) is performed on the swing (*preṅkha*).

Objects related to funeral rites

Except for Brahmanic and Buddhist ritual suicides by drowning, which require neither ceremony nor funeral apparatus, there are three methods of disposing of dead human bodies: cremation, stripping of the flesh, and inhumation, performed with or without embalming. These methods have coexisted and still coexist throughout the world. The preparation of the corpse often depends on the method adopted, which, in turn, governs the objects and instruments used. In Japanese sects, particularly in the Shingon and other Buddhist sects, a razor (made of gold in the Jōdo sects) is used for an actual or simulated tonsure of the head of the deceased. A mirror, used in magic to detect evil spirits, figured in the judgment of souls in ancient China. A copper mirror was placed under the head of the dead of pre-Hellenistic Egypt; one of bronze was placed near the head in Buddhist Japan. In Vedic and Brahmanic India, thin pieces of gold were used to close the facial and bodily orifices, and pieces of jade served the same purpose in ancient China. Mortuary masks made of gold, bronze, hard stone, many-coloured terra-cotta, and other materials were used at Mycenae, in pre-Hellenistic and later in Coptic (early Christian) Egypt, in Peru, and other places to cover the face and sometimes the chest. Elsewhere, a cloth covering the face or a shroud, which often was red, was considered sufficient. Pieces of money to pay for the passage from this world to the next were placed in the mouth of corpses in ancient Mycenae, Greece, and Rome and in a pouch in Japan.

Corpses have been borne to funeral sites by various means. Among primitive peoples and in Tibet, they are carried on the back or in the arms, and among the Jews, Muslims, Parsis, Slavs, and Hindus they are carried on biers, which are sometimes richly decorated and are either put in a tomb or destroyed. In modern Western countries, the funeral chariots of Rome and elsewhere have been transformed into motor hearses, while the contemporary Chinese and Vietnamese use carts that have been specially fitted out. Funeral boats were used in pre-Hellenistic Egypt, in ancient Scandinavia, and in the Pacific islands; Venetians of Italy still use gondolas for funeral rites. The sledge was used in the Kurgans of southern Russia.

Cremation practices and objects

When cremated, the corpse is often burned with its bier. In the Buddhist world, as, for example, in Cambodia and Thailand, it is burned in a wood and paper coffin made in the form of a sacred animal, with a cloth canopy surmounting the pyre. If the ashes are dispersed after cremation, as in India, they are collected in a cinerary urn. The form and composition of such urns have varied considerably, being made of terra-cotta, stone, porphyry, alabaster, bronze, silver, gold, ceramic ware, and other materials. The urn is placed in the grave, as in ancient Assyria and elsewhere, on a bronze or terra-cotta support (usually an armchair) and lowered into a large jug, as among the Etruscans, or in the niches of the cineraria (places containing ashes of cremated bodies), columbaria (vaults containing urns of cremated bodies), or catacombs, as in Etruria (in Italy), Greece, and Rome. Among the Zapotec of Mexico, the ceramic urn was placed in the niches of cells, the *mogotes*, made beneath hills set aside for the purpose, a practice also observed by the Mosquito Indians of Nicaragua. In Buddhist countries the urn is often displayed on the domestic altar, and in Tibet the imperfectly calcined bones are ground up and mixed with clay and the mixture is molded into the form of a votive offering (*tsha-tsha*), which is placed in the niches of the funeral *stūpa (mchod-rten)*. In ancient southwestern India the terra-cotta "feminine" urns had a pair of "breasts" formed by two bowls stuck onto the bulge of the urn.

Stripping the flesh of the corpse generally does not require the use of specific objects, since it is the work of vultures or sometimes of pigs, dogs, or other animals. The Parsis, however, build "towers of silence" (*dakhma*) for the purpose, to which they accompany the deceased with a pot containing fire.

Use of coffins, burial boats, and sarcophagi

Bodies have been and still are sometimes buried without coffins, as in Rome, where they were put into pit tombs. Among primitive and prehistoric peoples, ancient Egyptians, and the people of the Harappā civilization (*c.* 2500–1700 BC) of the Indus, the corpse was wrapped in a mat made of plant fibres. Coffins are sometimes carved or painted, and the crudest ones—such as those used by ancient Romans and primitive peoples—are made from hollowed-out tree trunks. Some coffins are modelled according to the human form, such as the colourful wooden coffins of pre-Hellenistic Egypt or the Chinese coffins covered with jade mosaic of the 2nd-century-BC Han dynasty. The majority, however, are oblong and made of wood; in ancient Greece, coffins were made of cypress. Tibetan coffins (*ro-sgam*) and Japanese Buddhist and Shintō coffins, however, are cubical, with the corpse placed in a sitting or crouching position. Among certain coastal peoples—*e.g.*, the Vikings—the deceased is either buried in his boat or put out to sea and cremated with it. Sarcophagi—used in many civilizations—were made of various materials; terra-cotta in Etruria, Greece, southern India prior to the 2nd century BC, and Japan; wood and stone in Japan; and marble in late Rome and in the Christian world. They are often richly decorated with symbolic or allegorical carvings and are frequently very colourful. In ancient Egypt the viscera were placed separately in canopic (burial) jars. The Etrurians also used such jars, the covers of which were decorated with the portrait of the deceased.

From prehistoric times, the deceased was accompanied by ordinary objects placed either in the coffin or in the grave itself, the most common of which were drinking cups, pitchers, cups or vessels for solid food, weapons, tools and ornaments, and jewelry. Ancient Chinese collections of funerary objects of high quality have been exhumed, but the most complete outfitting of the dead was that of the Egyptian tombs, which is completed by scenes painted or carved on the interior walls of the rooms of the tomb. Funeral models of houses, wells, farms, herds, and armies were used in the Han (206 BC–AD 220), T'ang (618–907), and Ming (1368–1644) periods of China as well as in ancient Egypt. Figurines representing the deceased were included among Egyptian funerary objects, along with figurines representing his retinue; in China the retinue figurines included dancers, musicians, and soldiers (*ming-ch'i*). The models were probably substitutes for the servants who formerly had been sacrificed in the royal tomb. For a long time the Chinese figurines were made of ceramic decorated in many colours, but in more recent periods (*i.e.*, after the revolution of 1911 and during the 19th century) they were straw effigies.

Some of the individual objects used in funeral rites include situlae, Roman and Egyptian bronze libation jars with a handle on the tops; Indian Brahmanic terra-cotta jars with perforated bases, which are broken after their use in the aqueous purification of the pyre; and cages containing birds (Buddhist Japan), sometimes eagles (ancient Rome), released near the tomb after burial. There are also the objects used in postmortem rites, such as the tablet of the ancestors (Japanese *ihai*) in China, Japan, and Vietnam and the miniature straw boat, flat-bottomed and with a curved prow, which is set afloat with a bit of candle and food during the Japanese Shinto festival of lights (Bon Matsuri), returning the spirit of the ancestor to the land of souls after three days' visit (see also *Rites of Passage*, above).

Sacrificial sites

Objects used in sacrifices and in sacred meals. The most elementary type of site in which a sacrifice is performed is

simply a massive rock or a hilltop, with no accoutrements. Menhirs (*e.g.,* the Hebrew *matzeva,* a conical stele rubbed with oil at the top), megaliths, and sacrificial posts (*e.g.,* the Vedic *yūpa*)—which are widespread throughout the primitive world—are also quite rudimentary. Altars, properly speaking, are set up either on sacrificial sites or in temples and may be either hollowed out in the earth or raised or constructed. Both of these categories are unknown in Africa and South America, where sacrifices are made on the ground or on a bed of sand. The first category includes the *vedi* ("altar") of Vedic rites, trenches, pits, and ditches dug in the earth. Some of the hollowed-out sites are used for a sacrificial fire and some for collecting victims' blood, as in Greece, pre-Sāsānid Iran, and pre-Islāmic Arabia. The altar is most often a table with one, three, four, or more legs. The top may be smooth, or it may be provided with drains for blood and liquid libations or with dishes to hold solid offerings, such as the firstfruits—*e.g.,* the *kernoi* (small sacrificial pots) of the pre-Hellenic Aegean civilization.

The altar may be round or oblong or may imitate other forms, such as the Indian Vedic altar, which was made in the form of a bird with spread wings. Altars are usually fixed in place and are made of various materials: clay (pre-Columbian religions of Central America); terra-cotta (*kernos*) and stucco-covered sun-baked bricks (religions of ancient Greece); fired bricks (Vedism in ancient India); wood (Buddhism and Shintō of Japan, primitive religions of Polynesia, and Christianity in Western and Nestorian—an Eastern independent church—churches until the 10th or 11th century); wood plated with metals, such as bronze and gold (the religions of the Hebrews and Byzantine Christians); and metals, such as iron (Germanic religion), bronze (ancient Near Eastern religions), and gold (5th- and 6th-century Byzantine Christianity). Most commonly, however, altars are made of stone slabs resting horizontally on legs, columns, or lateral supports, although the pre-Sāsānid Iranian slab altar (*āḍōshi*) rested on a pedestal. The Christian altar is square or oblong; that used in Greek hero worship was rectangular, as was the altar of pre-Hellenistic Egypt, whic was made of alabaster. Some altars, such as the marble Altar of the Earth at Peking, are cubical, and others, such as the Altar of Heaven at Peking and ancient Phoenician altars, are cylindrical. Occasionally, as in Greece, they are hollow and contain the ashes of burnt offerings. The Roman Catholic altar is required to contain a stone, no matter what the predominant material may be.

A throne may be a special form of altar and may be either a true piece of furniture fashioned in wood or metal or a seat carved out in rock. It also may surmount a stele, as in northern Vietnam and Bali.

Equipment for sacrifices

Sacrificial weapons, like the utensils, vary according to the nature of the sacrifice. The most common weapon is the knife, which is used to slit the throat of the human or animal victim, a practice observed, for example, by Semites, Muslims, and ancient Greeks. Sometimes the knife is cast into the sea after use. An ax involved in the Athenian Bouphonia ("Ox-Slaughtering Festival") was carried to the tribunal of the Prytaneum (the town hall, containing a community altar or hearth), inspected, and then submerged in the same way. Sometimes a poniard or dagger was used, such as in the Mithraic sacrifice of a bull; a ritual knife (*khaḍga*) shaped like a sickle, with the outer edge forming the cutting edge, is used in the sacrifice of black goats to Kālī (a Hindu goddess who is the consort of Śiva) in Calcutta. In the great imperial sacrifice of the horse (*aśvamedha*) of Vedic India, a gold-ornamented knife was used to sacrifice the horse, but knives of copper and iron were used for other animals. In the sacrificial rites of contemporary primitive peoples, a sword, which varies in size and form, generally is used. In ancient Iran the victim was slaughtered with a log or pestle. In all sacrificial rites, it should be noted that a flow of blood is always necessary, even when the victim is clubbed.

Sacrificial victims are also very frequently burned or else are cooked for a communal meal. Vessels for holding and maintaining the sacrificial fire may be used in such situations. Two such vessels have been well described in religious literature: the Vedic Indian vessel (*ukhā*) made of earth and fired in a pit on the sacrificial grounds and the urn (*ātash-dān*) of pre-Sāsānid Iranian fire altars. Sometimes the ashes were collected in cauldrons (the ancient Hebrews), and occasionally the viscera were placed separately in a gourd (Africa) or on a tray (pre-Hellenistic Egypt and contemporary Africa). When intoxicating beverages—such as the Avestan Iranian *haoma* and the Vedic Indian *soma*—are made at the same time as the sacrifice, the inventory of ritual objects necessarily includes the stones for pressing the plants, a wooden vat, a filter, and a libation cup at the fire.

Types of objects in libation rites

Three types of objects used in ablution and libation rites may be distinguished. First are the containers for storing liquids, such as water, fermented liquor, wine, and blood. A second type includes utensils—*e.g.,* spoons and ladles—used for drawing off liquids, which are fashioned out of pieces of wood of different, although ritualistically defined, varieties. The third type comprises the containers used directly for ablutions, libations, and oblations—*e.g.,* the ewers of Sumer, Egypt, and Vedic India; gold, silver, copper, or iron *pātra* of the Vedic and Brahmanic world; Hebrew goblets; cups of various forms, such as the Vedic and Tantric skull cup; the phial (bowl) and patera (shallow libation dish) of the Roman and early Christian worlds, made of gold, chased and engraved metal, semiprecious stones, or glass; the Australian bark *pitchi;* and the ciborium (covered container for the consecrated bread) and chalice (cup containing the consecrated wine) of Roman Catholic, Anglican, and Lutheran worship. The cup of the chalice must be made of gold, silver, or vermeil (gilded silver, bronze, or copper).

Other objects pertaining to sacrifice

The sickle for harvesting plants, a winnowing basket for preparing grain offerings, a reed broom for cleaning the sacrificial area, the scoop for collecting ashes used in Vedic India and by the Hebrews (who made it of gold or bronze), and baskets for presenting offerings of fruit or cakes are among the many other objects used in sacrificial rites. In order to consecrate such offerings, a priest of ancient Egypt touched them with a sceptre (*kherep*).

Ornaments used in sacrificial rites are of many different types. The adornment of the victim before sacrifice may take the form of gilding the horns, as in ancient Greece, or putting a necklace or garland of flowers on it. The priest may wear a breastplate, as in Egypt, Etruria, and Jerusalem, or a gold ornament—*e.g.,* the Vedic Indian *nikṣa*—around his neck. Divine statues also may be adorned with jewels, diadems, tiaras, and garments consisting of goldworked covers, a practice still observed in southern India, or with ceremonial apparel, a Christian practice observed in the veneration of saints, particularly in Czechoslovakia (Prague), Poland, and France (Brittany). Altars are permanently or occasionally decorated with incense burners, candelabra, and vases of flowers. Artificial flowers have been used on altars in Japan since the 7th century.

Finally, many sacrifices are accompanied by music, which may be viewed either as a protective measure or as an offering of sound. The musical instruments used in worship do not necessarily assume any special form, but they are often played by the priests themselves, as among Hindus, Tibetan Tantrists, and Hebrews, or are reserved for the accompaniment of particular rites. The silver trumpets of the Hebrews and the conches of Indian-influenced countries are used in this way.

Objects used in temple, state, and private ceremonies. A large number of ordinary objects produced especially for the god have been used in the daily worship of divine statues. The most complete and best described rites were those practiced in ancient Assyria and Egypt and those still observed in the Vaiṣṇavite temples of southeastern India. Such objects are identical in form to those ordinarily used by men, although the materials may vary: earthenware jars for "pure" water; table service, which may include plates, trays, bowls, cups, and pitchers; clothing; pots and flasks for salves and perfumes; jewels, ornaments, flower garlands, and metal mirrors; thrones and platforms; a swing; palanquins (enclosed litters), processional chariots, and boats for the god's journeys outside the temple; musi-

cal instruments, such as drums of all sizes, lutes, clarinets, and conches; and parasols, fans, flyswatters, standards, and oriflammes (banners).

The principal ceremony that pertains to the state is the coronation of the king or emperor. In addition to the pomp displayed on such occasions, the most significant objects generally are the containers used in baptizing or anointing the king, such as the sacred conches or antelope horns used for the lustral water in Indian-influenced countries and the Holy Ampulla (flask) for consecration oil, used particularly in France; the throne, which is the essential object of the ceremony in almost all civilizations; and the crown, the sceptre, the hand of justice, and the globe of the Byzantine, Iranian, and Western worlds.

Domestic rites were observed daily in ancient Rome, Brahmanic India, the Buddhist world, China, Japan, and other areas, as they still are in many places. The objects involved in such ceremonies are the same as those used in temple worship. Permanent altars, which are often placed near the entrance, contain statues, the tablets of the ancestors, and offerings of flowers, incense, fruits, and lights.

CONCLUSION

Ceremonial and ritual objects in past times have held and still hold, in many cases, a very important place in the civilizations of the world. From prehistoric times, they have played an integral part in the evolution of the various civilizations on two levels: (1) on the level of rites and rituals practiced in everyday life and (2) on the level of the more solemn and rare cultic and communal rites. From a merely functional standpoint, such objects serve sacred or symbolic purposes; their construction, forms, dimensions, and styles have been, from earliest times, codified. Some have been so closely associated with the divine or the sacred that they have been considered either a symbolic manifestation of the deity or an actual manifestation of the deity itself. In general, however, they lose in the course of time this particularistic characteristic. In this process, they generally survive only in a formal sense, and thus henceforth are devoid of any sacred power. (J.Au.)

Religious dress and vestments

Religious dress and vestments, broadly understood, include a wide range of attire, accoutrements, and markings used in religious rituals that may be corporate, domestic, or personal in nature. They comprise types of coverings all the way from the highly symbolic and ornamented eucharistic (Holy Communion) vestments of Eastern Orthodox Christianity to tattooing, scarification, or body painting of members of primitive (preliterate) societies. Some types of religious dress may be used to distinguish the priestly from the lay members of a religious group, or they may also be used to signify various orders or ranks within a priesthood. Some religious communities may require that religious personages (*e.g.*, priests, monks, nuns, shamans, priestesses, and others) garb themselves with appropriate types of religious dress at all times, whereas other religious communities may only request that religious dress be worn during rituals.

The meaning of religious dress and vestments

In theocratic societies, such as Judaism and Islām, religious sanctions govern what may and may not be worn by members of the community; and religious dress embraces not only what is worn by a prayer leader but also what is worn by his congregation outside as well as inside a place of worship. In many traditions, habits serve to identify monastic groups. Indeed, in the latter case, the function of religious dress is more akin to heraldry as a form of symbolic identification than to liturgy, with its ritualistic symbolic motifs.

In a more restricted sense, religious vestments articulate a liturgical language as part of a figurative idiom shared with other religious symbols—*e.g.*, icons (images), statues, drama, music, and ritual. According to the richness of the liturgical or ritual vocabulary employed, the more feasibly can a symbology of vesture be attempted. This is especially the case with Eastern Orthodoxy, whose predilection for symbolical theology has spread from sacraments to sacramentals and everything associated with worship, including dress. With allegory paramount in the Middle Ages, the Western Church could not escape attributing symbolical values to garments whose origin may have owed little to symbolism. From the liturgical writer Amalarius of Metz in the 9th century to the theologian Durandus of Saint-Pourçain in the 13th–14th century sacerdotal vestments, in particular the stole and the chasuble, were viewed as symbols and indeed operated as such in a way that still influences current usage. Thus, because the stole is a yoke around the neck of the priest and he should rejoice in his servitude, on donning or doffing it he kisses the emblem of his servile status.

The notion of dress as a substitute skin and, hence, as an acquired personality temporarily assumed has been widespread in primitive religion; such practices in shamanism have been widely observed in Arctic and Siberian regions. The use of a substitute skin in religious ritual is also explicit in the cultic actions of some advanced cultures, such as in the rite of the Aztec maize goddess Chicomecoátl. A virgin chosen to represent Chicomecoátl, after having danced for 24 hours, was then sacrificed and flayed; and the celebrant, dressed in her skin, re-enacted the same ritual dance to identify with the victim, who was viewed as the goddess.

Religious dress may also serve a memorial function, as in the case of the religious leaders (mullahs) of the Shīʿites (Muslim members of the party of ʿAlī), whose black gowns allude to the sufferings of Ḥusayn (ʿAlī's son by Fāṭimah, Muḥammad's only surviving daughter), who was martyred at Karbalāʾ, in modern Iraq, in AD 680. In the Eucharist, which is both a thanksgiving and a reenactment of the sacrifice of Christ on Golgotha, the chasuble (outer garment) worn by the celebrant depicts scenes from the Passion on the orphrey, the name given to the elaborately embroidered strips stitched on the chasuble. The fringes on the Jewish prayer shawl witness to "the commandments of the Lord" in Numbers, chapter 15, and remind the worshipper that he has covenanted to observe them.

TYPES OF DRESS AND VESTMENTS IN WESTERN RELIGIONS

Judaism. *Early sacerdotal dress.* Jewish vesture is an amalgam of very ancient and extremely modern religious dress. Originally, sacerdotal dress was probably varied and complex, but, after the destruction of the Second Temple in AD 70 and the subsequent disappearance of the Temple offices, many garments associated with priestly functions passed into oblivion. Chief among these offices was that of the high priest. In addition to the usual Levitical garments (those of the priestly class), the high priest, while officiating, wore the *meʿil* (mantle), the ephod (an upper garment), a breastplate, and a headdress. The *meʿil* was a sleeveless robe of purple the lower hem of which had a fringe of small gold bells alternating with pomegranate tassels in red, scarlet, purple, and violet. The ephod—an object of much controversy—probably consisted of a wide band of material with a belt to secure it to the body, and it was worn over the other priestly garments. Most important was the breastplate (*ḥoshen*), which was square in outline and probably served as a pouch in which the divinatory devices of Urim and Thummim were kept. Exodus, chapter 28, verse 15, specifies that it was to be woven of golden and linen threads dyed blue, purple, and scarlet. Because of its oracular function, it was called the "breastpiece of judgment." On the face of the breastplate were set 12 gems in four rows, symbolizing the 12 tribes of Israel. These stones were a sardius, a topaz, and a carbuncle in the first row; an emerald, a sapphire, and a diamond in the second; a jacinth, an agate, and an amethyst in the third; and a beryl, an onyx, and a jasper in the fourth. The identity, sequence, and objects of representation of these stones are matters of controversy. Worn over the ephod, the breastplate was slung from the shoulders of the wearer by golden attachments. On his head the high priest usually wore a *mitzenfet* (either a tiara or a turban), except on Yom Kippur ("Day of Atonement"), when he wore nothing but white linen garments upon entering the Holy of Holies (the inner sanctuary).

High priestly garments

Later religious dress. Later religious dress of Judaism after the fall of the Temple in AD 70 reflects usages

The use of phylacteries and fringes

that predate that event but were continued in Judaism at the synagogue. Included among such garments are *tefillin* (phylacteries) and *tzitzit* (fringes), which have certain features in common. The name phylacteries is sometimes thought to point to a prophylactic origin, but the term is actually a translation of the Hebrew word for "frontlets" (*totafot*). Phylacteries are worn in obedience to the commandment found in Deuteronomy, chapter 11, verse 18, and Exodus, chapter 13, verses 9 and 16: "And you shall bind them [*i.e.,* the words of God] as a sign upon your hand, and they shall be as frontlets between your eyes." This implies that there should be two phylacteries: one to be worn on the arm, the other on the head. Both kinds consist of a small black box of hide containing a manuscript and are secured to the respective parts of the body by leather thongs. On the sides of the head *tefilla* is the Hebrew letter ש, the first letter of Shaddai (Almighty). Both boxes are secured by leather thongs. The practice can be dated at least as far back as the 3rd century BC. The knotted thongs indicate a prophylactic purpose—*i.e.,* to protect the wearer against demons. Likewise, the wearer of these objects was, for the prayer's duration, under the protection of the Almighty, whose name he bore. The importance of knots in Semitic magic is also alluded to in the Qur'ān (the Islāmic scripture).

Something similar obtains in the case of the *tzitzit* (fringes), or "twisted cords." The wearing of fringes is in obedience to a commandment in Numbers, chapter 15, verses 38–40: "It shall be to you a tassel to look upon and remember all the commandments of the Lord, [and] to do them." The fringes were attached to the outer garment with no attempt at or reason for concealment. Later, because of persecution, they became an inner garment, enabling the wearer to observe the Law clandestinely. This garment, which is not entirely obsolete, is styled *arba'kanfot* ("four corners") in allusion to Deuteronomy, chapter 22, verse 12 ("you shall make yourself tassels on the four corners of your cloak with which you cover yourself"), although no literary reference to its use can be traced further back than the 14th century.

The tallith, or prayer shawl, has the four fringes also, but it is confined to synagogal use and, even there, is limited to the morning service, whereas the *arba'kanfot* is worn all day. Both silk and wool are used, but the woollen tallith is preferable, with white as its ground colour. In the 20th century the tallith is worn like a scarf and is sometimes pulled over the head to aid in concentrating during prayer. Formerly, however, it was always wrapped around the head. In orthodox Judaism, the head is invariably covered during worship, usually by a skullcap known as a *yarmulka* or *kappel.*

Because a Jewish male is not supposed to walk more than four cubits (six feet) with his head uncovered, a religious Jew will wear the skullcap clipped to his hair, and indeed he may wear it all day because he believes himself to be in the presence of God at all times.

The dress of rabbis never conformed to precise standards. Current practice approximates modern Genevan (Protestant) conventions (gown and bands). The Jewish Reform movement, which began in Germany, further emphasized the Protestant character of rabbinical dress, and Reform rabbis differ little in this respect from ministers of various Protestant churches. Both cantor (*hazzan*) and rabbi now use the black gown and round black hat, which came into use during the 19th century.

Garments used on Yom Kippur

On Yom Kippur, it was the custom for participants to wear a *sargenes,* or white garment, emphasizing that Yom Kippur was an occasion not only of repentance but also of grace, for which festal wear was appropriate. Emphasis on the atoning aspect of the occasion, however, led to the *sargenes* being interpreted as *takhrikhim,* or graveclothes, which are worn to aid the worshipper toward a mood of repentance, a practice also adopted by the *hazzan* on two other occasions and by the host at the seder (meal) on Passover (a feast celebrating the Exodus of the Hebrews from Egypt in the 13th century BC). Officiants at the Yom Kippur service still dress in white robes. Shrouds are normally of unadorned white linen, following the sumptuary ruling of the 1st-century-AD rabbi Gamaliel the Elder. To the shroud may be added the tallith used by the deceased, but with the fringes removed or cut, because the prescription governing their use applies only to the living. Both liturgical vesture and everyday clothing must conform to the Mosaic requirement that forbids the combination of linen and wool in the same garment (see also JUDAISM).

Christianity. In the pre-Constantinian church (before the early 4th century), no distinctive liturgical dress was worn, and the Eucharist (Holy Communion) was celebrated by priests whose dress did not differ from that worn by lay members of their congregations. Present liturgical vestments in Roman Catholic and Orthodox churches derive from a common origin—*i.e.,* the garments that were fashionable in the late Roman Empire. After the Schism of 1054, however, they each followed separate courses (see also CHRISTIANITY).

Roman Catholic religious dress. A distinction is made between the insignia of ecclesiastical and sacerdotal office in the hierarchy and the functionally and symbolically significant liturgical robes. After the barbarian invasions of the Roman Empire from the 4th century on, fashions in secular dress changed, and thus the clergy became distinct in matters of dress from the laity. Certain robes indicate a position in the hierarchy; others correspond to function and may be worn by the same individual at different times. The most important vestment among the insignia is the stole, the emblem of sacerdotal status, the origin of which is the ancient *pallium.* The stole originally was a draped garment, then a folded one with the appearance of a scarf, and, finally, in the 4th century, a scarf. As a symbol of jurisdiction in the Roman Empire, the supreme pontiff (the pope, or bishop of Rome) conferred it upon archbishops and, later, upon bishops, as emblematic of their sharing in the papal authority.

Liturgical garb

The distinctive garb of the liturgical celebrant is the chasuble, a vestment that goes back to the Roman *paenula.* The *paenula* also was the Orthodox equivalent of the chasuble, the *phelonion,* and perhaps also the cope (a long mantle-like vestment). In its primitive form the *paenula* was a cone-shaped dress with an opening at the apex to admit the head. Because ancient looms were not wide enough to make the complete garment, it was made in several parts sewn together with strips covering the seams. These strips, of contrasting material, developed into the orphrey (embroidery), on which much attention was later lavished. Next in the hierarchical order after the priesthood were the diaconate and subdiaconate, whose characteristic vestments were, respectively, the dalmatic (*dalmatica*), a loose-fitting robe with open sides and wide sleeves, and the tunic (*tunica*), a loose gown. A priest wore all three, one over another. Under these he wore the alb (a long white vestment), held round the waist by a girdle, and around the neck the amice (a square or oblong, white linen cloth), with the maniple (originally a handkerchief) on the left arm. Although the deacon used a stole, the subdeacon did not. In the formative period of liturgical dress, these practices were in the process of becoming normative. During the 9th to the 13th century the norms now familiar were established. The chasuble became an exclusively eucharistic garment; the cope, excluded from the Eucharist, became an all-purpose festive garment.

Next in importance to the chasuble is the cope, a garment not worn during the celebration of the mass but rather a processional vestment. It is worn by the celebrant for rites of a non-eucharistic character, such as the Asperges, a rite of sprinkling water on the faithful preceding the mass. The origins of the cope are not known for certain by liturgical scholars. According to one theory, it derives from the open-fronted *paenula,* just as the chasuble derives from the closed version of the same garment. (The subsequent wide divergence between the two vestments need not preclude a common origin.) Unlike the chasuble, the form of which has never stopped changing, the evolution of the cope was complete before the end of the Middle Ages. Cope chests, based on the quadrant of a circle and designed to preserve the embroidered surfaces by keeping the copes flat, were a common feature of medieval cathedrals. When it is worn, the two sides of the garment are held together by a morse (a metal clasp). The cope occupied an intermediate

position between liturgical and nonliturgical vestments, the most important of which was the cassock, the normal dress of the priesthood outside church ceremonies. When engaged in religious ceremonies, the officiant would wear the liturgical vestments over his cassock.

The tiara, the papal diadem or crown apostolic, emerged in the early medieval period; and the mitre (the liturgical headdress of bishops and abbots), the most conspicuous of the episcopal insignia, began as a mark of favour accorded to certain bishops by the supreme pontiff at a somewhat later date.

Like the cope, the surplice (a white outer robe) entered liturgical usage in the Middle Ages as a late modification of the alb. By the 14th century its present role as a choral or processional garment was established. With the passage of time, the length of the garment grew progressively shorter.

Monastic garb

The surplice was also associated with the monastic orders, but vesture distinguished only the order and not the kind of order. Eremitical (hermitic) monasticism allowed no standard form of dress to develop, and only communal monasticism, beginning with the Rule of St. Benedict of Nursia in the 6th century, enabled standardization to become possible. Monastic dress included habit, girdle or belt, hood or cowl, and scapular (a long narrow cloth worn over the tunic). The salient characteristics of monastic dress have always been sobriety and conservatism. The orders proved even more retentive of archaic fashions than the hierarchy, and, in contrast to the deliberate splendour of ecclesiastical vestments, monastic dress was expressive of a renunciation of luxury. The contrast was functional in origin: the menial tasks of the monk related him sartorially to the peasant, whose humble avocations he often duplicated, rather than to the princes and prelates of the church, whose dress reflected the splendour of the ceremonies in which they engaged.

Because of the diversity of the monastic orders, only a summary account of their vesture may be given. The Benedictine mantle was black, fastened with a leather belt; but the Cistercians—reformed Benedictines—eschewed any dyed material and instead dressed in undyed woollen material, which was off-white in colour. In the course of time this became white, a tacit relaxation of the primitive austerity adopted as a protest against "luxury." Carthusians, a contemplative order founded in the 11th century, likewise wore white. In the 13th century the mendicant orders (friars) emerged. The Franciscans, founded by St. Francis of Assisi, first used a gray habit, which in the 15th century was exchanged for a brown one; in spite of this change they continued to be known as the Grey Friars. The Carmelites, an order founded in the 12th century, became known as White Friars. Dominicans, founded by St. Dominic from Spain, adhered from the beginning to a black robe over a white gown. Canons regular (communal religious persons living under vows), although ordained, lived like the orders under a rule, and the Augustinians (several orders following the Rule of St. Augustine) are styled Black Canons in contradistinction to the Premonstratensians, or White Canons, an order founded by St. Norbert in the 12th century. Because the office (prescribed prayers) took up so much of a monk's time, his choir robes were almost as important as his day clothes. Surplices were worn in choir with an almuce over; this last was a lined shoulder cape designed to help the wearer resist the cold of medieval churches.

Nuns' costumes were similar to those of monks, the chief difference consisting in the replacement of the hood by a wimple (collar and bib) and head veil. Habits are white or black or mixed, and this remained unaltered until the 17th century, when the Sisters of St. Vincent de Paul introduced blue. This exception remained unique; nuns' habits retained a markedly medieval aspect until reformed by the second Vatican Council (1962–65).

The cassock has its origin in the barbarian *caracalla*, a robe favoured by the Roman emperor Bassianus (reigned 211–217), who came to be known as Caracalla because of the garment he habitually wore. Worn by the clergy as early as the 5th century, it became in time the standard day wear for prelates and priests, hierarchical rank being indicated by colour: bishops, archbishops, and other prelates wore purple; cardinals, red; the pope, white; and ordinary clergy, black (see also ROMAN CATHOLICISM).

Garb of priests, bishops, and monks

Eastern Orthodox religious dress. The Middle Ages also witnessed the evolution of Eastern Orthodox vestments into approximately their present form. The eucharistic garment corresponding to the chasuble was the *phelonion*, with variant forms in the Greek and Russian churches. The *sticharion*, which is held by the *zōnē*, or girdle, corresponds to the alb. The cuffs, or *epimanikia*, which fit over the *sticharion*, bear little or no resemblance to the maniple. The *epitrachēlion* is the Orthodox equivalent of the stole, but it hangs straight instead of being crossed over the chest, as is the case with the stole in Western churches. On the deacon, the *epitrachēlion* is pinned to the left shoulder and hangs in front and behind; with this exception, the deacon's vesture is identical with the priest's. The bishop wears an *omophorion*, whose shape and manner of wearing are closer to the original *pallium* than either the stole or the *epitrachēlion*. In place of the *phelonion*, since the 16th century, the bishop uses a dalmatic known as the *sakkos*. The *epigonation*, or rhombus-shaped portion of silk hanging to below the right knee, is common both to bishops and archimandrites (head abbots).

The monastic habit of the monk differs according to which of the three grades he occupies. The fully professed monk wears the great, or angelical, habit, which consists of the inner and outer rhasons, girdle, cowl (with veil), *analvos*, and *mandyas* (mantle). The inner rhason corresponds to the cassock and, like it, is used by the secular clergy. The outer rhason, a wide-sleeved garment, is black in the Greek Church but variable in colour in the Russian Church among the secular clergy (*i.e.*, those who minister in parishes). The *analvos* (shaped like the Western scapular, although historically unconnected with it) differentiates the full, or perfect, monk from the other grades, and its substance must be of animal, nonvegetable origin to remind the wearer constantly of death. The *mandyas* is the bishop's cloak (for non-eucharistic occasions), and in the Russian Church its use is granted to monks of the intermediate grade, although this license does not obtain in the Greek Church. In neither church may the *mandyas* or *analvos* be worn by monks of the lowest grade. Unlike Western orders, Orthodox monks dress only in black, but they share the same sartorial conservatism, their habits having remained unchanged in essentials from medieval times to the present (see also EASTERN ORTHODOXY).

Protestant religious dress. The Reformation of the 16th century varied in intensity from one country to another, and the fate of liturgical vesture suffered accordingly. With the rejection of the dogma of transubstantiation (the Roman Catholic teaching that in the Eucharist the substance of the bread and wine is changed into the body and blood of Christ, with the properties of the bread and wine remaining the same), the use of the mass garments might have been expected to be eliminated, but, wherever an altered eucharistic doctrine survived, an attenuated liturgical vesture contrived to survive with it. In the case of the Anglican and Lutheran churches, a paradoxical situation emerged whereby, in the latter, pre-Reformation practices (*e.g.*, use of crucifixes) survived alongside a Reformation theology, whereas, in Anglicanism, a Catholic theology survived along with a repudiation of Catholic rites. The Lutherans rejected the insignia of a celibate clergy but retained the chasuble for Communion services and the surplice and alb for other services.

Anglican and Lutheran liturgical vesture

Bishops in both Lutheran and Anglican communions retained the cope. The different editions of *The Book of Common Prayer* (the Anglican liturgical book) attest to 16th-century reforms and the rising power of Puritanism, a 17th-century reform movement; the use of vestments declined in consequence. The cathedrals, however, maintained liturgical vestment standards to a certain degree, even when the last vestiges of liturgical propriety had been extinguished in the parishes in the 18th century. The cope became the High Church (liturgically oriented) vestment *par excellence*, worn by bishops not only processionally but even during Communion. Many views about the ceremonial revival of the 19th century have not in all respects been accurate; and followers of Edward Pusey, a leader

of the Catholic revival known as the Oxford Movement, and ritualists sometimes blundered not from excess of archaeological zeal as has been commonly supposed but rather because they were inordinately influenced by their sociocultural environment. This may be less immediately obvious in the case of vesture than in architecture, but one result of overreacting was the loss, in the 19th century, of the customary dress of the clergy. The gown and cassock, as street attire, were allowed to fall into desuetude because in Puseyite views the gown was Genevan, whereas in reality it was the reverse. Another instance lay in the adoption of the (local) Roman biretta, introducing an Italian fashion even though adequate indigenous precedents were not lacking.

The gown, now inseparably associated in the popular mind with Genevan (Reformed) divines, was in fact opposed by these same divines in England and Scotland in the 17th century. In spite of this, standard vesture in Presbyterian churches is now the black gown and white linen bands over cassock and cincture, with the academic hood added for preaching services as a mark of learning appropriate to the pulpit, and a stole or scarf (see also PROTESTANTISM).

Modern changes in religious dress and vestments. With a change in emphasis, chiefly expressed in the episcopal use of the cope, Episcopalian usage in the first half of the 20th century differed little from Catholic rules except in Anglo-Catholicism, in which deliberate archaism imposed an adhesion to Baroque (17th to early 18th century) models, themselves superseded within Roman Catholicism. The Liturgical Movement of the 20th century has exercised an influence beyond the boundaries of the church in which it originated, and modern clerics of different denominations increasingly resemble one another sartorially because all have had recourse to the same sources of liturgical inspiration.

In Roman Catholicism, the formative period of religious dress was over before the Reformation, and Reformation influence was indirect—via the impetus supplied by the Counter-Reformation, which made Baroque its official art style. The emphasis on richness of material, excessive decoration, and preoccupation with surface set in motion a process of decline that was not arrested till the 20th century. The degeneration of the Gothic chasuble with its pointed folds into a stiff fiddle-backed, overembroidered vestment had begun as early as the 13th century with the practice of elevating the Host (sacrificial elements) in the mass. The elevation of the Host entailed the folding back on the celebrant's shoulders of the sides of the chasuble. The flexibility of the early chasuble permitted this, but, to facilitate the elevation, more and more material was removed from the sides until the garment became a caricature of its primitive form, distorted beyond recognition and its vestigial portions—dorsal (back) and pectoral (front)—came to be viewed simply as canvases for the display of virtuoso embroidery. Undergarments also became what is now viewed as effeminate with the addition of lace, and, although the Liturgical Movement began with a new theology of the Eucharist, its repercussions forced a decline of the Baroque style in dress.

The influence of the second Vatican Council

From the late Middle Ages to the 20th century, the history of religious dress in the Roman Catholic Church has been the history of its rubrical evolution: the regional variants of patristic (early church) and early medieval times were eliminated in the interest of ultramontanism (a theory that advocated a greater authority for the papacy), until the second Vatican Council reversed the process of eight centuries, again sanctioning regional divergences. Council rulings also simplified the use of the mitre and suppressed the use of the maniple altogether. Increased lay participation in the liturgy has led to an extension of lay religious dress in more than one communion. To lay offices such as the verger, who wears a gown over cassock, and chorister, who wears a surplice, Anglicans have added that of the lay reader, who vests in cassock and surplice, with a scarf as his ensign.

The upheavals of the 16th, 19th, and 20th centuries have not had much effect on Eastern Orthodox vesture, and the same canons (rules) prevail today in Orthodoxy as obtained prior to the fall of Constantinople in the 15th century. To ascribe this condition in Eastern Orthodoxy solely to the effects of cultural isolation would be an oversimplification. Suppression of vestments or their alteration is less likely to occur in a church in which such vestments have higher symbolic value attributed to them than in other traditions.

Islām. Islām attaches less importance to liturgical vestments than do most religions, but the social emphasis of the Islāmic faith finds expression in the universal application of the regulations governing dress; *e.g.,* all who enter the mosque remove their footwear, and all going on pilgrimage must wear the same habit, the *iḥrām,* and thus appear in the holy places in the guise of a beggar.

Because Islām recognizes no priesthood in a sense of a class sacramentally set apart, "clerical" functions are discharged by the *ʿulamāʾ,* or "the learned (in the Law)," whose insignia is the *ʿimāmah* (a scarf or turban). The garb of the *ʿulamāʾ* exhibits geographical variations, but the *ʿimāmah* is found everywhere. Two broad regional distributions obtain, with Iraq as the area of confluence between the two. In the western part of the Muslim world, "clerical" dress has tended to become standardized according to the Azhar (Egyptian) pattern: a long, wide-sleeved gown (*jubbah*) reaching to the feet and buttoned halfway down its total length over a striped garment (caftan); and the headgear consists of a soft collapsible cap (*qalansūwah*) of red felt around which is wound a white muslin *ʿimāmah.* In Syria a hard *ṭarbūsh* of the same red shade replaces the *qalansūwah.* Both the *qalansūwah* and the *ṭarbūsh* are provided with a blue tassel. The *jubbah* is usually a sober shade of blue, gray, or brown, and seldom black. Among the Sunnites—from Iraq eastward—the *jubbah* is worn in association with an *ʿabāʾ* (a long, full garment), traditionally of camel's hair and brown or black in colour. This is sometimes secured by a *ḥijām,* or cummerbund. In this second regional variant, the *ʿimāmah* becomes a full turban replacing the cap, or fez. A green turban usually denotes a *sharīf,* or descendant of the Prophet Muḥammad; and among the Shiʿites (the party of ʿAlī) the entire "clerical garb" is black, as a symbol of mourning for the death of Ḥusayn at Karbalāʾ.

The Ottoman Turks, as strict Sunnites, preferred turbans of other colours, which, elaborately wound, served to distinguish the wearer from a non-Muslim. On conquering Constantinople in 1453, they adopted the Byzantine cap and wound the turban around it in demonstration of conquest. The elaborately wound turbans of Persia and India also have a skullcap as a foundation for their folds. The art of winding a turban required no small degree of skill, the wearer fitting the cap over his knee and winding it in that position, whereafter the cap kept the folds in place. To the Prophet Muḥammad is attributed the saying "What differentiates us [in appearance] from the polytheists is the turban." In India the turban has also been worn by non-Muslims, but the Muslim turban has remained distinguishable from the Hindu by the use of the skullcap as its foundation.

Religious dress prohibitions

For all Muslim males, whether Sunnite or Shiʿite, clerical or lay, the wearing of gold or silk is forbidden in consequence of a prescription (Ḥadīth) of the Prophet, whereby the wearing of either was rendered "*ḥarām* [forbidden] for the males of my nation." Footwear must be removed on entering a mosque for fear of defiling the interior with ritually impure substances that may have adhered to the sole of the shoe. This rule applies also to entering a grave; thus, gravediggers and stonemasons must be unshod on such occasions. Because covering the head is a Near Eastern way of showing respect, a head covering should properly be worn in the mosque and even when praying outside the mosque.

When a Muslim purposes to visit the holy city of Mecca at the time of the major pilgrimage (*ḥajj*), he enters on a state of consecration and robes himself in two white seamless garments (*iḥrām*), which may not be exchanged for normal dress until he deconsecrates himself after the conclusion of the pilgrimage ceremonies. To these two garments women may add a veil.

Many of the mystical dervish orders (*ṭuruq*) wear dis-

tinctive robes, frequently with hierarchical differences. In Turkey, headstones are carved in the shape of the headdress distinctive to the order to which the deceased belonged and are tinctured in the appropriate colours. Particularly interesting are the ceremonial robes of the Mawlawiyah order (popularly known in the West as the Whirling or Dancing Dervishes), in which the symbolism of the robes is central to the mysteries of the order. The dervishes wear over all other garments a black robe (*khirqah*), which symbolizes the grave, and the tall camel's hair hat (*sikke*) represents the headstone. Underneath are the white "dancing" robes consisting of a very wide, pleated frock (*tannūr*), over which fits a short jacket (*destegül*). On arising to participate in the ritual dance, the dervish casts off the blackness of the grave and appears radiant in the white shroud of resurrection. The head of the order wears a green scarf of office wound around the base of his *sikke*.

For all Muslims of whatever sect the standard grave-clothes are the threefold linen shroud, or *kafan*: the *izār*, or lower garment; the *ridāʾ*, or upper garment; and the *lifāfah*, or overall shroud. Martyrs, however, are buried in the clothes in which they die, without their bodies or their garments being washed, because the blood and the dirt are viewed as evidences of their state of glory (see also ISLĀM).

TYPES OF DRESS AND VESTMENTS IN EASTERN RELIGIONS

Indian religions. The distinction between ordinary dress and religious dress is difficult to delineate in India because the ordinary members of the various socioreligious groups may often be distinguished by their costumes. For example, Parsi (Zoroastrian) women wear the *sārī* (robe) on the right shoulder, not the left.

Hindu men frequently wear short coats (*angarkhā*), and the women wear a long scarf, or robe (*sārī*), whereas typical Muslim attire for men and women is a long white cotton shirt (*kurtah*) and trousers (*pāʾijamah*). Muslim women also wear a veil called the *burqah*, which not only hides the face but also envelops the entire body. Traditional Sikh (a religion combining Hindu and Muslim elements) dress is an ordinary *kurtah* and cotton trousers, covered by a long hanging coat (*choghah*). The male Sikh is recognized especially by his practice of wearing his hair and beard uncut, the former being covered by a particularly large turban and the latter often restrained by a net.

Hindu religious garments

The Brahmin (Hindu priest) is distinguished primarily by the sacred thread (*upavīta*), which is bestowed on him during his boyhood investiture and worn diagonally across the body, over the left shoulder, at all times. During the water offering to saints, it is worn suspended around the neck and, during ancestor rites, over the right shoulder. Devotees may also wear a tonsure that leaves a tuft of hair longer than the rest (*śikhā*). The *pravrajyā* ("going forth") associated with some *Upaniṣads* (Hindu philosophical treatises) involved a ritual rejection not only of homelife but also of the *upavīta* and *śikhā*. Ascetics usually wear the ordinary loincloth, or *dhotī*, for meditation or yoga (a physical and psychological meditation system), but there is also a tradition of naked asceticism. A teacher (*swāmī*) traditionally wears a yellow robe (see also HINDUISM; SIKHISM; ZOROASTRIANISM AND PARSIISM).

Buddhism. Buddhism became more widespread in Asia than other ascetic and meditational movements, partly because of the strong organization of its monastic communities (*saṅgha*). One of the main outward signs of the *saṅgha*, along with the tonsure and the begging bowl, has always been the monk's robe; "taking the robe" became a regular expression for entering the *saṅgha*. The *saṅgha* was organized in accordance with the traditional code of discipline (*vinaya*), which includes the basic rules regarding robes in all Buddhist countries. These rules are all linked to the authority of the Buddha himself, but at the same time they allow considerable flexibility to cater to changing circumstances.

The Buddhist robe

The robe (*cīvara*) illustrates two main types of religious action, each symbolized by the character of the materials used. First, the wearing of "cast-off rags" was one of the "four resources" of a monk, being an exercise in ascetic humility similar to the other three, which are living on alms, dwelling at the foot of a tree, and using only cow's urine as medicine. The use of rags was later formalized into making the robes out of separate strips or pieces of cloth, but the rough patchwork tradition was carried over into China, where hermit monks in modern times wore robes made of old rags. In Japan, robes have been preserved with designs imitating the effect of patchwork, and robes sewn from square pieces of cloth were nicknamed "paddy-field robe" (*densōe*). This latter term is reminiscent of an old Indian Buddhist tradition according to which the Buddha instructed his disciple Ānanda to provide robes for the monks made like a field in Magadha (in India), which was laid out in "strips, lines, embankments, and squares." In general, whatever the degree of formalization, the rag motif ensured that the robe was to be "suitable for recluses and not coveted by opponents." The second type of religious action associated with the robe stemmed from the permission granted to monks to receive robes or the materials for making them from the laity. This meant that the laity "became joyful, elated, thinking: 'Now we will give gifts, we will work merit . . . ' " (*Mahāvagga* VIII, 1, 36). Thus, the presentation of materials for robes was thought to have the same beneficial karmic effects (toward a better birth in the future) as the offering of food. The practice meant that various good materials were offered as well as rags, and in due course six types were allowed on the authority of the Buddha, namely, linen, cotton, silk, wool, coarse hempen cloth, and canvas.

There are three types of *cīvara (i.e., tricīvara)*: the inner robe (Pāli, *antaravāsaka*), made of five strips of cloth; the outer robe (*uttarāsaṅga*), made of seven strips; and the great robe, or cloak (*saṃghāṭi*), made of nine, 15, or 25 strips.

In order to avoid the primary colours, Buddhist robes are of mixed colours, such as orange or brown. Another common term for the robe, *kasāya*, originally referred to the colour saffron, though this meaning is lost in the Chinese and Japanese derivatives, *chia-sa* and *kesa*. The robe is normally hung from the left shoulder, leaving the right shoulder bare, though some ancient texts speak of disciples arranging their robes on the right shoulder before approaching the Buddha with a question. In cooler climates, both shoulders may be covered with an inner robe, and the outer robe is hung from the left shoulder, as in China.

Sandals are allowed if they are simple and have one lining only, or they may have many linings if they are cast-off sandals. The rules for nuns' robes are similar, but they also wear a belt and skirt. Some special vestments are worn by Tibetan Buddhists, including various hats characteristic of the different sects (see also BUDDHISM).

Chinese religions. Court dress, sacrificial dress, and ordinary dress were all influenced in ancient China by the Confucian-inspired civil religion. The classical text for the Confucian ideal of deportment and dress is Book X of the *Analects*, in which the emphasis is on propriety in every detail, whether at home or in affairs of state or ceremony. The undergarment, for example, was normally cut wide at the bottom and narrow at the top to save cloth, but it had to be made full width throughout for court and sacrificial purposes.

The influence of the Confucian ideal of propriety

Confucius was also said to have insisted on the primary, or "correct," colours—blue, yellow, red, white, and black—rather than "intermediate" colours, such as purple or puce, and to have avoided red for himself because it was more appropriate for women.

Garments used in sacrifices to former kings and dukes were prepared from silk grown in a special silkworm house. According to the "Doctrine of the Mean," the clothes used by ordinary people at sacrifices were "their richest dresses." The fully developed Imperial costume for sacrifices was a broad-sleeved jacket and a pleated apron around the waist. Decorative symbols represented the universe in microcosm and thus the universal sovereignty of the emperor.

Funeral dress was generally white, although the *Shu Ching* ("Classic of History") refers to a funeral at which those who officiated wore hempen caps and variously coloured skirts. According to the *I Li*, mourning dress consists of "an untrimmed sackcloth coat and skirt, fillets

of the female nettle hemp, a staff, a twisted girdle, a hat whose hat string is of cord, and rush shoes." For Mencius, a 4th–3rd-century-BC philosopher, the wearing of a coarse cloth mourning garment was an important aspect of traditional filial piety.

Buddhist robes in China followed Indian tradition fairly closely, though they were noted under the T'ang dynasty (AD 618–907) for being black in colour. Taoist robes, in contrast, were yellow. That this is an old tradition may be seen from the example of the 2nd-century-AD Yellow Turban movement, in which the missionaries and priests wore yellow robes and the followers yellow headdresses.

Japanese religions. The priestly robes of Shintō are an example of the way in which rather normal garments of a formative age became the specialized religious vestments of later times.

Shintō religious garments

They consist of an ankle-length divided skirt (*hakama*) in white, light blue, or purple, depending on rank; a kimono in white, symbolizing purity, and of which there are various types; and a large-sleeved outer robe of various colours that is frequently a *kariginu,* or hunting garment, as used in the Heian period (794–1185). The headgear is a rounded black hat (*eboshi*). The more elaborate "crown" (*kammuri*) has a flat base, a protuberance rising forward from the back of the head, and a flat band curving down to the rear. Within a shrine, stiff white socks with a divided toe (*tabi*) are worn, and, when proceeding to or from a shrine, officiants wear special black lacquered clogs (*asagutsu*) of paulownia wood. Shintō priests carry a flat, slightly tapered wooden mace (*shaku*), which symbolizes their office but otherwise has no precisely agreed upon significance. The dress of *miko* (girl attendants at shrines), whose main function is ceremonial dance, also typically consists of a divided skirt and a white kimono. They carry a fan of cypress wood. Young male parishioners bearing a portable shrine through the streets may wear a kimono marked with the crest of the shrine and a simple *eboshi.*

Buddhist robes continued the general Buddhist tradition, but of particular interest are the ornate ceremonial robes of high-ranking monks, especially in the Shingon and Nichirenite sects; the white robes worn by devotees in the syncretistic Shugen-dō tradition (famous for its *yamabushi,* or mountain priests) during lustrations and similar rituals, symbolizing purity, as in Shintō; and the deep, inverted bowl-shaped hats of woven straw (*ajirogasa*) worn by Zen monks during begging tours.

Many new religions in Japan have carefully manufactured ceremonial vestments based on Shintō or Buddhist models or of mixed or original design. A common feature is the use of fairly simple uniform clothing for all believers during dedicated labour, mass rallies, or acts of worship. In Tenri-kyō, a religion founded in the 19th century by Nakayama Miki, the name of the religion figures prominently on the back of the garment, and, in Nichiren movements, the central symbol *namu Myōhō renge-kyō* ("Homage to the Lotus of the Good Law") may be displayed on a stole hanging from the left shoulder.

(J.Di./M.Py.)

Feasts and festivals

Throughout the history of human culture, certain days or periods of time have been set aside to commemorate, ritually celebrate or re-enact, or anticipate events or seasons—agricultural, religious, or sociocultural—that give meaning and cohesiveness to an individual and to his religious, political, or socioeconomic community. Because such days or periods generally originated in religious celebrations or ritual commemorations that usually included sacred community meals, they are called feasts or festivals.

The terms feast and festival usually—though not always in modern times—involve eating or drinking or both in connection with a specific kind of rite: passage rites, death rites, sacrificial rites, seasonal observances, commemorative observances, and rites celebrating the ending of fasts or fast periods. Fasting, the opposite of feasting, has often been associated with purification rites or as a preparatory discipline for the celebration of feasts and associated rites. Festivals often include not only feasting but also dramatic dancing and athletic events, as well as revelries and carnivals that at times border on the licentious. Depending upon the central purpose of a feast or festival, the celebration may be solemn or joyful, merry, festive, and ferial.

Another term associated with the events and activities of days of sacred significance is "holy day," from which is derived the word holiday. This term has come to mean a day or period of special significance not only in religious calendars (*e.g.,* the Christian Christmas and the Jewish Ḥanukka) but also in the secular (*e.g.,* May Day in Russia and Labor Day in the United States and Canada, both of which holidays celebrate especially the accomplishments of the working class).

This section, though it will concentrate on feasts and festivals in the history of religions, will also give attention to the holidays of what has been termed the secular (or profane) sphere. Most secular holidays, however, have some relationship—in terms of origin—with religious feasts and festivals. The modern practice of vacations—*i.e.,* periods in which persons are "renewed" or participate in activities of "recreation"—is derived from the ancient Roman religious calendar in a reverse fashion. More than 100 days of the year were feast days dedicated to various Roman gods and goddesses. On the days that were sacred festivals, and thus holy days, persons rested from their routine daily activities. Days that were not considered sacred were called *dies vacantes,* vacant days, during which people worked. In modern times, however, vacations (derived from the term *dies vacantes*) are periods of rest, renewal, or recreation that may be sacred or secular holidays—or simply periods of time away from everyday work allowed by modern business or labour practices.

Feasts and festivals, originating in the dim past of man's social, religious, and psychic history, are rich in symbols that have only begun to be investigated in the 19th and 20th centuries by anthropologists, comparative folklorists, psychoanalysts, sociologists, historians of religion, and theologians. Such investigations will not only elucidate mythological, ritualistic, doctrinal, aesthetic, and psychic motifs and themes but will also provide educative insights to modern man, who has been caught up in social and religious forces that he has found difficult to understand. Feasts and festivals in the past have been significant informational and cohesive devices for the continuity of societies and religious institutions. Even when the feasts or festivals have lost their original meanings in doctrinal or mythological explanations, the symbols preserved in the rites, ceremonies, and arts (*e.g.,* pictorial, dramatic, or choreographic) have enabled persons in periods of crisis or transition to preserve an equanimity despite apparent evidences of disintegration within their cultures or societies. Thus, the scholarly investigations of the many and various facets of feasts and festivals will provide different forms of information that will be of help to modern man in achieving an understanding of his origin, identity, and destiny.

The value of studies of feasts and festivals

NATURE AND SIGNIFICANCE

Concepts of sacred times. By their very nature, feasts and festivals are special times, not just in the sense that they are extraordinary occasions but more so in the sense that they are separate from ordinary times. According to Mircea Eliade, a Romanian-American historian of religion, festival time is sacred; *i.e.,* it participates in the transcendent (or supernatural) realm in which the patterns of man's religious, social, or cultural institutions and activities were or are established. Through ritualistic reenactment of the events that inform man about his origin, identity, and destiny, a participant in a festival identifies himself with the sacred time:

> Religious man feels the need to plunge periodically into this sacred and indestructible time. For him it is sacred time that makes possible the other time, ordinary time, the profane duration in which every human life takes its course. It is the *eternal present* of the mythical event that makes possible the profane duration of historical events.

In religions and cultures that view time as cyclical—and this applies to most non-monotheistic religions and the cultures influenced by them—man understands his

Man's view of himself in cultures that see time as cyclical

status in the cosmos, in part, through special times (*e.g.*, New Year's festivals) celebrating the victory of order in nature over chaos. New Year's festivals have been celebrated in recorded history for more than five millennia. In ancient Mesopotamia, for example, Sumerians and Babylonians celebrated the renewal of the life-sustaining spring rains in the month of Nisan—although some cities of Mesopotamia retained an ancient custom of celebrating a second similar festival when the rains returned in the month of Tishri (autumn). Sacrifices of grain and other foods were dedicated to the gods Dumuzi (or Tammuz) or Marduk, major fertility deities, at a ziggurat (tower temple), after which the people participated in feasting, dancing, and other appropriate ritualistic activities.

In the 20th century, the view that New Year's Day is a time significant in the victory of order over disorder has been celebrated, for example, in areas influenced by Chinese religions. In order to frighten the *kuei* (evil or unpredictable spirits), which are believed to be dispersed by light and noise, participants in the New Year's festival light torches, lanterns, bonfires, and candles and explode firecrackers. In 1953, when the first day of the lunar New Year coincided with a solar eclipse, the government of the People's Republic of China (which has been anti-religious in its propaganda and official activities) expressed an anxiety that the repressed "religious popular superstitions" might encourage some form of anti-government activity. According to the views of Confucius (6th–5th centuries BC) and Mencius (4th–3rd centuries BC), two of China's great religious teachers, whose social and ethical influences have extended into the 20th century, a solar eclipse during the New Year's festival is a sign of a coming disaster and of a lack of favour by Shang Ti, the Heavenly Lord, who sends omens to indicate his disapproval of man's evil activities.

Man's view of himself in cultures that see time as linear

In religions and cultures that conceive of time as linear, progressing from a beginning toward an end time, when the whole cosmos will be renewed or changed, man understands his status (*i.e.*, his origin, identity, and destiny) in relationship to particular events in history that have a significance similar to those expressed in the myths of people who view time as cyclical. The Jew understands his status as a member of the "people of God," who were "chosen" during the Exodus of the Hebrews from Egypt in the 13th century BC to be witnesses to the liberating love of Yahweh (their God). His being one of the chosen "people of God" is celebrated especially during the Passover festival—in which the Exodus is ritually re-enacted and commemorated—in the month of Nisan (spring). Similarly, the Christian understands his status as a member of the "new people of God." He believes that he has been chosen by Christ, who was crucified and resurrected by God in the 1st century AD, to work for the Kingdom of God that was inaugurated in the first advent of Christ and will be consummated at the Parousia, the Second Coming of Christ as king and judge. The festival of the Resurrection, or Easter, is ritually re-enacted every year in order that the believer may participate in the present and future kingdom of peace. The eucharistic feast (the Holy Communion), though celebrated at many and various times during the year, originated in the event (namely, the Lord's Supper on Holy Thursday preceding Christ's Passion) that has been interpreted as a commemoration of the crucifixion and Resurrection. Just as the New Year's festivals of the religions that interpreted sacred time as cyclical incorporated both remorse and joy in their celebrations, so also the feasts of the Passover and the Resurrection include sorrow for the sins of the individual and of mankind and joy and hope for the salvation of man and the world (see also CALENDAR; JUDAISM; CHRISTIANITY).

Times of seasonal changes. *The significance of seasonal renewal in prehistoric times.* Before the development of agriculture, with its associations with solar and lunar calendars, ritual feasts were probably celebrated by hunters and gatherers of tubers and fruits. Paleolithic (Old Stone Age) peoples from about 30,000–10,000 BC and those living in what are called "Stone Age" cultures in the 20th century, such as the Aborigines in Australia and New Guinea, have celebrated various rites in which feasts have assumed positions of significance. Seasonal variations—important in the maintenance of the food supply—were associated with the migrations and fertility of animals and the growth and decay of tubers and fruits upon which the clan or tribe depended for its very existence. Thus, out of an acknowledgment of seasonal change, rituals—often including ceremonial feasts—most likely developed in relationship to beliefs that the continuance of the food supply depended on the sacred or holy powers that controlled various aspects and facets of nature: *e.g.*, animals, vegetation, the change in climatic conditions, weather phenomena, mountains, and rivers.

Access to the sacred or holy powers was obtained and maintained by certain religious personages (*e.g.*, shamans, or persons having healing and psychic transformation powers, priests, clan or tribal leaders, and other persons having special learned or inherited powers). Though interpretations by scholars vary and the evidence is still subject to much speculating, Paleolithic cave paintings—such as that of the "sorcerer" (a bearded figure wearing a mask on the top of which were antlers of a deer) at Les Trois Frères in France—and rock paintings of the Aruntas of central Australia—such as totemic animals (symbolizing clan and animal relationships) or mythological nature heroes (*e.g.*, Katuru, the "lightning man")—may indicate that fertility of animals and vegetation has been a primary concern (though not the only concern) in the ritual control of the food supply. Rituals connected with controlling the food supply generally centre on a feast in which eating, drinking, dancing, and the chanting of efficacious formulas play important symbolic roles.

Probable reasons for feast and fast days

At some point in human history (about 8,000–6,000 BC in the ancient Near East), when calendrical seasons were associated with planting and harvesting, special days or periods most likely were set aside for fasting (because of a paucity in the food supply) or for feasting (because of an increase in the food supply). Thus some calendrical periods inspired feelings of discouragement and remorse (when the food supply was low) or feelings of encouragement or joy (when the food supply was sufficient to meet immediate and future needs). Certain days were set aside during these periods for special rituals (often including feasts) that celebrated seasonal renewal, later interpreted in terms of individual spiritual or social renewal. In Zoroastrianism and Parsiism, for example, the annual seasonal renewal festival of Nōrūz (New Year) in the spring, dedicated to Rapithwin (the time of the midday meal), is at the same time a solemn and joyful celebration of new life in nature and the anticipated resurrection of the body when the world will be restored to its original and intended goodness—after the defeat of Ahriman (the spirit of evil and chaos) and his demons.

The pharaoh in Egyptian feasts and festivals

The significance of seasonal renewal in ancient Egypt. Seasonal-renewal motifs in ancient Egypt were often incorporated into other aspects of sacred times—such as times of passage rites (*e.g.*, ascension of the pharaoh to the throne), of death rites (*e.g.*, the transformation of the dead person into a glorified person), and of commemorating certain historical events (*e.g.*, military victories in which the pharaoh preserved *maʿat*—*i.e.*, order, truth, and justice—which was active in the realms of nature and society).

In Egypt during the 5th millennium BC, astronomers in the Nile Delta region associated the annual inundation of the river—which covered wide areas with fertile soil—with celestial movements, especially that of the star Sirius (*i.e.*, Sothis) and the sun. From such observations the Egyptians developed a solar calendar of 365 days, with 12 months of 30 days each and five festival days at the end of the year. Though priests assumed important functions at the festivals centred about the fertility of the soil irrigated by the Nile and the life-giving warmth of the sun, the pharaoh, the sacred king, embodied the continuity between the realm of the sacred (*i.e.*, the transcendent sphere) and the realm of the profane (*i.e.*, the sphere of time, space, and cause and effect). The pharaoh was believed to be the son of the sun god Horus of the Horizon (Harakhte), symbolized by the falcon; the sun god was also known as Re, among other names. The eastern horizon was viewed as the meeting point of the underworld of the dead and the

world of the living. The sun god also was known as Atum, which means "to be at the end," or the west. Osiris, the god of the afterlife (the world of the dead) was believed to be embodied in the recently deceased pharaoh, who passed on his sacred powers and position to the new pharaoh, his son. At the *śd* festival, the new pharaoh, as the son of Horus and of Re, as well as of Osiris, was invested with both kingly and priestly powers. At his coronation festival the pharaoh was believed to gain the power to restore *ma'at* after the death of the previous pharaoh, and also to restore economic prosperity.

During the royal festivals—*i.e.,* ascension to the throne, the coronation, and the *śd* festival—feasting presumably occurred. Festivals associated with seasonal renewal, however, involved sacrifices, eating, drinking, and sometimes dramatic or carnival-like events. Some scholars hold that the Egyptian terms for festival, however, contain concepts that became extremely significant in later Hellenistic (Greco-Roman) religions—*e.g.,* the mystery, or salvatory, religions, such as those of Mithra, Isis, and the Eleusinian mysteries—and Semitic-based religions—*e.g.,* Judaism, Christianity, and Islām. According to this view Egyptian terms for festival, such as *ḥb, ḥ', and pr.t,* all contain concepts of resurrection and epiphany (*i.e.,* the manifestation of a god). In Eastern Orthodox Christianity, for example, the festival of the Epiphany (January 6) celebrates Christ's manifestation to the Magi of the East (presumably followers of Zoroaster, a 6th-century BC Iranian prophet) and his Baptism in the Jordan River. The usual Greek designation for Epiphany is "the day of the light" (*hē hēmera tou phōtou*), in reference to the words in the Bible, in John 1:4, that Jesus is the "light of men." Under the influence of the Christian Catechetical school at Alexandria (led by Clement and Origen in the 2nd and 3rd centuries AD), the earlier religious speculations of the Egyptians concerning their festivals were enhanced by further mystical and spiritual interpretations that affected Christian worship, piety, doctrine, and iconography, especially in Eastern Christianity.

Egyptian terms for and concepts of festivals in later religions

The Egyptians celebrated many festivals that were connected with seasonal renewal, some of which became elaborated into sacred times of cosmic significance. Among their more popular festivals were those dedicated to Osiris, Amon-Re (the sun god), Horus, and Hathor (the sky goddess, represented by a cow).

Of special interest is the festival dedicated to Min, celebrated during the harvest month of Shemou (April). A statue of Min, represented as an ithyphallic god of fertility in iconography, was placed on an inclined pedestal, which was the symbol of *ma'at.* This pedestal represented the primordial mountain, a symbol of resurrection, renewal, and rebirth. During the processional honoring Min, hymns were sung and ritual dances and perhaps other types of dances were performed. The pharaoh and his queen entered the shrine and presumably enacted a sacred marriage rite. After the pharaoh's enthronement at the harvest Festival of Min, four arrows were shot toward the north, east, south, and west; and birds also were released in the directions of the four cardinal points of the compass. The releasing of the birds and arrows announced the harmonious union of man—both as an individual and as a corporate being—with the divine powers of nature inherent in the pharaoh as "Horus son of Min and Osiris." Though the pharaoh was symbolically significant in the feasts and festivals of ancient Egypt, the priests of the various cults officiated in the rituals and sacrifices to the many gods and announced the proper times for the differing forms of celebrations (see also SACRED OFFICES AND ORDERS).

The Festival of Min

The significance of seasonal renewal in ancient Mesopotamia. In ancient Mesopotamia, in Babylon, where the king was viewed not as the son of a god but as a god's agent, or representative, on earth, the New Year's festival (Akitu), in the spring month of Nisan, contained not only seasonal renewal motifs but also themes centring on the renewal of man and his community. The *Enuma elish,* the epic of creation, was read at the festival in order to remind the participants that cosmos (order) arose out of chaos by means of a struggle between Marduk, the god of heaven, and Tiamat, the goddess of the deep and the powers of chaos. The New Year's festival was sometimes celebrated over a period of 10 to 12 days in Babylon. On the fifth day, a sheep was beheaded; the body of the sheep was thrown into the river, and the head was taken into the wilderness. This ritual act, in which an exorcist (*mashmashu*)—one who casts out demonic powers—participated, symbolized the ridding of the community of the powers of chaos. (It was similar to the scapegoat ritual of the ancient Hebrews, in which the sins of the community were ceremonially transferred to a goat, which was later led to a wilderness area to wander about far from the community.)

The New Year's festival

Before sunrise of the third day following the scapegoat ceremony, the Babylonian king, as the representative of a sinful people as well as the agent of the god, had to submit to ritual acts of humiliation: his symbols of power were removed, and the priest (*urigallu*) hit him in the face and enjoined him to pray for the forgiveness of his sins and the sins of his people. After a profession of innocence, the priest absolved the king, restored his regal insignia, and performed ceremonies with the king to ensure the continuous support of the powers of order in nature. During the three days between the sacrifice of the sheep and the reinvestiture of the king, the populace of the city engaged in chaotic activities, perhaps of a carnival-like nature, to symbolize the presence of chaos in nature and society during this period of the apparent absence of the king and the god. When the king reappeared to his people, with his royal symbols of office and in the presence of the statue of Marduk, a procession of statues of the various gods together with their adoring devotees then took place, leading to a sanctuary (*bītakītu*) outside the city. On the 10th day, a banquet involving the king, priests, temple functionaries, and the gods was held to celebrate the renewal of nature, man, and society.

The significance of seasonal renewal in areas of other religions. Among the pre-Columbian Maya, the first month (*uinal*), Pop, of the New Year—which would be July in the presently used calendar—became a time for several renewal ceremonies. Old pottery and fibre mats were destroyed, and new clothes were put on. The temple was renovated to meet the needs of the god that was especially venerated during a particular year (the annual god changed from year to year). New wooden and clay idols were made, and the portals and implements of the temple were reconsecrated with blue paint, the sacred colour. The god of the year entered the sacred precincts according to the cardinal point of the compass that he represented (and thus there were only four New Year's gods). The purpose of the processional rite was to ward off the forces of evil that might prevail against the people of the area. Dances by old women and sacrifices of live dogs (by throwing them down from the temple pyramid) were some of the activities that occurred during the Maya New Year's festival.

In Japan, among those engaged in agriculture, the *taasobi* ("rice-field ritual") festival is celebrated at the beginning of the year to ensure a plentiful harvest. Dances, songs sung with a *sasara* (musical instrument), sowing of seeds, and feasting play important roles in securing the aid of the *kami* (gods or spirits). Divination by means of archery, in which the angle of the arrow on the target is significant, has been used in shrines to help determine the methods that should be used in securing a good crop. In Hinduism, the Makara-Saṃkrānti, a New Year's festival in the month of Māgha (January–February), is celebrated with a fair that continues for a month's duration, with much rejoicing. The Śrī Pañcamī, a festival (*utsava*) of seasonal renewal on the fifth day of Māgha, symbolizes the ripening of crops. Feasts and festivals centring on seasonal renewal can be found among all peoples of the world, both past and present. Rogation festivities (Days of Asking), originally held by the ancient Romans to counteract the effectiveness of the deity (Robigus) of red mildew on wheat, were reinterpreted by early medieval Christians of the West from the 5th century on as litanies for the blessing of the seed. Rogation Day, the fifth Sunday after Easter, is still practiced in the 20th century in rural Roman Catholic, Anglican, and Lutheran churches.

Other sacred times. *Crucial stages of life.* Birth, puberty, marriage, and death have been times of sacred significance for peoples of all races from time immemorial. They signify changes in the status of a person's being in terms of a person's relationship with fellow members of his or her society and the realm of the sacred or holy that informs the person of the practical and symbolic ramifications of the new status. These times of change, therefore, have become occasions for feasts and festivals. Some are very elaborate and of long duration; others, especially under the influence of modern secularization, have been abruptly shortened or eliminated.

Rites and feasts connected with birth

Birth, a most sacred time in the religions of the world, is celebrated by rites and festivities that appear to be incongruous or inconsistent in many religions. Mothers of newborn children are considered both as participants of the sacred by having brought forth a new being into the world and as persons who are ritually unclean (*e.g.*, among the Israelites and Zoroastrians), probably because of the presence of blood at birth, the loss of which may symbolize the loss of some of the life-sustaining force. Among Brazilian Indians, however, both the father and the mother participate in a ceremony of seclusion for five days (eating only certain foods) in order to protect the sacredness and health of the new mother and child. Seclusion, thus, need not be interpreted negatively. Among the Kikuyu of eastern Africa, seclusion is a symbol of death and resurrection. The mother and child symbolically die and rise again during and after a ceremony of seclusion, after which a feast is held in which a goat is sacrificed and prayers are said. The whole community rejoices that a new child has become a part of the family of man.

The Christian celebration of birth culminates in the sacrament of Baptism, a symbol of the death of the old person and the rebirth of the new person in Christ. As such, it is a rite of purification, using water and the words of institution by Christ. After the sacrament has been solemnized, Christians in many areas have engaged in much feasting to emphasize the joy inherent in the "new birth."

Among the ancient pre-Christian Norsemen, baptism by means of water was believed to impart divine and eternal life to men and even to preserve men from death—so that they "will not perish in war" nor "fall before any sword." Thus, when St. Boniface baptized members of Germanic tribes in the 8th century, he was ordered by Pope Gregory III to do so only according to the formula "in the name of the Father, and of the Son, and of the Holy Spirit." Because whole tribes became Christian en masse during this period, the feasts celebrating the incorporation of the tribe into the church often lasted for several days and included folk customs of which the church did not especially approve, such as those connected with merrymaking (*e.g.*, the drinking of mead).

Rites and feasts connected with puberty

Puberty, the transition into adulthood, has been celebrated since ancient times by various rituals and festivals. In the secular sphere, it is celebrated in democratic countries by the granting of the right to vote to persons upon the attainment of a certain age. In ancient Greece, young men of the ages of 16 or 17 were admitted as full members of the city-state; but before they were granted voting privileges, they had to swear allegiance to the religion of the city; this made them religious citizens and subsequently adults. After he had attained adulthood, a young Greek could participate in military service and could marry. In the United States in the early 1970s, citizens having attained the age of 18 were granted the right to vote; but the ceremony commemorating this right has been a secularized de-emphasis of this important rite of passage: the mere signing of one's name on a registration certificate.

Puberty rites are celebrated in various ways according to the prevailing religious and social customs. Among the Masai of eastern Africa, youths pass from childhood to adulthood by the rite of circumcision. After various preliminary activities, the boys (12 to 16 years of age) are circumcised and the blood released from the operation is later placed on their heads. After four days of seclusion and a period during which they are dressed in female attire, their heads are shaved and they attain the status of adults and thus can become warriors. Girls attain adulthood by means of similar practices: the cutting or piercing of sexual organs. Among the Kamba of eastern Africa, who perform similar puberty rites of passage, those initiated into adulthood are given presents, and offerings are made to the ancestors. A significant aspect of the festival celebrating the rite of passing from childhood to adulthood is the return from seclusion; this return to their communities symbolizes a type of resurrection and renewal as new persons—adults.

Among the churches of the 16th-century Reformation, the rite of confirmation in the Anglican and Lutheran churches has been a type of puberty rite. The child, who had been a baptized member of the church, became, in effect, an adult, assuming personal responsibility and the privilege of participating in the Eucharist. In the early 1970s, however, the instructional aspect of confirmation—important in almost all pre-puberty practices—has been diminished, especially in some Lutheran churches in the United States, thus de-emphasizing the importance of confirmation as a rite of passage. As the church has become increasingly influenced by secularization processes in the 20th century, the customary feasting to celebrate the rite of confirmation has decreased in practice.

Rites and feasts connected with marriage

Marriage, the rite of passage from the single to the united state, has been celebrated with many forms of feasts and festivals. Connected with the *hieros gamos* ("sacred marriage") of the Mesopotamian Akitu (New Year's festival), and of the Israelite Sukkot (Feast of Tabernacles)—during the month of Tishri (the first month of the year)—which had both sexual and covenantal overtones, the rite of marriage developed into a legal and religious act in Judaism and into a sacrament in Roman Catholic and Eastern Christianity. In most religions the married state is considered superior to the single, though tensions between these two states of existence exist in most religions. Monks and nuns who vow to live in a celibate state often celebrate a symbolic marriage to the founder of their religion (*e.g.*, to Christ) or to a religious institution (*e.g.*, the church). In the Talmud, a compendium of Jewish law, lore, and commentary, the statement is made that "He who does not marry is like a murderer and he mutilates (violates) the image of God." In the Avesta, the sacred book of Zoroastrianism, a similar statement is made: "The man who is married stands above him who is not married." Thus, the wedding has become the most significant domestic festival in both the secular and religious realms, in spite of the ascetic tendencies that exist in certain sectors of Christianity, Buddhism, and other religions. The wedding ceremony has often been accompanied by feasting and gift-giving to express the concern of the community for a successful participation within the community and an extension of the community through the procreation of children. Among African religions, marriage as a rite of passage is incomplete if procreation is avoided or not accomplished. After a wedding among the Batoro of Uganda in Africa, dancing and feasting last until the following morning. Later on, gifts are given to the bride's family in order to show gratitude, to compensate for her absence, and to legalize the marriage agreement.

Rites and feasts connected with death

The final rite of passage, death, has brought about numerous festival customs, all the way from the ritual sacrifice of the widow in Hinduism (until the 19th century) to the commercialization of death rites in Western societies. Just as the early Hebrews believed that life passes on to death when the breath (*ruaḥ*) leaves the body, so also do Eskimos in the 20th century believe that death occurs when breath (soul) leaves the body and that death may be a moment when one is translated into another form of life. Among the ancient Greeks, Thanatos (death) is the twin brother of Hypnos (sleep), and from this conceptional relationship may come the view that death is merely a sleeping state in the passage from this life to an afterlife. Festivities surrounding rites include the customs of playing mournful (and, sometimes, joyful) music, speaking eulogies, performing sacramental acts (*e.g.*, extreme unction in the Roman Catholic Church), performing elaborate or simple embalming practices (*e.g.*, the lengthy procedural techniques of the ancient Egyptians and the rapid techniques of modern morticians), utilizing appropriate and expected

bodily gestures and vocal expressions, and feasts of varied elaborateness, depending on the economic or social circumstances of the deceased or his next of kin. Flowers often play important roles in the festivities connected with death rites. In the 20th century, a change from mourning to joyful expectation has occurred in the funeral rites of some Christian churches. Among some African tribes, such as the Ndebele of Zimbabwe, funeral processions, sacrifices, ceremonial washings, and protective medicine are included in the festivities that symbolically celebrate man's conquest over death (see also *Rites of passage* and *Death rites and customs,* above).

Times of commemoration and remembrance. Festivals of commemoration are among the most important of the sacred times. Some festivals commemorate important events in mythology or the birth, inauguration, or victory of a founder of a religion, a god, or a hero. In Hinduism, for example, the Vaikuṇṭha-ekādaśī festival in December–January commemorates the victory of the goddess Ekādaśī Devī in her killing of a demon; and the Gaṇeśacturthī commemorates the birthday of Gaṇeśa, the elephant-headed god of fortune. Another major Hindu festival, Navarātri, commemorates the victory of the goddess Durgā over the buffalo-headed demon Mahiṣa; and Rāma-navamī commemorates the birth of Rāma, the hero of the *Rāmāyana,* one of India's great epics. In Chinese Buddhism, the birthdays of Kuan-yin (or Avalokiteśvara), Amitābha, and Śākyamuni (the first two being *bodhisattvas,* or buddhas-to-be, and the last being the Buddha himself) were celebrated before the 1950s with much ceremony. The nativity of Christ (or Christmas) is the most widely celebrated "birthday" of a divine being, though in the 20th century Christmas has been subjected to a wide variety of secular influences.

TYPES AND KINDS OF FEASTS AND FESTIVALS

National and local festivals. Feasts and festivals vary greatly in type. Though most are religious in background and character, other types have flourished in both ancient and modern civilizations. Included among such types are social and cultural festivals: *e.g.,* New Year's Day in the 20th century, sword-dance festivals in Scotland, the Olympic festivals in ancient Greece and the modern world, the Great Dionysia of ancient Greece during which dramatic contests took place, and May Day celebrations. National festivals in the United States include Thanksgiving Day (in November), which commemorates colonial celebrations following successful harvests; Independence Day (July 4), which commemorates the Declaration of Independence of the American colonies from the British crown; St. Patrick's Day (March 17), celebrated mainly in Chicago and New York City as a secular–religious feast; Mother's Day (in May); Memorial Day (in May), commemorating those who have died, especially in war; and Flag Day (June 14). National or local festivals in other countries include: Bastille Day (July 14), commemorating the beginning of the French Revolution in 1789; Dominion Day (July 1) in Canada; and independence days in many countries. Birthdays of national founders or heroes are also types of commemorative festivals. In some Protestant countries, Reformation Day has assumed the position of a holiday either nationally or locally. In Israel, Holocaust Day commemorates the systematic destruction of European Jews by Nazi Germany in the 1930s and '40s.

Secular modernist festivals. Secular modernist festivals are often mixed with previous religious festivals. May Day, once mainly a springtime fertility festival that can be traced back to the Magna Mater (Great Mother) festivals of Hellenistic (Greco-Roman) times, has become a festival of the labouring class in Socialist countries. Football games in the United States have all the external trappings of religious festivals. A person from a preliterate culture would see a large congregation witnessing a ritual combat, conducted according to precise ritualistic rules. The participants are dressed in appropriate identifiable costumes as they engage in their ritual combat—one side representing evil and the other good, depending upon the viewpoint of the audience. Leading the congregation are priestesses (cheerleaders) dressed in appropriate garb, participating in ritualistic dances, and chanting supposedly efficacious formulas. Operating on the principle of sympathetic magic, the priestesses attempt to transfer the crowd's enthusiasm to the appropriate combatants. In Western countries, according to some critics, lay participation in congregational worship has for a long time been little more than a spectator sport, and this may well have contributed to the festival character of weekend sports activities.

May Day

Carnivals and saturnalias. Some feasts and festivals provide psychological, cathartic, and therapeutic outlets for persons during periods of seasonal depression. The Holī festival of Hinduism during February–March was once a fertility festival. Of early origin, the Holī festival incorporates a pole, similar to the Maypole of Europe, that may be a phallic symbol. Bonfires are lit; street dancing, accompanied by loud drums and horns, obscene gestures, and vocalized obscenities, is allowed; and various objects, such as coloured powders, are thrown at people.

One of the best-known festivals of ancient Rome was the Saturnalia, a winter festival celebrated on December 17–24. Because it was a time of wild merrymaking and domestic celebrations, businesses, schools, and law courts were closed so that the public could feast, dance, gamble, and generally enjoy itself to the fullest. December 25—the birthday of Mithra, the Iranian god of light, and a day devoted to the invincible sun, as well as the day after the Saturnalia—was adopted by the church as Christmas, the nativity of Christ, to counteract the effects of these festivals.

Carnival-like celebrations were held in England on Shrove Tuesday, the day before the Lenten fast began, until the 19th century. Originating as a seasonal renewal festival incorporating fertility motifs, the celebrations included ball games that often turned into riots between opposing villages. Feasts of pancakes and much drinking followed the contests. This tradition of merrymaking continues, for example, in the United States in the Mardi Gras festival on Shrove Tuesday in Louisiana.

CONCLUSION

Feasts and festivals, whether religious or secular, national or local, serve to meet specific social and psychological needs and provide cohesiveness to social institutions: *e.g.,* church, state, and esoteric or socially nonaccepted groups. The cohesiveness engendered in the feasts and festivals of minority groups (*e.g.,* Christians in the early Roman Empire) often provides these groups with the strength to influence the institutions of the society and the culture of the majority. When a particular religion triumphs over other religions, it often incorporates elements from the feasts and festivals of the previously predominant religions into its own religious calendar. This has been an important practice of all the world religions in their attempts to bring about social solidarity, order, and tranquility. Similarly, individuals can gain a sense of psychological cohesiveness through participation in feasts and festivals.

During periods of crisis in society, feasts and festivals may lose some of the impact of their interpretive and cohesive functions. The sacraments of the medieval Western Church lost some of their earlier interpretive values in the 16th century during the Reformation, and the month of fasting before the Feast of Bēma ("judge's seat")—a festival commemorating the death of Mani, a 3rd-century-AD Iranian prophet who founded the syncretistic Manichaean religion—probably became the prototype of the Muslim fast month of Ramaḍān after Islāmic invasions of the 7th century AD. So also can persons living in the 20th century expect reinterpretations of the feasts and festivals to which they have become accustomed. Reinterpretations of feasts and festivals may thus provide impulses for institutional changes, which generally occur in times of crisis and transition. (L.F.)

BIBLIOGRAPHY

Sacred or holy: The socio-anthropological analyses written near the beginning of the 20th century that are still useful for their interpretations of the sacred in preliterate societies include HENRI HUBERT and MARCEL MAUSS, *Essai sur la nature et le fonction du sacrifice* (1899; Eng. trans., *Sacrifice: Its Nature and Function,* 1964); and ÉMILE DURKHEIM, *Les Formes élé-*

mentaires de la vie religieuse, le système totémique en Australia (1912; Eng. trans., *The Elementary Forms of the Religious Life,* 1965). More recently, in the same vein, are E.O. JAMES, *Sacrifice and Sacrament* (1962), a comparative analysis of sacred ritual from many different religious traditions; and ROGER CAILLOIS, *L'Homme et le sacré* (1939; Eng. trans., *Man and the Sacred,* 1960), a general reflective interpretation of various social expressions of the sacred. The following combine philosophical and theological concerns: RUDOLF OTTO, *Das Heilige* (1917; Eng. trans., *The Idea of the Holy,* 1923), an appeal to an a priori preconceptual knowledge of the holy; MAX F. SCHELER, *Vom Ewigen im Menschen,* 2nd ed. (1923; Eng. trans., *On the Eternal in Man,* 1960), an intuitive philosopher's argument for the eternal reality of the sacred prior to man's awareness or social expression of it; NATHAN SODERBLOM, "Holiness," *Encyclopedia of Religion and Ethics,* 6:731–41 (1928, reprinted 1955), which stressed the quality of holiness in all religion four years prior to Otto's more famous statement, and his *Living God: Basal Forms of Personal Religion* (1933), a comparative study of religion organized according to various ways through which man encounters God; and JOACHIM WACH, *The Comparative Study of Religions,* ed. by J.M. KITAGAWA (1958), a systematic analysis of the modes (thought, action, fellowship) used to express the religious experience. Two Dutch phenomenologists of religion who have made notable contributions to the interpretation of forms that express man's relation to the sacred are GERARDUS VAN DER LEEUW, whose *Phänomenologie der Religion* (1933; Eng. trans., *Religion in Essence and Manifestation,* 1963) organizes a wide spectrum of data into three foci: the object of religion, the subject of religion, and their reciprocal relation; and W. BREDE KRISTENSEN, who wrote *The Meaning of Religion* (1960), a series of lectures given during the 1930s on the sacredness of man's cosmological, anthropological, and cultic awareness as expressed in the preliterate cultures and those of the ancient Mediterranean area. An extensive analysis of the forms and modes in which the sacred is recognized is found in the writings of MIRCEA ELIADE, for whom the apprehension of the sacred is a unique kind of experience in which the creative power(s) of life appear(s) in particular symbols, myths, and rites. Four of his works that deal with the nature and meaning of the sacred in different types of expression are: *Le Mythe de l'éternel retour* (1949; Eng. trans., *The Myth of the Eternal Return,* 1954, reissued 1989); *Myth and Reality* (1963); *Traité d'histoire des religions* (Eng. trans., *Patterns in Comparative Religion,* 1958); and *The Sacred and the Profane: The Nature of Religion* (1959).

Worship: MIRCEA ELIADE, *Traité d'histoire des religions* (1949; Eng. trans., *Patterns in Comparative Religion,* 1958), is a standard work with much information on worship. JAMES G. FRAZER, *The Worship of Nature* (1926), is a classic work but is now out of date. WALTER HARRELSON, *From Fertility Cult to Worship* (1969); and F.H. HILLIARD, *How Men Worship* (1965), are popular, brief treatments of worship in the major religions. JOHN S. MBITI, *Concepts of God in Africa* (1970), has valuable source materials but is difficult to use because of the variety of materials placed together. GEOFFREY PARRINDER, *Worship in the World's Religions* (1961), is a popular work, accurate and helpful, but brief. H.H. ROWLEY, *Worship in Ancient Israel* (1967), is a standard and comprehensive work. EVELYN UNDERHILL, *Worship* (1936), is a standard work, still valuable, but now out of date. GEOFFREY WAINWRIGHT, *Doxology: The Praise of God in Worship, Doctrine and Life* (1980), argues that all Christian themes come to a focus in worship.

Ritual: WILLIAM LESSA and EVON Z. VOGT, *Reader in Comparative Religion,* 3rd ed. (1971), is a good general anthology on classical and modern positions on religion, ritual, and myth (mainly concerned with nonliterate cultures), with an excellent bibliography. *Gods and Rituals,* ed. by JOHN MIDDLETON (1967), contains a good collection of essays on ritual practices in nonliterate cultures, also with a fine bibliography. Among the classic texts dealing with the origin of ritual and religion, there are three authors who have enduring influence: W. ROBERTSON SMITH, *Lectures on the Religion of the Semites* (1889); ÉMILE DURKHEIM, *Les Formes élémentaires de la vie religieuse* (1912; Eng. trans., *The Elementary Forms of the Religious Life,* 1965); and SIGMUND FREUD, *Totem und Tabu* (1913; Eng. trans., *Totem and Taboo,* 1918). Among the classic positions on a functional approach to ritual are those of BRONISŁAW MALINOWSKI, *Coral Gardens and Their Magic,* 2 vol. (1935); and A.R. RADCLIFFE-BROWN, *The Andaman Islanders* (1922). More recent examples of the functional approach are the anthropological texts of E.E. EVANS-PRITCHARD, *Nuer Religion* (1956); and EDMUND LEACH, *Political Systems of Highland Burma* (1954). MELFORD E. SPIRO, *Burmese Supernaturalism* (1967), is one of the best critical texts using data from Burmese Buddhism as support for a revised approach. VICTOR W. TURNER, *The Forest of Symbols* (1967), represents a novel analysis of dominant symbols in belief and ritual. Among valuable approaches by theologians and historians of religion are RUDOLF OTTO, *Das Heilige* (1917; Eng. trans., *The Idea of the Holy,* 1923); and JOACHIM WACH, *The Comparative Study of Religions* (1958). JANE E. HARRISON, *Themis,* 2nd ed. rev. (1927); and S.H. HOOKE (ed.), *Myth, Ritual and Kingship* (1958), are good examples of the myth-ritual school. An excellent critique of this school may be found in JOSEPH E. FONTENROSE, *The Ritual Theory of Myth* (1966). HENRI HUBERT and MARCEL MAUSS, *Essai sur la nature et le fonction du sacrifice* (1899; Eng. trans., *Sacrifice: Its Nature and Function,* 1964), remains a standard analysis of sacrifice as ritual. ARNOLD VAN GENNEP, *Les Rites de passage* (Eng. trans., *The Rites of Passage,* 1960), although written in 1909, continues to be an important work on ritual as a marker of passage. MIRCEA ELIADE, *Birth and Rebirth* (1958), is an excellent historical study of ritual as initiation, with a good bibliography. See also BRUCE LINCOLN, *Emerging from the Chrysalis: Studies in Rituals of Women's Initiation* (1981).

Prayer: The classical work is still that of F. HEILER, *Das Gebet,* 5th ed. (1923), which includes a bibliography and defends the theory of religious syncretism. R. BOCCASSINO (ed.), *La preghiera,* 3 vol. (1967), is a historical and psychological study of prayer and includes a useful bibliography. G. VAN DER LEEUW, *Phänomenologie der Religion* (1933; Eng. trans., *Religion in Essence and Manifestation,* 1963), is an excellent overview and general introduction of the psychology of prayer. Also of a psychological bent are WILLIAM JAMES, *The Varieties of Religious Experience* (1902); and H.U. VON BALTHASAR, *Das betrachtende Gebet* (1955; Eng. trans., *Prayer,* 1961). For prayer in various religions, in addition to *La preghiera,* one should refer to M.P. NILSSON, *Geschichte der griechischen Religion,* 2 vol. (1941–50; Eng. trans., *A History of Greek Religion,* 2nd ed., 1963); A. FALKENSTEIN and W. VON SODEN, *Summerische und akkadische Hymnen und Gebete* (1953); G.C. LOUNSBERY, *La Méditation bouddhique* (1935; Eng. trans., *Buddhist Meditation in the Southern School,* 1950); E. CONZE, *Buddhist Meditation* (1956); and R.C. ZAEHNER, *The Teachings of the Magi* (1956) and *The Dawn and Twilight of Zoroastrianism* (1961). For a bibliography on biblical prayer, see "Prière," *Dictionnaire de la Bible,* suppl. 8, pp. 604–606 (1968). I. ELBOGEN, *Der jüdische Gottesdienst in seiner geschichtlichen Entwicklung,* 4th ed. (1962), is a classic on Jewish prayer. This can be supplemented by A.Z. IDELSOHN, *Jewish Liturgy and Its Development* (1967). On the origins of Christian prayer, see A. HAMMAN, *La Prière,* 2 vol. (1959–63). KENNETH LEECH, *True Prayer* (1980), is an Anglo-Catholic introduction to Christian spirituality.

Creed and confession: A comprehensive treatment of creeds in all religions is "Creeds and Articles," *Encyclopaedia of Religion and Ethics,* 4:231–248 (1912, reprinted 1955). For anthropological, sociological, and phenomenological considerations, see the *International Encyclopedia of the Social Sciences,* 13:398–414 (1968); G. VAN DER LEEUW, *Phänomenologie der Religion* (1933; Eng. trans., *Religion in Essence and Manifestation,* 1963); and J. WACH, *Sociology of Religion* (1944). Works devoted to creedal and confessional formulations are rare for most religions, but see S. SCHECHTER, "The Dogmas of Judaism," *Studies in Judaism,* pp. 147–181 (1896); and A.J. WENSINCK, *The Muslim Creed: Its Genesis and Historical Development* (1932). For Christianity, the fullest collection of texts remains P. SCHAFF, *The Creeds of Christendom,* 3 vol., 6th ed. (1919); for Roman Catholicism, H.J.D. DENZINGER and A. SCHONMETZER, *Enchiridion Symbolorum* (1963); and W.M. ABBOTT (ed.), *The Documents of Vatican II* (1966); for Protestantism, T.G. TAPPERT (ed. and trans.), *The Book of Concord* (1959); and A.C. COCHRANE (ed.), *Reformed Confessions of the 16th Century* (1966). On the Ecumenical movement, see L. VISCHER (ed.), *A Documentary History of the Faith and Order Movement 1927–1963* (1963). Brief but representative collections are B.A. GERRISH, *The Faith of Christendom: A Source Book of Creeds and Confessions* (1963); and J.H. LEITH (ed.), *Creeds of the Churches* (1963). Secondary works on early creeds include O. CULLMANN, *Die ersten christlichen Glaubensbekenntnisse* (1943; Eng. trans., *The Earliest Christian Confessions,* 1949); J.N.D. KELLY, *Early Christian Doctrines,* 2nd ed. (1960); A.E. BURN, *The Athanasian Creed,* 3rd impression (1930); D.L. HOLLAND, "The Earliest Text of the Old Roman Symbol," *Church History,* 34:262–281 (1965), and "The Creeds of Nicaea and Constantinople Reexamined," *Church History,* 38:248-261 (1969). On later confessions, a full treatment with good bibliographies is E. MOLLAND, *Christendom* (1959). This is usually supplemented by W.A. CURTIS, *A History of Creeds and Confessions of Faith in Christendom and Beyond* (1911); and C.A. BRIGGS, *Theological Symbolics* (1914).

Sacrament: The standard pioneer work on the wider pre-Christian occurrence and interpretations is W.R. SMITH, *Lectures on the Religion of the Semites,* 3rd ed. (1927), stressing sacramental communion with the deity. A. GARDNER, *History of Sacrament in Relation to Thought and Progress* (1921), applied the sacramental principle to various aspects of life and belief;

and R.R. MARETT, *Sacraments of Simple Folk* (1933), wrote on sacraments in primitive culture. In his *Dawn and Twilight of Zoroastrianism* (1961), R.C. ZAEHNER discussed the anticipation of the Christian eucharistic sacramental rite in the *Yasna* ceremony in the Avestan liturgy. The basic sacramental beliefs and cults throughout the ages are examined anthropologically in E.O. JAMES, *Sacrifice and Sacrament* (1962).

A good general study of the Christian doctrine of sacraments is P.T. FORSYTH, *Lectures on the Church and the Sacraments* (1917). O.C. QUICK, *The Christian Sacraments* (1927, reprinted continually to 1952), is one of the most comprehensive surveys. The positions of various religious bodies are presented in the following: A.J. TAIT, *Nature and Functions of the Sacraments* (1917), the evangelical viewpoint; B. LEEMING, *The Principles of Sacramental Theology,* new ed. (1960), the Roman Catholic viewpoint; D.M. BAILLIE, *The Theology of the Sacraments and Other Papers* (1957), the Protestant viewpoint; and A. SCHMEMANN, *Sacraments and Orthodoxy* (1965), the Orthodox viewpoint. J.H. SRAWLEY, *Liturgical Movement: Its Origin and Growth* (1954), examines the development of sacramental worship with special reference to lay participation. See also JOSEPH MARTOS, *Doors to the Sacred: A Historical Introduction to Sacraments in the Catholic Church* (1981).

Sacrifice: Classic theories of the origin and nature of sacrifice are found in the following: EDWARD B. TYLOR, *Primitive Culture,* 2 vol. (1871, reprinted 1958), a presentation of the gift theory of sacrifice; W. ROBERTSON SMITH, *Lectures on the Religion of the Semites,* 3rd ed. (1927), the clearest formulation of the author's theory of communion through a sacrificial meal; JAMES G. FRAZER, *The Golden Bough,* 3rd ed., 12 vol. (1907–15; abridged ed., *The New Golden Bough,* 1964), a famous and influential treatise on ancient religion that presents sacrifice as a means for rejuvenating a god; and HENRI HUBERT and MARCEL MAUSS, "Essai sur la nature et la fonction du sacrifice," *L'Année sociologique* (1899; Eng. trans., *Sacrifice: Its Nature and Function,* 1964), a sociological explanation of the sacrificial victim as a buffer between man and the god. More recent formulations include GERARDUS VAN DER LEEUW, *Phänomenologie der Religion* (1933; Eng. trans., *Religion in Essence and Manifestation,* 1963), an expansion of the notion of the sacrificial gift by a phenomenologist of religion; ADOLF E. JENSEN, *Mythos und Kult bei Naturvölkern,* rev. ed. (1960; Eng. trans., *Myth and Cult Among Primitive Peoples,* 1963), which correlates types of cultures and their sacrifice; RAYMOND FIRTH, "Offering and Sacrifice: Problems of Organization," in W.A. LESSA and E.Z. VOGT (eds.), *Reader in Comparative Religion,* 3rd ed., pp. 185–194 (1971), an economic interpretation of sacrifice; and E.O. JAMES, *Sacrifice and Sacrament* (1962), a good survey. FRANCES M. YOUNG, *Sacrifice and the Death of Christ* (1978), is an overview of the theology of sacrifice in the early Christian church.

Brief articles on several religions are found in "Sacrifice," *Encyclopaedia of Religion and Ethics,* 11:1–39 (1928, reprinted 1955). On Vedic religion, A.B. KEITH, *The Religion and Philosophy of the Veda and Upanishads,* 2 vol. (1925), is still a standard work; and LOUIS RENOU, *Religions of Ancient India* (1953), offers a brief survey. On Chinese sacrificial rites, see C.K. YANG, *Religion in Chinese Society* (1961); on ancient Egypt, J.H. BREASTED, *The Elder Development of Religion and Thought in Ancient Egypt* (1912); and on ancient Greek and Roman religions, R.K. YERKES, *Sacrifice in Greek and Roman Religions and Early Judaism* (1952), a clearly written, well-documented work; and M.P. NILSSON, *Geschichte der griechischen Religion,* 2 vol. (1941–50; Eng. trans., *A History of Greek Religion,* 2nd ed., 1963), a good handbook on Greek religion. On sacrificial rites in Judaism there is extensive literature, including "Sacrifice," *Encyclopedia Judaica,* 14:599–615 (1971), a good survey with a bibliography; ROLAND DE VAUX, *Les Sacrifices de l'Ancien Testament* (1964; Eng. trans., *Studies in Old Testament Sacrifice,* 1964); and YERKES (above). On ancient Scandinavian rites, see E.O.G. TURVILLE-PETRIE, *Myth and Religion of the North* (1964). WALTER KRICKEBERG *et al., Die Religionen des Alten Amerika* (1961; Eng. trans., *Pre-Columbian American Religions,* 1968), discusses the rites of the ancient civilizations of the American continents. On the religions of the peoples of Africa, JOHN S. MBITI, *Concepts of God in Africa* (1970), is an introduction with extensive bibliography. Important specific studies include MELVILLE HERSKOVITS, *Dahomey,* 2 vol. (1938); E.E. EVANS-PRITCHARD, *Nuer Religion* (1956); E.B. IDOWU, *Olódùmarè: God in Yoruba Belief* (1962); and GEOFFREY PARRINDER, *West African Religion,* 2nd ed. rev. (1961).

Rites of passage: ARNOLD VAN GENNEP, *Les Rites de passage* (1909; Eng. trans., *The Rites of Passage,* 1960), is a pioneering study and standard work on passage rites. D.M. SCHNEIDER and K. GOUGH (eds.), *Matrilineal Kinship* (1961), is also, in comparison, a discussion of patrilineal kinship. BRUNO BETTELHEIM, *Symbolic Wounds* (1954), is a Freudian-inspired work interpreting ritual acts of circumcision and other genital operations. E.D. CHAPPLE and C.S. COON, *Principles of Anthropology* (1942), has useful information on social interaction, social equilibrium and disruption, and the role of rites of passage in restoring equilibrium. J.G. FRAZER, *The Golden Bough,* 3rd ed., 12 vol. (1907–15), is a classic work that discusses rites of passage and many other features of religion. A.M. HOCART, *Social Origins* (1954), is an interesting interpretive work although somewhat dated. FRANK W. YOUNG, *Initiation Ceremonies* (1965), concerns rites of coming-of-age, interpreting their significance in relation to the social roles of males and females and the organization of social groups.

Death rites and customs: E. BENDANN, *Death Customs: An Analytical Study of Burial Rites* (1930, reprinted 1974), is a useful account of relevant ethnological material. P.C. ROSENBLATT, *Grief and Mourning in Cross-Cultural Perspective* (1976); and R. HUNTINGTON, *Celebrations of Death: The Anthropology of Mortuary Rituals* (1979), are two more-recent anthropological studies. Four works by S.G.F. BRANDON are helpful: *Man and His Destiny in the Great Religions* (1962), with extensive bibliographies and documentation; *The Judgment of the Dead* (1967), a comprehensive study of the subject; *Man and God in Art and Ritual* (1972), a profusely illustrated study that deals with mortuary rituals, conceptions of burial, and funerary iconography; and "The Personification of Death in Some Ancient Religions," *Bull. John Rylands Library,* 43:317–385 (1961). J. MARINGER, *Vorgeschichtliche Religion* (1956; Eng. trans., *The Gods of Prehistoric Man,* 1960), discusses Paleolithic and Neolithic burial practices. E.A.W. BUDGE, *The Mummy,* 2nd ed. (1894, reprinted 1974), is a handbook on Egyptian funerary archaeology. J. ZANDEE, *Death As an Enemy, According to Ancient Egyptian Conceptions* (1960, reissued 1977), also includes Coptic evidence. M. LAMM, *The Jewish Way in Death and Mourning,* rev. ed. (1972); and J. JEREMIAS, *Heiligengräber in Jesu Umwelt (Mt. 23, 29; Lk. 11, 47)* (1958), an account of Jewish mortuary beliefs, are valuable studies. Sources focusing on Greek and Roman civilizations include E. ROHDE, *Psyche: Seelenkult und Unsterblichkeitsglaube der Griechen,* 8th ed. (1921; Eng. trans., *Psyche: The Cult of Souls and Belief in Immortality Among the Greeks,* 1925, reprinted 1972); D.C. KURTZ, *Greek Burial Customs* (1971); F.V.A. CUMONT, *After Life in Roman Paganism* (1922); J.M.C. TOYNBEE, *Death and Burial in the Roman World* (1971); and A.K. FORTESCUE, *The Ceremonies of the Roman Rite Described,* 6th ed. rev. by J.B. O'CONNELL (1937). The following are also recommended: R. EKLUND, *Life Between Death and Resurrection According to Islam* (1941); J.D.C. PAVRY, *The Zoroastrian Doctrine of a Future Life: From Death to the Individual Judgment,* 2nd ed. (1929, reissued 1975); J.J. MODI, *The Religious Ceremonies and Customs of the Parsees* (1922, reprinted 1979; 2nd ed., 1937); J.A. DUBOIS, *Hindu Manners, Customs and Ceremonies,* 3rd ed. (1906, reissued 1968), invaluable for its descriptions; M. GRANET, *La Civilisation Chinoise* (1929; Eng. trans., *Chinese Civilization,* 1930, reprinted 1974); *La Religion des Chinois,* 2nd ed. (1951); P. ARIÈS, *Western Attitudes Toward Death: From the Middle Ages to the Present* (1974), as reflected in ceremonies, customs, literature, and art; W.K.L. CLARKE (ed.), *Liturgy and Worship: A Companion to the Prayer Books of the Anglican Communion* (1932); T.S.R. BOASE, *Death in the Middle Ages: Mortality, Judgment, Remembrance* (1972); and E. PANOFSKY, *Tomb Sculpture* (1964), an illustrated survey of funerary iconography from ancient Egypt to the Renaissance.

Purification rites and customs: Few works deal directly with purification rites. MARY DOUGLAS, *Purity and Danger* (1966), is a major work dealing with the problems of purity and impurity. HUTTON WEBSTER, *Taboo: A Sociological Study* (1942); and FRANZ STEINER, *Taboo* (1956), deal with pollution taboos as part of the general field of ritual prohibitions. In the *Encyclopaedia of Religion and Ethics,* 10:455–505 (1919, reprinted 1955), the article "Purification" has many examples. Religious texts are among the best available sources: the Old Testament; the Egyptian Book of the Dead; and FRIEDRICH MAX MÜLLER (ed.), *The Sacred Books of the East,* 51 vol. (1879–1904), available in many later editions. The latter includes texts of Hinduism, Buddhism, Islām, Zoroastrianism, and Taoism. For a good summary of Zoroastrian purification rites, see J.J. MODI, *The Religious Ceremonies and Customs of the Parsees* (1922, reprinted 1979; 2nd ed., 1937). For ancient Greece, see LOUIS MOULINIER, *Le Pur et l'impur dans la pensée des Grecs, d'Homère à Aristote* (1952); and JANE HARRISON, *Prolegomena to the Study of Greek Religion,* 3rd ed. (1922). For excellent synopses of African thought systems, see DARYLL FORDE (ed.), *African Worlds* (1963); and, for a North American tribe, see GLADYS REICHARD, *Navaho Religion,* 2nd ed. (1963). MARGARET MEAD, *Sex and Temperament in Three Primitive Societies* (1935; reprinted, 1968), brings together material on three Pacific Islands societies.

Dietary laws and food customs: General works include DON-

ALD E. CARR, *The Deadly Feast of Life* (1971), a popular account of food habits and nutritional behaviour; MARY DOUGLAS, *Purity and Danger: An Analysis of Concepts of Pollution and Taboo* (1966), a definitive source; and CRAIG MCANDREW and ROBERT B. EDGERTON, *Drunken Comportment: A Social Explanation* (1969), an exploration of the ways people are expected to behave under the influence of alcohol in different cultures.

The following discuss food customs and dietary laws in tribal societies: RAYMOND FIRTH, *We, the Tikopia: A Sociological Study of Kinship in Primitive Polynesia* (1936); MEYER FORTES, "Pietas in Ancestor Worship," *Jl. R. Anthrop. Inst.,* 91:166–191 (1961), reprinted in *Man in Adaptation,* vol. 3, *The Institutional Framework,* ed. by YEHUDI A. COHEN, pp. 207–226 (1971); and MARGARET MEAD, *The Mountain Arapesh,* vol. 2 (1970).

The basic sources for Judaism and Christianity are, of course, the Old Testament (especially Leviticus 11, Deuteronomy 14, and the prophets) and the New Testament (especially Acts, Luke, Mark, and Romans). See also JOHANNES PEDERSEN, *Israel: Its Life and Culture,* 4 vol. (1926–40); and MARK ZBOROWSKI and ELIZABETH HERZOG, *Life Is with People* (1952, reprinted 1962), on the shtetl.

The following Islāmic sources may be consulted: the Qurʾān; AMEER ALI, *Mohammedan Law,* 5th ed., 2 vol. (1929); and CHARLES C. TORREY, *The Jewish Foundation of Islam* (1933).

Sources on Indian systems include LOUIS DUMONT, *Homo hierarchicus, essai sur le système des castes* (1967; Eng. trans., *Homo Hierarchicus: The Caste System and Its Implications,* 1970); EDWARD B. HARPER (ed.), *Religion in South Asia* (1964), especially Harper's "Ritual Pollution As an Integrator of Caste and Religion," pp. 151–196; EDMUND R. LEACH (ed.), *Aspects of Caste in South India, Ceylon, and North-west Pakistan* (1960); DAVID G. MANDELBAUM, *Society in India,* 2 vol. (1970); MCKIM MARRIOTT, "Caste Ranking and Food Transactions: A Matrix Analysis," *Structure and Change in Indian Society,* ed. by MILTON B. SINGER and BERNARD S. COHN, pp. 133–171 (1968); KENNETH K.S. CHʾEN, *Buddhism: The Light of Asia* (1968); and CHARLES NORTON ELIOT, *Hinduism and Buddhism: An Historical Sketch,* 3 vol. (1921).

For the dietary laws and customs of Japan and China, see ROBERT N. BELLAH, *Tokugawa Religion* (1957); GEORGE DE VOS and HIROSHI WAGATSUMA (eds.), *Japan's Invisible Race: Caste in Culture and Personality* (1966); KENNETH K.S. CHʾEN, *Buddhism in China* (1964); and ARTHUR F. WRIGHT, *Buddhism in Chinese History* (1959).

Ceremonial and ritualistic objects: JAMES HASTINGS (ed.), *Encyclopaedia of Religion and Ethics,* 13 vol. (1908–26, reprinted 1955), although outdated, is a very complete general source. See also the *Histoire générale des religions,* 2nd ed., 2 vol. (1960); *Symbolisme cosmique et monuments religieux,* 2 vol. (1953), texts and illustrations from an exhibit at the Musée Guimet, Paris; and *Le Symbolisme cosmique des monuments religieux* (1957), the proceedings of an international conference of the Istituto per il Medio ed Estremo-Oriente, Rome. For the ancient period, see CHARLES V. DAREMBERG and EDMOND SAGLIO (eds.), *Dictionnaire des antiquités grecques et romaines . . .,* 5 vol. (1877–1919); PIERRE LAVEDAN, *Dictionnaire illustré de la mythologie et des antiquités grecques et romaines* (1931); and MIRCEA ELIADE, *Le Mythe de l'éternel retour* (1949; Eng. trans., *The Myth of the Eternal Return,* 1954, reissued 1989).

JEANNINE AUBOYER, *Introduction à l'étude de l'art de l'Inde* (1965), is a basic work on the Holy Place; for the principal components, see JAMES FERGUSSON, *Tree and Serpent Worship,* 2nd ed. (1873), which uses Indian facts as a base but makes many comparisons with data from antiquity. This work is complemented by ODETTE VIENNOT, *Le Culte de l'arbre dans l'Inde ancienne* (1954). JEANNINE AUBOYER, *Le Trône et son symbolisme dans l'Inde ancienne* (1949), makes many references to the role and the morphology of the throne (royal and divine) in ancient and modern civilizations. D.R. SHASTRI, *Origin and Development of the Rituals of Ancestor Worship in India* (1963), is helpful. JEAN PRZYLUSKI, "Le Symbolisme du pilier de Sarnath," in *Mélanges d'Orientalisme,* vol. 2 (1932), deals with the gnomon and the cosmic pivot. See also LEOPOLD M. CADIÈRE, *Croyances et pratiques religieuses des Annamites dans les environs de Hue,* 2 vol. (1918–19).

For architectural symbolism, GEORGE COEDES, *Pour mieux comprendre Angkor,* rev. 2nd ed. (1947; Eng. trans., *Angkor: An Introduction,* 1963), contains pertinent information in the chapters on temples and tombs and on architectural symbolism. ROLF A. STEIN, "Architecture et pensée religieuse en Extrême-Orient," *Arts Asiatiques,* 4:163–186 (1957), deals with tents used in Central Asia, especially Siberia, and Rupestral temples.

Icons and ritual symbols are discussed in J.N. BANERJEA, *The Development of Hindu Iconography,* 2nd ed. (1956), particularly ch. 2, "The Antiquity of Image-Worship in India," ch. 5, "Deities and Their Emblems on Early Indian Seals," and ch. 8, "Canons of Iconometry"; GEORGE COEDES (*op. cit.*), ch. 3; and PIERRE FRANCASTEL (ed.), *Emblèmes, totems, blasons* (1964), an exhibition catalog produced by the Musée Guimet, Paris.

Cultic and ritual objects are discussed in HENRIETTE DEMOULIN-BERNARD, *Masques . . . exposés dans l'annexe du Musée Guimet en décembre 1959* (1965); for Judaism, see JAMES HASTINGS, *A Dictionary of the Bible,* rev. ed. (1963); for Christianity, *Historia Religionum,* vol. 1, *Religions of the Past* (1969); and OSCAR CULLMANN, *Urchristentum und Gottesdienst,* 2nd ed. (1950; Eng. trans., *Early Christian Worship,* 1953); for Hinduism, *The Cultural Heritage of India,* vol. 1, *Vedic Rituals,* rev. ed. (1958); and PAUL E. DUMONT, *L'Aśvamedha: descriptions du sacrifice solennel du cheval dans le culte védique* (1927); for Buddhism, *Hōbōgirin: dictionnaire encyclopédique du Bouddhisme . . .,* 4 vol. to date (1937–67); and GEORGE P. MALALASEKERA (ed.), *Encyclopaedia of Buddhism* (1961–), appearing in fascicles; for the Indian world, JAN GONDA, *Die Religionen Indiens,* 3 vol. (1960–64); for Indonesia, WALDEMAR STOEHR and PIET ZOETMULDER, *Die Religionen Indonesiens* (1965); for the Islāmic world, the *Encyclopaedia of Islam,* 5 vol. (1908–38; new ed., 1960–); for Tibet, ROBERT B. EKVALL, *Religious Observances in Tibet* (1964); HELMUT HOFFMANN, *Symbolik der tibetischen Religionen und des Schamanismus* (1967); ROLF A. STEIN, *La Civilisation tibétaine* (1962); and TURRELL WYLIE, "Apropos of Tibetan Religious Observances," *Journal of the American Oriental Society,* 86:39–45 (1966); and for Japan, WILLIAM G. ASHTON, *Shinto* (1905).

Religious dress and vestments: HILAIRE and MEYER HILER, *A Bibliography of Costume* (1939, reprinted 1967), furnishes the widest bibliographical account. For Christian dress, see HERBERT NORRIS, *Church Vestments: Their Origin and Development* (1949); and CYRIL E. POCKNÉE, *Liturgical Vesture: Its Origins and Development* (1960), a succinct, well-illustrated account. All previous research was superseded by JOSEPH BRAUN, *Die liturgische Gewandung im Occident und Orient . . .* (1907), which marked a turning point in liturgiological studies. JOHN B. O'CONNELL, *The Celebration of Mass,* new ed. (1956), a handbook for priests, is a study of the rubrics. This can be supplemented by ADRIAN FORTESCUE, *Ceremonies of the Roman Rite Described,* 12th rev. ed. (1962), which indicates the appropriate use of garments and is a standard work on Roman Catholic ritual. Concerning Orthodox vesture, N.F. ROBINSON, *Monasticism in the Orthodox Churches* (1916, reprinted 1971); and A.N. SVIRIN, *Drevnerusskoe shitʹe* (1963), focussing on embroidery for vestments, are descriptive rather than historical or analytical in methodology. Protestant vesture has not attracted the attention of liturgists, but PERCY DEARMER, *The Parson's Handbook,* 13th rev. ed. (1965); and CHARLES WALKER, *The Ritual Reason Why,* new ed. (1931), treat the subject from a High Anglican standpoint. On religious dress in Judaism, *The Universal Jewish Encyclopaedia,* 10 vol. (1939–43); and WILLIAM OESTERLEY and G.H. BOX, *The Religion and Worship of the Synagogue* (1907), are useful. On Islāmic dress, there is MOHAMMAD BAQIR OLMAJLISI, *Helyet ol-motaqqīn* (1952). On East Asian religions, JEAN HERBERT, *Shintô* (1967), is a standard study of the subject and incorporates drawings of priestly attire. S. ONO, *Shinto: The Kami Way* (1960), also illustrates priestly attire. On Buddhist religious dress and vestments, *The Book of the Discipline,* vol. 4, trans. by I.B. HORNER (1951), gives the early Buddhist traditions about "the robe." HOLMES WELCH, *The Practice of Chinese Buddhism,* vol. 1, *1900–1950* (1967), contains much incidental material about monastic vestments, with photographs.

Feasts and festivals: E.O. JAMES, *Seasonal Feasts and Festivals* (1961), is a classic work on feasts and festivals of seasonal renewal and on Western folk festivals and customs from prehistoric times to the early 20th century. MIRCEA ELIADE, *The Sacred and the Profane* (1959), is a classic treatment of the concept of sacred time. W. BREDE KRISTENSEN, *The Meaning of Religion* (1960), is a phenomenological treatment of feasts and festivals. C. JOUCO BLEEKER and GEO WIDENGREN (eds.), *Historia Religionum: Handbook for the History of Religions,* 2 vol. (1969–71), incorporates feasts and festivals into the general framework of particular religions. H.W. PARKE, *Festivals of the Athenians* (1977), describes these ancient celebrations. WALTER KRICKEBERG *et al., Die Religionen des Alten Amerika* (1961; Eng. trans., *Pre-Columbian American Religions,* 1968), includes treatment of various feasts, festivals, and associated rites. JOHN S. MBITI, *African Religions and Philosophy,* pp. 110–165 (1969), covers feasts and festivals associated with African passage rites. DERK BODDE, *Festivals in Classical China* (1975), has important information on the Han dynasty. JEAN HERBERT, *Aux Sources du Japon: le Shintô* (1964; Eng. trans., *Shinto,* 1967), covers the feasts and festivals of Shintô in detail (pp. 147–224). JOHN B. NOSS, *Man's Religions,* 6th ed. (1980), provides the best single-volume coverage of the feasts and festivals of the various religions of the world. P. RAFAEL AVILA, *Worship and Politics* (1981), is a historical overview of religious feasts.

Rivers

By original usage, a river is flowing water in a channel with defined banks (ultimately from Latin *ripa,* "bank"). Modern usage includes rivers that are multichanneled, intermittent, or ephemeral in flow and channels that are practically bankless. The concept of channeled surface flow, however, remains central to the definition.

The word stream (derived ultimately from the Indo-European root *srou-*) emphasizes the fact of flow; as a noun it is synonymous with river and is often preferred in technical writing. Small natural watercourses are sometimes called rivulets, but a variety of names—including branch, brook, burn, and creek—are more common, occurring regionally to nationally in place-names. Arroyo and (dry) wash connote ephemeral streams or their resultant channels. Tiny streams or channels are often referred to as rills or runnels.

Rivers are nourished by precipitation, by direct overland runoff, through springs and seepages, or from meltwater at the edges of snowfields and glaciers. The contribution of direct precipitation on the water surface is usually minute, except where much of a catchment area is occupied by lakes. River water losses result from seepage and percolation into shallow or deep aquifers (permeable rock layers that readily transmit water) and particularly from evaporation. The difference between the water input and loss sustains surface discharge or streamflow.

The amount of water in river systems at any time is but a tiny fraction of the Earth's total water; 97 percent of all water is contained in the oceans, and about three-quarters of fresh water is stored as land ice; nearly all the remainder occurs as groundwater. Lakes hold less than 0.5 percent of all fresh water on the Earth, soil moisture accounts for about 0.05 percent, and water in river channels for roughly half as much, 0.025 percent, which represents only about one four-thousandth of the Earth's total fresh water.

Water is constantly cycled through the systems of land ice, soil, lakes, groundwater (in part), and river channels, however. The discharge of rivers to the oceans delivers to these systems the equivalent of the water vapour that is blown overland and then consequently precipitated as rain or snow—*i.e.,* some 7 percent of mean annual precipitation on the globe and 30 percent of precipitation on land areas.

Rivers are 100 times more effective than coastal erosion in delivering rock debris to the sea. Their rate of sediment delivery is equivalent to an average lowering of the lands by 30 centimetres (12 inches) in 9,000 years, a rate that is sufficient to remove all the existing continental relief in 25,000,000 years.

Rock debris enters fluvial systems either as fragments eroded from rocky channels or in dissolved form. During transit downstream, the solid particles undergo systematic changes in size and shape, traveling as bed load or suspension load. Generally speaking, except in high latitudes and on steep coasts, little or no coarse bed load ever reaches the sea. Movement of the solid load down a river valley is irregular, both because the streamflow is irregular and because the transported material is liable to enter temporary storage, forming distinctive river-built features that range through riffles, midstream bars, point bars, floodplains, levees, alluvial fans, and river terraces. In one sense, such geomorphic features belong to the same series as deltas, estuary fills, and the terrestrial sediments of many inland basins.

Rates of erosion and transportation, and comparative amounts of solid and dissolved load, vary widely from river to river. Least is known about dissolved load, which at coastal outlets is added to oceanic salt. Its concentration in tropical rivers is not necessarily high, although very high discharges can move large amounts; the dissolved load of the lowermost Amazon averages about 40 parts per million, whereas the Elbe and the Rio Grande, by contrast, average more than 800 parts per million. Suspended load for the world in general perhaps equals two and one-half times the amounts of dissolved load. Well over half of suspended load is deposited at river mouths as deltaic and estuarine sediment. About one-quarter of all the world's suspended load is estimated to come down the Ganges–Brahmaputra and the Huang Ho (Yellow River), which together deliver some 4,500,000,000 tons a year; the Yangtze, Indus, Amazon, and Mississippi deliver quantities ranging from about 500,000,000 to approximately 350,000,000 tons a year. Suspended sediment transport on the Huang Ho equals a denudation rate of about 3,090 tons per square kilometre (8,000 tons per square mile) per year; the corresponding rate for the Ganges–Brahmaputra system is almost half as great. Extraordinarily high rates have been recorded for some lesser rivers: for instance, 1,060 tons per square kilometre per year on the Ching and 1,080 tons per square kilometre per year on the Lo, both of which are loess-plateau tributaries of the Huang Ho.

This article concentrates on the distribution, drainage patterns, and geometry of river systems; its coverage of the latter includes a discussion of channel patterns and such related features as waterfalls. Considerable attention is also given to fluvial landforms and to the processes involved in their formation.

Additional information about the action of flowing water on the Earth's surface is provided in the articles GEOMORPHIC PROCESSES and CONTINENTAL LANDFORMS. Certain aspects of the changes in rivers through time are described in the article CLIMATE AND WEATHER, and the general interrelationship of river systems to other components of the Earth's hydrosphere is treated in HYDROSPHERE, THE. For information concerning the plant and animal forms that inhabit the riverine environment, see BIOSPHERE, THE.

For coverage of other related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 222, 231, and 232, and the *Index*.

The article is divided into the following sections:

Importance of rivers 844
  Significance in early human settlements 844
  Significance to trade, agriculture, and industry 844
  Environmental problems attendant on river use 845
Distribution of rivers in nature 846
  World's largest rivers 846
  Principles governing distribution and flow 847
    Variation of stream regime
    Determining factors
Drainage patterns 848
  Horton's laws of drainage composition 849
  Morphometry of drainage networks 849
    Relation of morphometric parameters and river flow
    Evolution of drainage systems
Geometry of river systems 850
  Hydraulic geometry 850
  River channel patterns 850
    Straight channels
    Meandering channels
    Braided channels
  Waterfalls 852
    World distribution of waterfalls
    Types of waterfalls
    Development of waterfalls
Streamflow and sediment yield 856
  Peak discharge and flooding 856

Sediment yield and sediment load 857
Sources of sediment and nature of deposition
Factors that influence sediment yield
Rivers as agents of landscape evolution 861
Valleys and canyons 861
Valley evolution
Formation of canyons and gorges
Floodplains 863
Floodplain deposits, origins, and features
Time and the floodplain system
Deltas 864
Classification of deltas
Morphology of deltas
Deposits and stratigraphy
Deltas and time
Estuaries 867
Origin and classification
Sedimentation in estuaries
The river system through time 868
Drainage diversion by stream capture 868
Non-fluvial invasion and deposition 868
Effects of climatic change 868
Bibliography 869

Importance of rivers

SIGNIFICANCE IN EARLY HUMAN SETTLEMENTS

The inner valleys of some great alluvial rivers contain the sites of ancestral permanent settlements, including pioneer cities. Sedentary settlement in Hither Asia began about 10,000 years ago at the site of Arīḥā (ancient Jericho). Similar settlement in the Tigris–Euphrates and Nile valleys dates back to at least 6000 BP (years before present). The first settlers are thought to have practiced a hunting economy, supplemented by harvesting of wild grain. Conversion to the management of domesticated animals and the cultivation of food crops provided the surpluses that made possible the rise of towns, with parts of their populations freed from direct dependence on food getting. Civilization in the Indus Valley, prominently represented at Mohenjo-daro, dates from about 4500 BP, while civilization in the Ganges Valley can be traced to approximately 3000 BP. Permanent settlement in the valley of the Huang Ho has a history some 4,000 years long, and the first large irrigation system in the Yangtze catchment dates to roughly the same time. Greek invaders of the Syrdarya, Amu Darya, and other valleys draining to the Aral Sea, east of the Caspian, encountered irrigating communities that had developed from about 2300 BP onward.

Sites of major prehistoric communities

The influence of climatic shifts on these prehistoric communities has yet to be worked out satisfactorily. In wide areas, these shifts included episodic desiccation from 12,000 or 10,000 BP onward. In what are now desert environments, increased dependence on the rivers may have proved as much a matter of necessity as of choice. All of the rivers in question have broad floodplains subject to annual inundation by rivers carrying heavy sediment loads.

Prehistoric works of flood defense and irrigation demanded firm community structures and required the development of engineering practice. Highly elaborate irrigation works are known from Mohenjo-daro; the ziggurats (temple mounds) of the Euphrates Valley may well have originated in ancient Egypt in response to the complete annual inundation of the Nile floodplain, where holdings had to be redefined after each flood subsided. It is not surprising that the communities named have been styled hydraulic civilizations. Yet, it would be oversimplistic to claim that riparian sites held the monopoly of the developments described. Elaborate urban systems arising in Mexico, Peru, and the eastern Mediterranean from about 4000 BP onward were not immediately dependent on the resources of rivers.

Where riverine cities did develop, they commanded ready means of communication; the two lands of Upper and Lower Egypt, for instance, were unified by the Nile. At the same time, it can be argued that early riverine and river-dependent civilizations bore the seeds of their own destruction, independent of major climatic variations and natural evolutionary changes in the river systems. High-consuming cities downstream inevitably exploited the upstream catchments, especially for timber. Deforestation there may possibly have led to ruinous silting in downstream reaches, although the contribution of this process to the eventual decline of civilization on the Euphrates and the Indus remains largely a matter of guesswork. An alternative or conjoint possibility is that continued irrigation promoted progressive salinization of the soils of irrigated lands, eventually preventing effective cropping. Salinization is known to have damaged the irrigated lands of Ur, progressively from about 4400 to 4000 BP, and may have ruined the Sumerian Empire of the time. The relative importance of environmental and social deterioration in prehistoric hydraulic civilizations, however, remains a matter of debate. Furthermore, defective design and maintenance of irrigation works promote the spread of malarial mosquitoes, which certainly afflicted the prehistoric hydraulic communities of the lower Tigris–Euphrates Valley. These same communities also may have been affected by bilharziasis, or schistosomiasis (blood fluke disease), which requires a species of freshwater snail for propagation and which even today follows many extensions of irrigation into arid lands.

At various intervals of history, rivers have provided the easiest, and in many areas the only, means of entry and circulation for explorers, traders, conquerors, and settlers. They assumed considerable importance in Europe after the fall of the Roman Empire and the dismemberment of its roads; regardless of political structures, control of crossing points was expressed in strongholds and the rise of bridge-towns. Rivers in medieval Europe supplied the water that sustained cities and the sewers that carried away city waste, and they were widely used, either directly or with offtakes, as power sources. Western European history records the rise of 13 national capitals on sizable rivers, exclusive of seawater inlets; three of them, Vienna, Budapest, and Belgrade, lie on the Danube, with two others, Sofia and Bucharest, on feeder streams above stem floodplain level. The location of provincial and corresponding capitals is even more strongly tied to riparian sites, as can be readily seen from the situation in the United Kingdom, France, and Germany. In modern history, in both North America and northern Asia, natural waterways directed the lines of exploration, conquest, and settlement. In these areas, passage from river system to river system was facilitated by portage along lines defined by temporary ice-marginal or ice-diverted channels. Many pioneer settlers of the North American interior entered by means of natural waterways, especially in Ohio.

Importance of rivers of the rise of urban centres

SIGNIFICANCE TO TRADE, AGRICULTURE, AND INDUSTRY

The historical record includes marked shifts in the appreciation of rivers, numerous conflicts in use demand, and an intensification of use that has rapidly accelerated during the 20th century. External freight trade became concentrated in estuarine ports rather than in inland ports when oceangoing vessels increased in size. Even the port of London, though constrained by high capital investment, has displaced itself toward its estuary. The Amazon remains naturally navigable by ocean ships for 3,700 kilometres (2,300 miles), the Yangtze for 1,000 kilometres, and the partly artificial St. Lawrence Seaway for 2,100 kilometres. Internal freight traffic on the Rhine system and its associated canals amounts to one-quarter or more of the total traffic in the basin and to more than half in some parts. After a period of decline from the later 1800s to about the mid-1900s, water transport of freight has steadily increased. This trend can in large part be attributed to advances in river engineering. Large-scale channel improvement and stabilization projects have been undertaken on many of the major rivers of the world, notably in the northern plain lands of Russia and in the interior of the United States (*e.g.*, various large tributaries of the Mississippi River).

Demand on open-channel water increases as population

and per capita water use increase and as underground water supplies fall short. Irrigation use constitutes a comparatively large percentage of the total supply. With a history of at least 5,000 years, controlled irrigation now affects roughly 2,000,000 square kilometres (770,000 square miles) of land, three-quarters of it in East and South Asia and two-fifths in mainland China alone. Most of this activity involves the use of natural floodwater, although reliance on artificially impounded storage has increased rapidly. Irrigation in the 1,300-kilometre length of the Indus Valley, for instance, depends almost exclusively on barrages (*i.e.*, distributor canals) running down alluvial fans and along floodplains.

Present-day demands on rivers as power sources range from the floating of timber, through the use of water for cooling, to hydroelectric generation. Logging in forests relies primarily on flotation during the season of meltwater high flow. Large power plants and other industrial facilities are often located along rivers, which supply the enormous quantities of water needed for cooling purposes (see below). Manufacturers of petrochemicals, steel, and woolen cloth also make large demands. Hydroelectric power generation was introduced more than 100 years ago, but the majority of the existing installations have been built since 1950. Many of the world's major industrial nations have developed their hydropower potential to the fullest, though a few like the United States still have some untapped resources. It has been estimated that 75 percent of the potential hydropower in the contiguous United States has been developed, and about 13 percent of the total annual electrical energy demands of the country are met by hydroelectric power plants. By contrast, there are some countries, such as Norway and Switzerland, that depend almost entirely on hydropower for their various electrical energy needs. There is great potential for supplies of hydropower in Central Asia and in many of the developing countries in the region of the Himalayas, Africa, and South America.

Hydroelectric power generation

Use demand of more immediate kinds are related to freshwater fisheries (including fish-farming), to dwelling in houseboats, and to recreational activities. Reliable data for these kinds of dependence on rivers do not exist; published estimates that freshwater and migratory fish provide up to about 15 percent of world catch may be too low. Certainly, millions of people are concerned with freshwater fishery and houseboat living, principally in the deltaic areas of East Asia, where dwelling, marketing, and travel can be located almost exclusively on the water. Furthermore, recreational use of rivers has increased over the years. In North America many waterways, particularly those with relatively light commercial traffic, support large numbers of recreational craft. In Europe pleasure cruisers transport multitudes of sightseers up and down the Rhine and Seine each year, while various derelict canals of such systems as the Thames have been restored for boating.

(G.H.D./Ed.)

ENVIRONMENTAL PROBLEMS ATTENDANT ON RIVER USE

The ever-increasing exploitation of rivers has given rise to a variety of problems. Extensive commercial navigation of rivers has resulted in much artificial improvement of natural channels, including increasing the depth of the channels to permit passage of larger vessels. In some cases, this lowering of the river bottom has caused the water table of the surrounding area to drop, which has adversely affected agriculture. Also, canalization, with its extensive system of locks and navigation dams, often seriously disrupts riverine ecosystems.

An even more far-reaching problem is that of water pollution. Pesticides and herbicides are now employed in large quantities throughout much of the world. The widespread use of such biocides and the universal nature of water makes it inevitable that the toxic chemicals would appear as stream pollutants. Biocides can contaminate water, especially of slow-flowing rivers, and are responsible for a number of fish kills each year.

Contamination by chemicals

In agricultural areas the extensive use of phosphates and nitrates as fertilizers may result in other problems. Entering rivers via rainwater runoff and groundwater seepage, these chemicals can cause eutrophication. This process involves a sharp increase in the concentration of phosphorus, nitrogen, and other plant nutrients that promotes the rapid growth of algae (so-called algal blooms) in sluggish rivers and a consequent depletion of oxygen in the water. Under normal conditions, algae contribute to the oxygen balance in rivers and also serve as food for fish, but in excessive amounts they crowd out populations of other organisms, overgrow, and finally die owing to the exhaustion of available nutrients and autointoxication. Various species of bacteria then begin to decay and putrefy the dead algal bodies, the oxidation of which sharply reduces the amount of oxygen in the river water. The water may develop a bad taste and is unfit for human consumption unless filtered and specially treated.

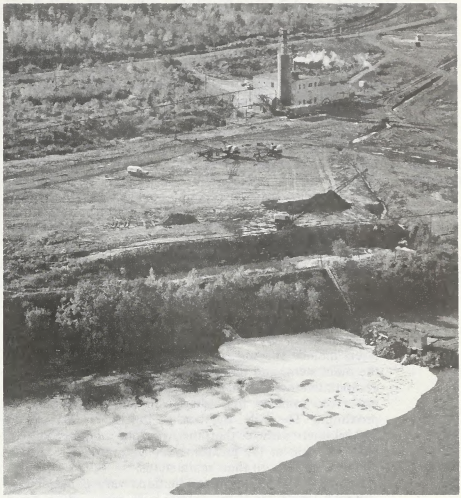


Figure 1: River polluted by untreated chemical wastes from a nearby industrial plant (right background).
Grant Heilman Photography

Urban centres located along rivers contribute significantly to the pollution problem as well. In spite of the availability of advanced waste-purification technology, a surprisingly large percentage of the sewage from cities and towns is released into waterways untreated. In effect, rivers are used as open sewers for municipal wastes, which results not only in the direct degradation of water quality but also in eutrophication.

Municipal and industrial wastes

Still another major source of pollutants is industry. Untreated industrial chemical wastes can alter the normal biological activity of rivers, and many of the chemicals react with water to raise the acidity of rivers to a point where the water becomes corrosive enough to destroy living organisms. An example of this is the formation of sulfuric acid from the sulfur-laden residue of coal-mining operations. Although upper limits for concentrations of unquestionably toxic chemicals such as arsenic, barium, cyanide, lead, and phenols have been established for drinking water, no general rules exist for the treatment of industrial wastes because of the wide variety of organic and inorganic compounds involved. Moreover, even in cases where a government-imposed ban checks the further discharge of certain dangerous substances into waterways, the chemicals may persist in the environment for years. Such is the case with polychlorinated biphenyls (PCBs), the chlorinated hydrocarbon by-products of various industrial processes that were routinely discharged into U.S. waterways until the late 1970s when the federal government not only prohibited the continued discharge of the chemicals

into the environment but their production as well. Since PCBs cannot be broken down by conventional waste-treatment methods and are degraded by natural processes very slowly, scientists fear that these compounds will continue to pose a serious hazard for decades to come. PCBs have been found in high concentrations in the fatty tissues of fish, which can be passed up the food chain to humans. An accumulation of PCBs in the human body is known to induce cancer and other severe disorders.

Thermal pollution

As noted above, many industrial facilities, including nuclear power plants, steel mills, chemical-processing facilities, and oil refineries, use large quantities of water for cooling and return it at elevated temperatures. Such heated water can alter the existing ecology, sometimes sufficiently to drive out or kill desirable species of fish. It also may cause rapid depletion of the oxygen supply by promoting algal blooms. (Ed.)

Distribution of rivers in nature

WORLD'S LARGEST RIVERS

Obvious bases by which to compare the world's great rivers include the size of the drainage area, the length of the main stem, and the mean discharge; however, reliable comparative data, even for the world's greatest rivers, do not exist. Some of the values listed in Table 1 are approximate. The Nile, the world's longest river, is about 250 kilometres longer than the Amazon. It is possible that well over 100 of the greatest rivers may exceed a 1,600-kilometre length on their main stems.

Drainage area, length, and discharge

Area–length–discharge combinations vary considerably, although length tends to increase with area and area and discharge to increase through their individual ranking series. On all counts except length, the Amazon is the world's principal river; the Congo and the Paraná are among the first five by area and discharge, but the Mississippi, fourth in length and fifth in area, is only seventh in discharge. The Ganges–Brahmaputra, third in discharge, is 13th (or lower) in area and well down the list of length for its two main stems taken separately.

Ranking in Table 1 is by drainage area. In combination, the rivers listed drain some 44,000,000 square kilometres, roughly 30 percent of the world's land area. If volume of discharge is taken to be the basis of comparison, then certain other rivers not tabulated also must be mentioned. The most important of these is the Orinoco, with a mean discharge of 19,800 cubic metres (700,000 cubic feet) per second and a basin of 948,000 square kilometres. Others are the Irrawaddy, discharge 13,000 cubic metres per second, basin 411,000 square kilometres; and the Mekong, 11,000 cubic metres per second, basin 795,000 square kilometres. The 20 greatest of these rivers, draining about 30 percent of the world's land area, discharge nearly 40 percent of total runoff, reckoned from a mean equivalent of 29.2 centimetres of precipitation. They deliver to the sea about 92 cubic kilometres of water per day, or roughly 33,325 cubic kilometres annually. The Amazon, the Paraná, the Congo, and the Ganges–Brahmaputra, combined, discharge more than 54 cubic kilometres a day and nearly 20,800 cubic kilometres a year, one-third of the world's total runoff to the oceans, with the Amazon alone accounting for almost one-fifth.

Higher-than-average discharges

World average external runoff is about 0.01 cubic metre per second per square kilometre (0.6 cubic foot per second per square mile). Great rivers with notably higher discharges are fed either by the convectional rains of equatorial regions or by monsoonal rains that are usually increased by altitudinal effects. The Huang Ho averages 0.046 cubic metre per second per square kilometre, the Irrawaddy 0.032 cubic metre per second per square kilometre, the Magdalena and the Amazon 0.026 cubic metre per second per square kilometre, the Orinoco 0.021 cubic metre per second per square kilometre, and the Ganges–Brahmaputra above 0.024 cubic metre per second per square kilometre. Very high mean discharges per unit area are also recorded for lesser basins in mountainous coastlands exposed to the zonal westerlies of mid-latitudes. Among great rivers with mean discharges near or not far below world averages per unit area are those of Siberia, the Mackenzie, and the Yukon (828,000 square kilometres, 5,900 cubic metres per second), all affected by low precipitation for which low evaporation rates barely compensate. The basins of the Mississippi, Niger, and Zambezi include some areas of dry climate. The Nelson illustrates the extreme effects of low precipitation in a cool climate, while the Nile, Murray–Darling, and Shaṭṭ al-'Arab (Tigris–Euphrates) experience low precipitation combined with high evaporation losses.

Table 1: The World's Principal Rivers, Ranked According to Drainage Area

river	drainage area		length (km)*	mean discharge			
	extent (000 sq km)	percent of world's land area		(000 cu m/sec)	rank order	percent of world total	cu m/sec/sq km
Amazon	7,050	4.8	6,400	180	1	19.2	.0255
Paraná	4,144	2.8	4,880	22	5	2.3	.0052
Congo	3,457	2.3	4,700	41	2	4.4	.0121
Nile	3,349	2.3	6,650	3	—	0.3	.0009
Mississippi–Missouri	3,221	2.2	6,020	18	8	2.0	.0057
Ob–Irtysh	2,975	2.0	5,410	15	10	1.7	.0053
Yenisey	2,580	1.7	5,540	19	6	2.0	.0073
Lena	2,490	1.7	4,400	16	9	1.7	.0065
Yangtze	1,959	1.3	6,300	34	4	3.6	.0174
Niger	1,890	1.3	4,200	6	—	0.7	.0032
Amur	1,855	1.3	2,824	12	10	1.3	.0066
Mackenzie	1,841	1.2	4,241	11	—	1.2	.0061
Ganges–Brahmaputra	1,621	1.1	2,897	38	3	4.1	.0237
St. Lawrence–Great Lakes	1,463	1.0	4,000	10	—	1.1	.0069
Volga	1,360	0.9	3,530	8	—	0.9	.0058
Zambezi	1,330	0.9	3,500	7	—	0.8	.0053
Indus	1,166	0.8	2,900	5	—	0.6	.0047
Shaṭṭ al-'Arab (Tigris–Euphrates)	1,114	0.8	2,800	1	—	0.1	.0012
Nelson	1,072	0.7	2,575	2	—	0.2	.0021
Murray–Darling	1,057	0.7	3,780	0.4	—	0.04	.0003
Tocantins	906	0.6	2,699	10	—	1.1	.0112
Danube	816	0.6	2,850	7	—	0.8	.0088
Columbia	668	0.5	2,000	7	—	0.7	.0104
Rio Grande	445	0.4	1,360	0.08	—	0.01	.0001
Rhine	160	0.1	1,392	2	—	0.2	.0137
Rhône	96	—	800	2	—	0.2	.0177
Thames	10	—	340	0.08	—	0.01	.0082

*Some figures are rounded to the nearest hundred kilometres.

The lower end of Table 1 lists comparative data for selected rivers in highly inhabited or otherwise hydrographically interesting valleys. The Rhine, Rhône, and Danube record regimes that vary along the length of their courses in response to glacier melt in the headwaters and the entry of contrasting tributaries downstream. The Rio Grande, like the Orange and the Colorado, suffers progressive downstream losses, both natural and irrigational. The Thames is special, as it experiences a very high tidal range in its estuary; this makes flood control especially difficult.

PRINCIPLES GOVERNING DISTRIBUTION AND FLOW

Moisture supply sufficient to sustain channeled surface flow is governed primarily by climate, which regulates precipitation, temperature, and evapotranspiration water loss caused by vegetation. In rainy tropical and exposed midlatitude areas, runoff commonly equals 38 centimetres or more of rain a year, rising to more than 102 centimetres. Negligible external runoff occurs in subtropical and rainshadow deserts; perennial, intermittent, and ephemeral lakes, expanding in response to local runoff, prevent the drainage of desert basins from finding escape routes.

Principal classes of regime

Variation of stream regime. Seasonal variation in discharge defines river regime. Three broad classes of regime can be distinguished for perennial streams. In the megathermal class, related to hot equatorial and tropical climates, two main variants occur; discharge is powerfully sustained throughout the year, usually with a double maximum (two peak values), but in some areas with a strong single maximum. In the mesothermal class some regimes resemble those of tropical and equatorial areas, with single or double summer maxima corresponding to heavy seasonal rainfall, while others include sustained flow with slight warm-season minima. Where mid-latitude climates include dry summers, streamflow decreases markedly and may cease altogether in the warm half of the year. In areas affected by release of meltwater, winter minima and spring maxima of discharge are characteristic. Microthermal regimes, which are influenced by snow cover, include winter minima and summer maxima resulting from snowmelt and convectional rain; alternatively, spring meltwater maxima are accompanied by secondary fall maxima that are associated with late-season thunder rain, or spring snowmelt maxima can be followed by a summer glacier-melt maximum, as on the Amu Darya. Megathermal regimes, which are controlled by systematic fluctuations in seasonal rain, and microthermal regimes, which are controlled by seasonal release of meltwater, may be more reliable than mesothermal regimes.

The regime can vary considerably along the length of a single river in timing and in seasonal characteristics. Spring maxima in the Volga headwaters are not followed by peak flows in the delta until two months later. The October seasonal peak on the upper Niger becomes a December peak on the middle river; the swing from tropical-rainy through steppe climate reduces volume by 25 percent through a 483-kilometre stretch. The seasonal headwater flood wave travels at 0.09 metre per second, taking some four months over 2,011 kilometres, but earlier seasonal peaks are reestablished on the lower river by tributaries fed by hot-season rains. The great Siberian rivers, flowing northward into regions of increasingly deferred thaw, habitually cause extensive flooding in their lower reaches, which remain ice-covered when upstream reaches have already thawed and are receiving the meltwaters of late spring and summer.

Extremes of regime characteristics come into question when streams are classified as perennial, intermittent, or ephemeral. These terms are in common use but lack rigid definition. Whereas the middle and lower reaches of streams in humid regions rarely or never cease flowing and can properly be called perennial, almost every year many of their upstream feeders run dry where they are not fed by springs. In basins cut in impermeable bedrock, prolonged droughts can halt flow in most channel reaches. Karst (limestone country) that has some surface drainage often includes streams that are spatially intermittent; frequently it also contains temporally intermittent streams that flow only when heavy rain raises the groundwater table and reactivates outlets above the usual level. Temporally intermittent streams also occur in dry areas where, at low stage, only some channel reaches contain flowing water.

There is a continuous progression from perennial streams through intermittent streams to ephemeral streams: the latter command much attention, especially because their effects in erosion, transportation, and deposition can be inordinately great and also because they relate closely to periods and cycles of gullying. Their channels generally have higher width–depth ratios than those of unbraided channels in humid areas; *e.g.*, 150:1 or more on small streams. In extreme cases, ephemeral streamflow merges into sheetflood. Streambeds, usually sandy, are nearly flat in cross section but contain low bars where gravel is available. These behave in many ways like riffles or braid bars elsewhere. Although beds and banks are erodible, the fine-material fraction is usually enough to sustain very steep channel banks and gully walls. Rapid downcutting produces flat-floored trenches, called arroyos, in distinction from the often V-shaped gullies of humid areas.

Arroyos

Discontinuous vegetation cover, well-packed surface soil, and occasionally intense rainfall promote rapid surface runoff, conversion of overland to channeled flow, and the multiplication of channels. Although reliable comparative data are scarce, it seems likely that ephemeral channel systems develop higher order ranking, area for area, than do perennial streams: channels as high as 11th order are recorded for basins of about 1,300 square kilometres, whereas the Mississippi is usually placed only in the 10th order (see below *Horton's laws of drainage composition*). This apart, geometry of ephemeral nets obeys the laws of drainage composition that apply to perennial streams: stream length, stream number, channel width, and water discharge can be expressed as exponential functions of stream order, and drainage area and channel slope as power functions, whereas slope and discharge can be expressed as power functions of width and drainage area.

At-a-station (a particular cross section) variations in width, depth, and velocity with variation in discharge in ephemeral streams resemble the corresponding variations in perennial streams. Differences appear, however, when downstream variations are considered. For a given frequency of discharge, the rate of increase in width differs little between the two groups, but ephemeral streams increase the more slowly in depth, becoming increasingly shallow in proportion in the downstream direction. This effect is compensated by a more rapid downstream increase in velocity, which reflects high concentrations of suspended sediment and a resultant reduction of friction. Ultimately, however, the ephemeral flood may lose so much water by evaporation and percolation that the stream is dissipated in a terminal mudflow.

Trenching, the extension of gullies, and their conversion into arroyo systems, implies valley fills of erodible surficial material. Like streams of humid regions, ephemeral stream systems record complex histories of cut and fill: it is reasonable to expect comparable timing for climatically controlled events. Whatever the effect upon stream erosion of historical settlement in the western United States, inland eastern Australia, and New Zealand, the present episode of gullying seems merely to have been intensified by man's use of the land. Accelerated channeling frequently involves three processes not characteristic of humid regions: piping, headcutting, and the formation of channel profiles that are discontinuous over short distances.

Piping and discontinuous gullies

In piping, water that has penetrated the topsoil washes out the subsoil where this is exposed in section, forming small tunnels that may attain lengths of many metres. Collapse of tunnel roofs initiates lateral gullying and lengthens existing cuts headward. Headcutting is commonly associated with piping, because headcuts frequently expose the subsoil. A headcut is an abrupt step in the channel profile, some centimetres to some metres high; it may originate merely as a bare or trampled patch in a vegetated channel bed but will increase in height (like some very large waterfalls) as it works upstream. At the foot of the headcut is a plunge pool, downstream of which occurs a depositional slope of low downstream gradient. Formation of successive headcuts, say at an average spacing of 150 metres, and

the construction of depositional slopes below each, causes the profile to become stepped. Ephemeral streams with stepped profiles are called discontinuous gullies. Speed of headcut recession varies widely with the incidence and intensity of rainfall; but ultimately, when the whole profile has been worked along and the bed widened, the original even slope is restored, though at a lower level than before.

Determining factors. Long-term effects expressed in mean seasonal regimes and short-term effects expressed in individual peak flows are alike affected by soil-moisture conditions, groundwater balance, and channel storage. Channeled surface flow begins when overland flow becomes deep enough to be erosive; and depth of overland flow represents a balance between short-term precipitation and soil infiltration. Rate and capacity of infiltration depend partly on antecedent conditions and partly on permeability. Seasonal assessments are possible, however; numbers of commercial crops can take up and transpire the equivalent of 38 centimetres of precipitation during the growing season. In many mid-latitude climates the rising curves of insolation and plant growth during spring and early summer cause soil moisture depletion, leading eventually to a deficit that is often strong enough to reduce runoff and streamflow. Soil-moisture recharge during colder months promotes high values of runoff frequently in the spring quite independently of the influence of precipitation regime or snowmelt.

Groundwater and glacier effects

Storage of water in groundwater tables, stream channels, on floodplains, and in lakes damps out variations in flow, whereas snow and ice storage exaggerate peaks. For the world as a whole, groundwater contributes perhaps 30 percent of total runoff, although the proportion varies widely from basin to basin, within basins, and through time. Shallow groundwater tables in contact with river channels absorb and release water, respectively at high and low stage. Percolation to greater depths and eventual discharge through springs delays the entry of water into channels; many groundwater reservoirs carry over some storage from one year to another. Similar carryover occurs with glaciers and to some extent also with permanent snowfields; water abstracted by the ice caps of high latitudes and by large mountain glaciers can be retained for many years, up to about 250,000 years in the central Antarctic cap. Temperate glaciers, however, with temperatures beneath the immediate subsurface constantly near the freezing (or the melting) point, can, like their associated snowfields, release large quantities of water during a given warm season. Their losses through evaporation are small.

Meltwater contributions to streamflow, however, can range from well above half the total discharge to well below the level of the snow line. They are vital to irrigation on alluvial fans rimming many dry basins, as in the Central Valley of California and the Tarim Basin of the Takla Makan Desert of China: meltwater is released during the planting or growing seasons. Within the limiting constraints of precipitation or meltwater input or both, and the outputs of evapotranspiration and percolation, the actual distribution of rivers in nature is affected by available drainage area, lithology, and vegetation. Vegetation is obviously climate dependent to a large extent but might well be capable of reaching thresholds of detention ability that do not match recognized climatic boundaries. It is, moreover, liable to the influence of climatically independent factors where it has been disturbed by human activity. Runoff on the plain lands of northern Asia, expressed as a percentage of mean annual precipitation, ranges from about 75 in the tundra, through about 70 in the boreal forest and 50 through boreal forest with perennially frozen ground, down through less than 40 in mixed forest, to five in semidesert. Clear felling of forest increases runoff in the short and medium term because it reduces surface detention and transpiration. In areas of seasonal snow cover, forest influences seasonal regime considerably. However, though there may be a jump in short-term runoff characteristics between areas of continuous vegetation (forest and grass sward) on the one hand and discontinuous vegetation (bunchgrass and scrub) on the other, comprehensive general studies of precipitation–temperature runoff characteristics suggest that mean annual runoff decreases, at a decreasing rate through the range that is involved, as temperature increases and as precipitation (weighted in respect of seasonal incidence) decreases.

Constraints of rock type and vegetation

Lithology is significant mainly in connection with permeability. The capacity of karst to swallow and to reissue water is well known, as is the role of permeable strata generally in absorbing water into groundwater tables. An extreme case of a special kind is represented by an artesian aquifer, which in favourable structural conditions can take water for a very long time from the surface and immediately connected circulations, returning it only if the artesian pressure becomes strong enough to promote the opening of flowing springs. Less directly, but with considerable effect on infiltration and short-term runoff, the mechanical grade of bedrock or of surficial deposits can considerably affect the response to individual storms.

Artesian aquifer

Both the ultimate possible extent of drainage basins and the opening of individual headwater channels are influenced by available drainage area. A hypothetical limit for very large basins could probably be constructed from considerations of stem length, basin shape, computed area, and continental extent. The Amazon probably approaches the hypothetical maximum. At the other extreme, basin morphometry (geometric aspects of basins and their measurement) can be made to indicate the limiting average area necessary to sustain a given length of channel; in large areas of the mid-latitudes, the ratio is close to 2.25 square kilometres of drainage area for a channel 1.6 kilometres in length. Estimates for the conterminous United States, an area of about 7,770,000 square kilometres, give some 5,230,000 kilometres of channel length. These estimates include 1,500,000 unbranched fingertip tributaries—each having an average length of 1.6 to 2.4 kilometres.

Drainage patterns

Distinctive patterns are acquired by stream networks in consequence of adjustment to geologic structure. In the early history of a network, and also when erosion is reactivated by earth movement or a fall in sea level, downcutting by trunk streams and extension of tributaries are most rapid on weak rocks, especially if these are impermeable, and along master joints and faults. Tributaries from those streams that cut and grow the fastest encroach on adjacent basins, eventually capturing parts of the competing networks therein. In this way, the principal valleys with their main drainage lines come to reflect the structural pattern.

Dendritic, trellis, and radial patterns

Flat-lying sedimentary rocks devoid of faults and strong joints and the flat glacial deposits of the Pleistocene Epoch (from 2,500,000 to 10,000 years ago) exert no structural control at all: this is reflected in branching networks (Figure 2). A variant pattern, in which trunk streams run subparallel, can occur on tilted strata. Rectangular patterns form where drainage lines are adjusted to sets of faults and marked joints that intersect at about right angles, as in some parts of ancient crustal blocks. The pattern is varied where the regional angle of structural intersection changes. Radial drainage is typical of volcanic cones, so long as they remain more or less intact. Erosion to the skeletal state often leaves the plug standing in high relief, ringed by concentric valleys developed in thick layers of ash.

Similarly, on structural domes where the rocks of the core vary in strength, valleys and master streams locate

Spence Air Photos

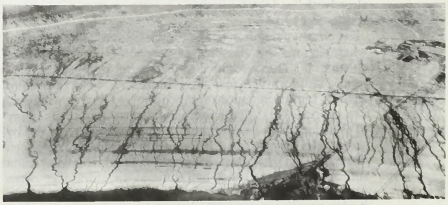


Figure 2: Dendritic drainage patterns in alluvium, Imperial Valley, California.

on weak outcrops in annular patterns. Centripetal patterns are produced where drainage converges on a single outlet or sink, as in some craters, eroded structural domes with weak cores, parts of some limestone country, and enclosed desert depressions. Trellis (or espalier) drainage patterns result from adjustment to tight regional folding in which the folds plunge. Denudation produces a zigzag pattern of outcrops, and adjustment to this pattern produces a stream net in which the trunks are aligned on weak rocks exposed along fold axes and small feeder streams run down the sides of ridges cut on the stronger formations.

Deranged patterns

Deranged patterns, in which channels are interrupted by lakes and swamps, characterize areas of modest relief from which continental ice has recently disappeared. These patterns may be developed either on the irregular surface of a till sheet (heterogeneous glacial deposit) or on the ice-scoured expanse of a planated crystalline block. Where a till sheet has been molded into drumlins (inverted-spoon-shaped forms that have been molded by moving ice), the postglacial drainage can approach a rectangular pattern. In glaciated highland, postglacial streams can pass anomalously through gaps if the divides have been breached by ice, and sheet glaciation of lowland country necessarily involves major derangement of river networks near the ice front. At the other climatic extreme, organized networks in dry climates can be deranged by desiccation, which breaks down the existing continuity of a net. The largely linear systems of ephemeral lakes in inland Western Australia have been referred to this process.

Adjustment to bedrock structure can be lost if earth movement raises folds or moves faults across drainage lines without actually diverting them; streams that maintain their courses across the new structures are called antecedent. Adjustment is lost on a regional scale when the drainage cuts down through an unconformity into an under-mass with structures differing greatly from those of the cover: the drainage then becomes superimposed. Where the cover is simple in structure and provides a regional slope for trunk drainage, remnants of the original pattern may persist long after superimposition and the total destruction of the cover, providing the means to reconstruct the earlier network.

HORTON'S LAWS OF DRAINAGE COMPOSITION

Great advances in the analysis of drainage nets were made by Robert E. Horton, an American hydraulic engineer who developed the fundamental concept of stream order:

Orders of streams

An unbranched headstream is designated as a first-order stream. Two unbranched headstreams unite to form a second-order stream; two second-order streams unite to form a third-order stream, and so on. Regardless of the entry of first- and second-order tributaries, a third-order stream will not pass into the fourth order until it is joined by another third-order confluent. Stream number is the total number of streams of a given order for a given drainage basin. The bifurcation ratio is the ratio of the number of streams in a given order to the number in the next higher order. By definition, the value of this ratio cannot fall below 2.0, but it can rise higher, since streams greater than first order can receive low-order tributaries without being promoted up the hierarchy. Some estimates for large continental extents give bifurcation ratios of 4.0 or more (see below *Sediment yield and sediment load*).

Although the number system given here, and nowadays in common use, differs from Horton's original in the treatment of trunk streams, Horton's laws of drainage composition still hold, namely:

1. Law of stream numbers: the numbers of streams of different orders in a given drainage basin tend closely to approximate an inverse geometric series in which the first term is unity and the ratio is the bifurcation ratio.

2. Law of stream lengths: the average lengths of streams of each of the different orders in a drainage basin tend closely to approximate a direct geometric series in which the first term is the average length of streams of the first order.

These laws are readily illustrated by plots of number and average length (on logarithmic scales) against order (on an arithmetic scale). The plotted points lie on, or close to, straight lines. The orderly relationships thus indicated are independent of network pattern. They demonstrate exponential relationships. Horton also concluded that stream slopes, expressed as tangents, decrease exponentially with increase in stream order. The systematic relationships identified by Horton are independent of network pattern: they greatly facilitate comparative studies, such as those of the influences of lithology and climate. Horton's successors have extended analysis through a wide range of basin geometry, showing that stream width, mean discharge, and length of main stem can also be expressed as exponential functions of order, and drainage area and channel slope as power functions. Slope and discharge can in turn be expressed as power functions of width and drainage area, respectively. The exponential relationships expressed by network morphometry are particular examples of the working of fundamental growth laws. In this respect, they relate drainage-net analysis to network analysis and topology in general.

MORPHOMETRY OF DRAINAGE NETWORKS

Relation of morphometric parameters and river flow. The functional relationships among various network characteristics, including the relationships between discharge on the one hand and drainage area, channel width, and length of main stem on the other, encourage the continued exploration of streamflow in relation to basin geometry. Attention has concentrated especially on peak flows, the forecasting of which is of practical importance; and since many basins are gaged either poorly or not at all, it would be advantageous to devise means of prediction that, while independent of gaging records, are yet accurate enough to be useful.

Problem of predicting peak discharge

A general equation for discharge maxima states that peak discharges are (or tend to be) power functions of drainage area. Such a relationship holds good for maximum discharges of record, but conflicting results have been obtained by empirical studies of stream order, stream length, drainage density, basin size, basin shape, stream and basin slope, aspect, and relative and absolute height in relation to individual peak discharges in the shorter term. One reason is that not all these parameters have always been dealt with. In any event, peak discharge is also affected by channel characteristics, vegetation, land use, and lags induced by interception, detention, evaporation, infiltration, and storage. Although frequency–intensity–duration characteristics (and, in consequence, magnitude characteristics) of single storms have been determined for considerable land areas, the distribution of a given storm is unlikely to fit the location of a given drainage basin. In addition, the peak flow produced by a particular storm is much affected by antecedent conditions, seasonal and shorter term wetting and drying of the soil considerably influencing infiltration and overland flow. Nevertheless, one large study attained considerable success by considering rainfall intensity for a given duration and frequency, plus basin area, and main-channel slope expressed as the height–distance relationship of points 85 and 10 percent of stem length above the station for which predictions were made. For practical purposes, the telemetering of rainfall in a catchment, combined with the empirical determination of its response characteristics, appears effective in forecasting individual peak flows.

Evolution of drainage systems. To empirical analysis of the morphometry of drainage networks has been added theoretical inquiry. Network plan geometry is specifically a form of topological mathematics. Horton's two fundamental laws of drainage composition are instances of growth laws. They are witnessed in operation, especially when a new drainage network is developing; and, at the same time, probability statistics can be used to describe the array of events and forms produced.

Random-walk analysis

Random-walk plotting, which involves the use of random numbers to lay out paths from a starting point, can produce networks that respond to analysis as do natural stream networks—*i.e.*, length and number increase and decrease respectively, in exponential relationship to order, and length can be expressed as a power function of area. The exponential relationship between number and

order signifies a constant bifurcation ratio throughout the network. A greater constancy in this respect would be expected from a randomly predicted network than from a natural network containing adventitious streams that join trunks of higher than one additional order. The exponential relationship between length and order in a random network follows from the assumption that the total area considered is drained to, and by, channels; the power relationship of length to area then also follows. The implication of the random-walk prediction of networks that obey the empirically derived laws of drainage composition is that natural networks correspond to, or closely approximate, the most probable states.

Geometry of river systems

HYDRAULIC GEOMETRY

Hydraulic geometry deals with variation in channel characteristics in relation to variations in discharge. Two sets of variations take place: variations at a particular cross section (at-a-station) and variations along the length of the stream (downstream variations). Characteristics responsive to analysis by hydraulic geometry include width (water-surface width), depth (mean water depth), velocity (mean velocity through the cross section), sediment (usually concentration or transport, or both, of suspended sediment), downstream slope, and channel friction.

Graphs of the values of channel characteristics against values of discharge usually display some scatter or departure from lines of best fit. One main cause is that values on a rising flood often differ from those on a falling flood, partly because of the reduction of flow resistance, and hence the increase in velocity, as sediment-concentration increases on the rising flood. Bed scour and bed fill are also related. Nevertheless, the variations for a given cross section can be expressed as functions of discharge, Q. For instance, width, depth, and velocity are related to discharge by the expressions: $w \propto Q^b$, $d \propto Q^f$, and $v \propto Q^m$, where w, d, v and b, f, m are numerical constants. The sum of the exponents $b+f+m=1$, because of the basic relation—namely that $Q = wdv$.

Width, depth, and velocity of flow

Similar functions can be derived for downstream variations, but, for downstream comparisons to be possible, the observed values of discharge and of channel characteristics must be referred to selected frequencies of discharge. When data are plotted on graphs with logarithmic scales for each of two discharge frequencies at an upstream and a downstream station, the four points for each channel characteristic define a parallelogram (Figure 3), whereby the hydraulic geometry of the stream is defined in respect of that characteristic. The values of exponents in the power equations differ considerably from one river to another: those shown here are theoretical optimum values. One common cause of difference is that many gaging stations are located where some channel characteristics are controlled, whether naturally as by rock outcrops or artificially as by bridge abutments. Constraints on variation in width, for instance, are mainly offset by increased variation in depth.

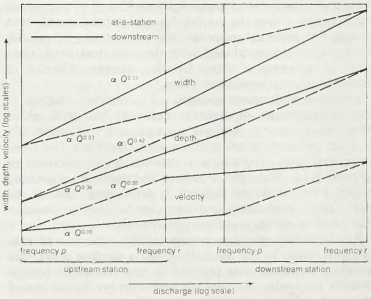


Figure 3: Hydraulic geometry for two frequencies of flow at an upstream and a downstream station, showing variation of width, depth, and velocity with discharge (see text).

Analyses of downstream variation in channel slope with discharge commonly reveal contrasts between field results and the theoretical optima. The discrepancy is probably due in considerable part to the fact that channel slope can vary in concert with channel efficiency, including channel habit, channel size, and channel form. Many past discussions of stream slope are invalidated by their restriction to the two dimensions of height and distance. In any event, the slopes of many natural channels are influenced by some combination of earth movement, change in base-level, glacial erosion, glacial deposition, and change of discharge and load characteristics that result from change of climate. Consequently, although natural profiles from stream source to stream mouth suggest a tendency toward a smooth concave-upward form, many actually are irregular. Even without a change of baselevel, degradational tendency, or discharge, a change in channel sinuosity can produce a significant change of channel slope.

A marked downstream lessening of slope does not imply a decrease in velocity at a given frequency of discharge; reduction of slope is accompanied, and offset, by an increase in channel efficiency due mainly to an increase in size. The lower Amazon, with a slope of less than 7.6 centimetres per 1.6 kilometres, flows faster at the bankfull stage than many mountain streams, at 2.4 metres per second. According to the assumptions made, an optimal velocity equation in hydraulic geometry can predict a slight increase, constancy, or a slight decrease in velocity downstream, for a given frequency of discharge. On the Mississippi, velocity at mean discharge (not a set frequency) increases downstream; velocity at the overbank stages of the five-year and 50-year floods is constant downstream. Constant downstream velocity may well be first attained at the bank-full stage. The fact that relationships are highly disturbed at and near waterfalls and other major breaks of slope (the Paraná just below the site of the former Guaíra Falls, for instance, ran at nine to 14 metres per second) has no bearing on the principles of hydraulic geometry, which apply essentially to streams in adjustable channels.

Downstream changes of velocity

The interrelationships and adjustments among width, depth, width–depth ratio, suspended-sediment concentration, sediment transport, deposition, eddy viscosity, bed roughness, bank roughness, channel roughness, and channel slope in their relation to discharge, both at-a-station and in the downstream direction, plus the tendency at many sections on many streams for variation to occur about some modal value, all encourage the conception of rivers as equilibrium systems. The designation quasi-equilibrium systems is usually used, since not all variances can be simultaneously minimized, and minimization of some variances (*e.g.*, of water-surface slope) can only be secured at the expense of maximizing others (*e.g.*, channel depth).

RIVER CHANNEL PATTERNS

Distinctive patterns in the plan geometry of streams correspond to distinctive combinations of cross-sectional form, calibre of bed load, downstream slope, and in some cases cross-valley slope, tendency to cut or fill, or position within the system. The full range of pattern has not been identified: it includes straight, meandering, braided, reticulate, anabranching, distributary, and irregular patterns (Figure 4). Although individual patterns are given separate names, the total range constitutes a continuum.

Basic types

Straight channels. Straight channels, mainly unstable, develop along the lines of faults and master joints, on steep slopes where rills closely follow the surface gradient, and in some delta outlets. Flume experiments show that straight channels of uniform cross section rapidly develop pool-and-riffle sequences. Pools are spaced at about five bed widths. Lateral shift of alternate pools toward alternate sides produces sinuous channels, and spacing of pools on each side of the channel is thus five to seven bed widths. This relation holds in natural meandering streams.

Meandering channels. Meandering channels are single channels that are sinuous in plan (Figure 5, left), but there is no criterion, except an arbitrary one, of the degree of

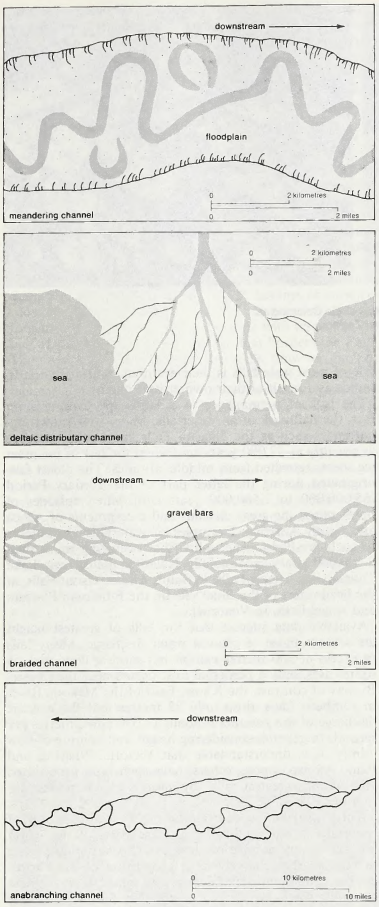


Figure 4: Common types of channel patterns.

sinuousity required before a channel is called meandering. The spacing of bends is controlled by flow resistance, which reaches a minimum when the radius of the bend is between two and three times the width of the bed. Accordingly, meander wavelength, the distance between two successive bends on the same side—or four-bend radii—tends to concentrate between eight and 12 bed widths, although variation both within and beyond this range seems to be related to variations in the cross-sectional form of the channel. Because bed width is related to discharge, meander wavelength also is related to discharge.

Meandering channels are equilibrium features that represent the most probable channel plan geometry, where single channels deviate from straightness. This deviation, and channel division in general, is related in part to the cohesiveness of channel banks and the abundance and bulk of midstream bars. When single channels are maintained, however, the meandering form is most efficient because it minimizes variance in water-surface slope, in angle of deflection of the current, and in the work done by the river in turning. This least-work property of meander bends is readily illustrated by the trace, identical with that of stream meanders, adopted by a bent band of spring steel. Meander plan geometry is simply describable by a sine function of the relative distance along the channel bend. The least-work and minimum-variance properties of the plan geometry, however, are secured only at the expense of maximizing the variance in depth. The longitudinal profile of the bed of a meandering stream includes pools at (or slightly downstream of) the extremities of bends and riffles at the inflections between bends. Increased tightness of bend, expressed by reduction in radius and increase in total angle of deflection, is accompanied by increased depth of pool. Where riffles are built of fragments larger than sand size, they behave as kinematic waves—*i.e.*, the speed of transport of material through a given riffle decreases as the spacing of surface fragments decreases, and the total rate of transport attains a maximum where the spacing is about two particle lengths. Numerous sand-bed streams in dry regions, however, fail to develop pool-and-riffle sequences, maintaining approximately uniform cross sections even at channel bends.

Effect of riffles

Irregularities in meanders developed in alluvium relate primarily to uneven resistance, which is often a function of varying grain size. Variations in total sinuosity are probably due in the main to adjustments of channel slope. The process of cutoff (short-circuiting of individual meanders) is favoured not only by the erosion of outer channel banks and by the tendency of meander trains to sweep down the valley but also by the stacking of meanders upstream of obstacles and by increases of sinuosity that accompany slope reduction.

Meandering streams that cut deeply into bedrock form entrenched meanders, the terminology of which is highly confused. It seems probable that, in actuality, the sole existing type of entrenched meanders is the ingrown type, where undercut slopes (river cliffs) on the outsides of bends oppose slip-off slopes (meander lobes) on the insides. For reasons not yet understood, lateral enlargement of ingrown meanders seems habitually to outpace downstream sweep, although the trimming of the upstream sides of lobes, and occasional cutoff, are well known. Many existing trains of ingrown meanders belong to valleys rather than to streams, relating to the traces of former rivers of greater discharge. Reconstruction of the original traces indicates approximate straightness at plateau level, as opposed to the inheritance of the ingrown loops from some former high-level floodplain.

In a broader context, meander phenomena cannot be understood as requiring cohesive banks of the kind usual in rivers. Meanders, with geometry comparable to that of rivers, have been recognized in the oceanic Gulf Stream and in the jet streams of the upper atmosphere. In this way, stream meanders are classed with wave phenomena in general.

Braided channels. Braided channels are subdivided at low-water stages by multiple midstream bars of sand or gravel (Figure 4 and Figure 5, right). At high water, many or all bars are submerged, although continuous downcutting or fixation by plants, or both, plus the trapping of sediment may enable some bars to remain above water. A single meandering channel may convert to braiding where one or more bars are constructed, as downstream of a tight bend where coarse material is brought up from the pool bottom. Each of the subdivided channels is less efficient, being smaller than the original single channel. If its inefficiency is compensated by an increase in slope (*i.e.*, by downcutting), the bar dries out and becomes vegetated and stabilized. However, many rivers that are largely or wholly braided along their length owe their condition to something more than local accidents. The braided condition involves weak banks, a very high width–depth ratio, powerful shear on the streambed (implied by the width–depth ratio), and mobile bed material. Thus, braided streams are typically encountered near the edges of land ice, where valleys are being filled with incoherent coarse sediment, and also on outwash plains, as the Canterbury Plains of South Island, New Zealand; width–depth ratios can exceed 1,000:1. Studies on terraced outwash plains

Sand and gravel bars



Figure 5: *River channel patterns.*
(Left) Depositional bars and oxbow cutoffs shown by the meandering reach of the Colorado River. (Right) Anastomosing (braided) reach of the Waimakariri River, New Zealand.
(Left) Spence Air Photos, (right) V.C. Browne

demonstrate that braided streams can readily excavate their valley floors—in other words, they are by no means solely an inevitable response to valley filling.

Distributary patterns

Distributary patterns, whether on alluvial fans or deltas (Figure 4), pose few problems. A delta pass that lengthens is liable to lateral breaching, whereas continued deposition, on deltas and on fans, raises the channel bed and promotes sideways spill down the least gradient. The branching rivers of inland eastern Australia, flowing across basin fills that range from thin sedimentary plains to thick fluvial accumulations, have affinities with deltaic distributaries even though their patterns are only radial in part. A branch may run for tens of kilometres before joining a trunk stream, whether its own or another. (G.H.D./Ed.)

WATERFALLS

Cascades and rapids

Waterfalls, sometimes called cataracts, arise from an abrupt steepening of a river channel that causes the flow of water to drop vertically, or nearly so. Waterfalls of small height and lesser steepness are called cascades; the term is often applied to a series of small falls along a river. Still gentler reaches of rivers that nonetheless exhibit turbulent flow and white water in response to a local increase in channel gradient are rapids.

Waterfalls are characterized by great erosive power. The rapidity of erosion depends on the height of a given waterfall, its volume of flow, the type and structure of the rocks involved, and other factors. In some cases the site of the waterfall migrates upstream by headward erosion of the cliff or scarp, whereas in others erosion tends to act downward to bevel the entire reach of river containing the falls. With the passage of time, by either or both of these means, the inescapable tendency of streams is to eliminate so gross a discordance of longitudinal profile as a waterfall. The energy of all rivers is directed toward the achievement of a relatively smooth, concave-upward, longitudinal profile; this is a common equilibrium, or adjusted condition, in nature.

Even in the absence of entrained rock debris that serves as an erosive tool of rivers, it is intuitively obvious that the energy available for erosion at the base of a waterfall is great. Indeed, one of the characteristic features associated with waterfalls of any great magnitude—with respect to volume of flow as well as to height—is the presence of a plunge pool, a basin that is scoured out of the river channel directly beneath the falling water. In some instances the depth of a plunge pool may nearly equal the height of the cliff causing the falls. Its depth depends not only on the erosive power of the falls, however, but also on the amount of time during which the falls remain at a particular place. The channel of the Niagara River below Horseshoe Falls, for example, contains a series of plunge pools, each of which represents a stillstand, or period of temporary stability, during the general upriver migration of the waterfall. The significance of this profile will be discussed below, but in general it may be said that the fate of most waterfalls is their eventual transformation to rapids as a result of their own erosive energy.

The transience of waterfalls

The lack of permanence as a landscape feature is, in fact, the hallmark of all waterfalls. Many well-known occurrences such as the Niagara Falls came into existence as recently as 12,000 years ago, when the last of the great ice sheets retreated from middle latitudes. The oldest falls originated during the latter part of the Tertiary Period (65,000,000 to 2,500,000 years ago), when episodes of uplift raised the great plateaus and escarpments of Africa and South America. Examples of waterfalls attributable to such pre-Pleistocene uplift (that occurring more than 2,500,000 years ago) include Kalambo Falls, near Lake Tanganyika; Tugela Falls, in South Africa; Tisisat Falls, at the headwaters of the Blue Nile on the Ethiopian Plateau; and Angel Falls, in Venezuela.

Height versus water discharge

Available data suggest that the falls of greatest height are seldom those of greatest water discharge. Many falls in excess of 300 metres exhibit but modest flow, and, in some cases, only a perpetual mist occurs near their bases. By way of contrast, the Khone Falls of the Mekong River in southern Laos, drop only 22 metres, but the average discharge of this cataract is about 11,330 cubic metres per second. In general, considering height and volume of flow jointly, it is understandable that Victoria, Niagara, and Paulo Afonso, among others, have each been proclaimed "the world's greatest falls" by various explorers and authorities (see Table 2).

World distribution of waterfalls. The distribution of waterfalls is not uniform, and large parts of the world are free of any notable occurrence. This is not surprising in view of the relatively large proportion of the Earth's land area that consists of deserts and semiarid areas; these are understandably devoid of modern falls on climatic grounds. Ice-covered polar regions and relatively unbroken, low-lying plains and plateaus also are unfavourable sites of development.

Considered on a global basis, waterfalls tend to occur in three principal kinds of areas: (1) along the margins of high plateaus or the great fractures that dissect them; (2) along fall lines, which mark a zone between resistant crystalline rocks of continental interiors and weaker sedimentary formations of coastal regions; and (3) in high mountain areas, particularly those that were subjected to glaciation in the recent past.

High plateaus. Notable falls along high plateaus include the world's highest, Angel Falls of the Churún River, Venezuela, with a drop of 979 metres and overall relief of more than 1,100 metres (Figure 6, right); Tugela Falls, issuing from the Great Escarpment, South Africa, which is 948 metres in height; Victoria Falls (108 metres) on the Zimbabwe–Zambia border; and Kalambo Falls (427 metres) on the Tanzania–Zambia border. The volume of flow at Victoria Falls is relatively large, approximately 1,080 cubic metres per second, but Guaíra Falls, a series of falls that until their submergence by the waters of

Itaipú Dam in 1982 totaled 114 metres along the Paraná River, Brazil–Paraguay, had the largest known average discharge—13,300 cubic metres per second. During flood stages, however, even this figure is exceeded at some falls along the Orange River and elsewhere. Angel Falls, Iguaçu Falls (82 metres), in Brazil, and several others occur along the margins of high plateaus, east of the Andes, between Venezuela and Argentina.

Fall lines. Waterfalls that occur along fall lines are in some cases relatively indistinguishable from plateau examples—the Aughrabies Falls (146 metres), for instance, which occur where the Orange River leaves resistant crystalline rocks of the plateau in southern Africa. The typical fall-line example, however, occurs at the junction of the crystalline rocks of the Appalachian Mountains and the sedimentary coastal plain along the eastern United States. A number of major cities, including Philadelphia, Baltimore, and Washington, D.C., are a geographic consequence of the existence of falls along this line or zone because they present barriers to further inland navigation. In England there is an analogous example with respect to the line of towns including Cambridge that borders the Fens. The most spectacular fall-line waterfalls, however, include Churchill (formerly Grand) Falls, Labrador, Canada (75 metres); Jog Falls (Gersoppa Falls), Karnātaka, India (253 metres [Figure 6, left]); and Paulo Afonso Falls, Brazil (84 metres).

Influence on establishment of cities

Glaciated mountains. The last category, mountainous and formerly glaciated regions, include such well-known waterfalls as Yosemite Falls, California (739 metres), with a three-section drop; Yellowstone Falls, Wyoming (94 metres [Figure 7, right]), with a two-section drop; Sutherland Falls, South Island, New Zealand (580 metres [Figure 7, left]); and Krimmler Waterfall, Austria (380 metres). Other falls of considerable height or volume of flow occur elsewhere in mountainous and formerly glaciated regions—namely, in the Alps, the Sierra Nevada and northern Rocky Mountains of North America, and South Island, New Zealand. The ice-free parts of Iceland and the fjord (drowned-valley) region of Norway also should be cited. Both areas contain numerous falls by reason of suitable topography and climate. Australia also has a few falls, notably the Wollomombi, in the Great Dividing Range, New South Wales (482 metres).

Types of waterfalls. The several types of waterfalls that occur in nature may be classified according to a variety of schemes. One of the simplest of these is based on principal region of occurrence—high plateaus, fall lines, and formerly glaciated mountains, as discussed above. More meaningful, however, is an alternate, threefold classification system that places more emphasis on the specific ways in which geologic and physiographic conditions produce and affect waterfalls. Thus, falls can be categorized as: (1) those attributable to natural discordance of river profiles, whether caused by faulting (vertical movements of the Earth's crust), glaciation, or other processes; (2) those attributable to differential erosion, which occurs whenever weak and resistant rocks are juxtaposed in some way; and (3) those attributable to constructional processes that create barriers and dams, over which water must fall. These three basic types will be discussed in turn.

Falls attributable to discordance of river profile. In one sense, all falls must be attributable to a discordance of river profile by their very definition. This category is here arbitrarily restricted, however, to exclude profile breaks that are caused by differential erosion and constructional processes. Remaining are waterfalls along fault scarps, uplifted plateaus and cliffs, glacial features of several kinds, karst topography—the caves and cave systems produced by solution of carbonate rocks—and falls that result from the issuance of springs from canyon walls high above valley floors.

The enormous rigid plates that make up the outer shell of the Earth continually move relative to one another, resulting in seafloor spreading, continental drift, and mountain building (see PLATE TECTONICS). These large-scale motions cause a buildup of strain within the rocks of the crust at some depth below the surface. Ultimately, the rocks must yield or shift in order to release this strain, and, when they suddenly do so, an earthquake results. Commonly, there will be some visible evidence of this sudden release at the Earth's surface, perhaps manifested by the creation of a cliff or series of cliffs along a line or zone. The sloping surfaces that form the cliff fronts are called fault scarps. The vertical movements that produce fault scarps seldom amount to more than about three metres during an individual earthquake. Repeated faulting along the same line or zone, however, can produce scarps that are thousands of metres in height in relatively brief periods of geologic time. Waterfalls occur where the faults cross established drainage systems. The ultimate height of such falls depends not only on the total height of uplift but also on the rate of downcutting by the affected rivers. Rates of uplift tend to exceed rates of downcutting considerably in those

Role of faulting

Shostal—EB Inc., (left) B. Bhansali, (right) Ray Manley



Figure 6: Waterfalls tend to occur in certain distinct regions. (Left) Jog Falls, Karnātaka (formerly Mysore) state, India. A waterfall associated with the fall line between resistant crystalline rocks and weaker sedimentary rocks. (Right) Angel Falls, Venezuela. The world's highest waterfall and an example of those associated with high plateaus throughout the world.

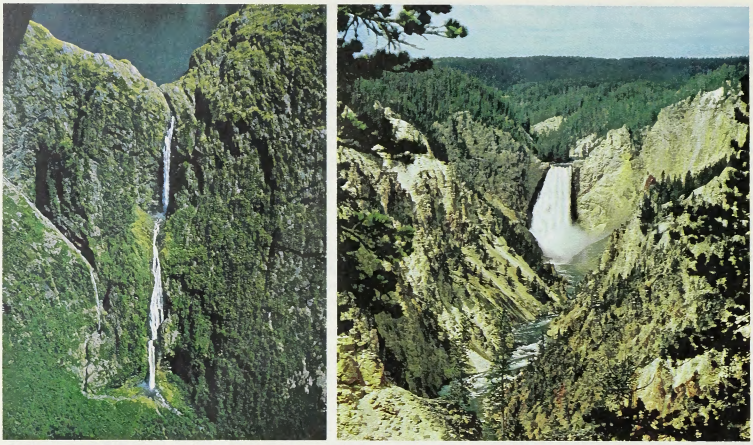


Figure 7: Waterfalls are formed by various geologic processes. (Left) Sutherland Falls, South Island, New Zealand. The falls issue from Lake Quill in a cirque produced by a former mountain glacier. (Right) Lower Yellowstone Falls, Yellowstone National Park, Wyoming. These are falls attributable to differential erosion in volcanic terrain; the Yellowstone River plunges over a near-vertical mass of more resistant rock.
(Left) Allyn Baum—Rapho/Photo Researchers, (right) Ray Manley—Shostal/EB Inc.

parts of the world where uplift is ongoing today. Hence, it is normal for high waterfalls to exist due to uplift in many areas. In addition, some plateaus are produced by broader, regional uplifts that are relatively continuous and are not associated with earthquakes. The heights attained are nevertheless comparable after suitable time intervals. Major rift (fracture) systems of continental or subcontinental scale, some sea cliffs, and other features of this nature also are attributable to some form of faulting. All of them provide suitable sites for waterfall development.

Relation of falls to glacial features

The processes of glaciation have served this same end. Mountain ranges that formerly were glaciated contain falls at the outlets of cirques, bowl-shaped depressions in the headwaters of drainage areas that were formed by the accumulation of ice and its erosive action on the underlying bedrock. In addition, waterfalls are most common where hanging valleys occur. Such valleys generally form when glacier ice deeply erodes a main or trunk valley, leaving tributary valleys literally hanging far above the main valley floor. After the glaciers have melted and withdrawn, streams from such tributary valleys must fall in order to join the main valley drainage system below. Hanging valleys also can occur in response to faulting and in some other non-glacial situations: the chalk cliffs of England, for example, where small streams cannot cut downward with sufficient rapidity to keep pace with backwearing of the cliffs by marine erosion.

Other features that may result from glaciation include glacial potholes and glacial steps. The former are thought to originate principally as a result of the plastic flow of ice at the base of a glacier; this permits the gouging of semicylindrical holes in the bedrock beneath the path of flow. The holes or depressions are subsequently enlarged and deepened by meltwater runoff that is heavily laden with gravels, and they have become the sites of modern cascades in many instances.

The steps (or glacial stairway, as this feature is sometimes called) consist of treads and risers on a relatively giant scale that have been produced by the passage of ice over bedrock, particularly when alternating rock properties or joints offer differential resistance to the flow of ice. Again, the establishment of runoff after wastage of the ice has occurred will lead to a series of waterfalls or cascades at the site of each riser in the stairway.

Most spectacular among glacial features, however, are the overdeepened valleys along formerly glaciated coasts, as in Norway. These fjords are intimately associated with falls because the valley walls typically are both high and steep and because hanging valleys are ubiquitous.

Like the potholes mentioned above, the solution of limestones and other carbonate rocks leads to the formation of pits, sinks, caves, and interconnected systems of caverns, which together are termed karst topography. Terrain of this kind commonly contains water in many of the included passages in the form of standing pools, streams, and, where discontinuities of cavern levels occur, waterfalls. There are a few parts of the world where karst topography and its associated drainage are prominent features of the landscape, but, on the whole, falls attributable to cave-forming processes are not numerous (see CONTINENTAL LANDFORMS: *Caves and karst landscape*). Springs that issue from canyon walls high above main valley floors are in the same category. Most of these artesian (free-flowing) systems result from the same type of solution phenomenon along joints and fractures that produce caves in carbonate rocks.

Resistance of rocks to erosion

Falls attributable to differential erosion. Rocks differ markedly with regard to their resistance to erosion by running water. Although no quantitative scales to express this difference have been developed, widespread agreement exists on certain generalities. Metamorphic rocks (those that are formed from preexisting rocks under the action of high temperatures and pressures), for example, are commonly more durable than are sedimentary rocks, and great differences can exist even among the latter because of a significant amount of variation in the degree of cementation and kinds of rock structure present in them. Thus, a quartz-rich sandstone whose constituent grains are cemented by silica tends to be much more resistant than a fissile shale, the clay-rich layers of which tend to split and separate. And the blocky character of some carbonate rocks (limestones and dolomites) and extrusive igneous rocks (formed by the cooling of lava flows) tends to enhance their resistance to fluvial erosion, notwithstanding their relatively low resistance to solution.

Regardless of the intrinsic toughness of any rock type, however, lengthy periods of weathering or the presence of intricate fracture patterns will render it easily erodible.

There are, in fact, a veritable legion of factors that influence rock resistance to erosion, and it is for this reason that generalities must be invoked. Suffice it to say that some rocks are weak whereas others are strong and that waterfalls are promoted where these occur in certain geologic arrangements.

There are three such arrangements that are common in nature: (1) horizontal or nearly horizontal strata in which rocks of greater resistance overlie weaker rocks, forming a protective cap rock; (2) inclined strata involving beds or layers of alternating resistance; and (3) various kinds of non-sedimentary rock arrangements in which dikes or veins of hard crystalline rocks are juxtaposed with weaker rocks. In each of these cases the weaker rocks are eroded more readily and more rapidly by running water, and the harder, resistant rocks, as a consequence, stand higher and are "falls makers." In the special case of the cap-rock arrangement, waterfalls migrate upriver because the protective upper layers break off as the weaker supporting strata are eroded from beneath. Niagara Falls is the most notable example involving sedimentary rocks (a blocky dolomite cap overlies a series of less-resistant shales and sandstones); more commonly, a lava flow caps erodible strata.

Effects of differential erosion

Falls attributable to constructional processes. There are four principal constructional processes that can lead to the creation of dams or barriers and, hence, to the formation of waterfalls. These processes are (1) precipitation of calcium carbonate from solution; (2) disruption of drainage by lava flows or the deposition of volcanic ash and other pyroclastic sediments; (3) ice damming and the construction of moraines, or ridgelike sedimentary deposits left at the sites of former glaciers; and (4) the deposition of landslide and avalanche debris.

The first of these, carbonate precipitation, can accumulate to considerable dimensions as spring deposits of travertine or calcareous tufa, often in a series of terraces. Where these ultimately block avenues of normal runoff,

Table 2: Selected Waterfalls of the World
(listed in declining order by height and by volume)

name	river	country	total height (metres)	height of greatest uninterrupted leap (metres)	average discharge by volume (cu m/sec)	number of falls C = cascade
Angel (Churún Merú)	Churún	Venezuela	979	807	...	2
Tugela	Tugela	South Africa	948	411	...	5
Mtarazi	Inyangombe	Zimbabwe	762	479	...	2
Yosemite	Yosemite	United States	739	436	...	3
Cuquenán	Cuquenán	Venezuela	610	317	...	...
Sutherland	Arthur	New Zealand	580	248	...	3
Kile	...	Norway	561	...	...	C
Kahiwa	...	United States	533	...	...	C
Mardal (Eastern)	Eikesdal	Norway	517	297	...	...
Ribbon	Ribbon	United States	491	491	...	...
King George VI	Utshi	Guyana	488	488	...	...
Wollomombi	Wollomombi	Australia	482	335	...	...
Mardal (Western)	Eikesdal	Norway	468	...	...	...
Kaliruwaa (Sacred)	Kalanui Stream	United States	463	80	...	C
Kalambo	Kalambo	Tanzania–Zambia	427	215	...	...
Gavarnie	Gave de Pau	France	422	...	...	C
Giessbach	Giessbach	Switzerland	391	...	...	...
Trümmelbach	Trümmelbach	Switzerland	391	...	...	...
Krimmler	Krimmler Ache	Austria	380	...	...	...
Vettis	Morkedola	Norway	371	...	...	...
Papalaua	Kawai Nui Stream	United States	366	...	...	...
Silver Strand	Merced	United States	357	...	...	C
Honokohau	Honokohau Stream	United States	341	...	...	C
Lofoi	Lofoi	Zaire	340	340	...	...
Serio	Serio	Italy	315	...	...	...
Barron	Barron	Australia	300	...	...	...
Belmore	Barrengarry Creek	Australia	300	...	...	3
Cannabullen	Cannabullen Creek	Australia	300	300	...	...
Horseshoe	Govetts Leap Creek	Australia	300	...	...	C
Wallaman	Stony Creek	Australia	300	...	...	...
Staubbach	Weisse Lutschine	Switzerland	290	290	...	...
Pungwe	Pungwe	Zimbabwe	277	277	...	...
Helena	Helena	New Zealand	271	...	...	1
Mollijus	Reisenelva	Norway	269	269	...	...
Austerkrok	Torrfjordelva	Norway	257	257	...	1
King Edward VIII	Semang	Guyana	256	...	...	...
Takakkaw	Yoho	Canada	254	...	...	...
Jog (Gersoppa)	Sharavati	India	253	253	...	1
Kaieteur	Potaro	Guyana	251	226	...	2
Waipio	Kekee Stream	United States	244	...	...	2
Tully	Tully	Australia	240	...	...	...
Feigum	Feigumelvi	Norway	218	...	...	...
Fairy	Fairy	United States	213	...	...	...
Fossa	Ullo	Norway	210	210	...	...
Feather	Fall	United States	195	...	...	...
Aurstapet	Aura	Norway	193	193	...	...
Maletsunyane (Semon Kong)	Maletsunyane	Lesotho	192	192	...	...
Sakaika	...	Guyana	192	140	...	...
Reichenbach	Reichenbach	Switzerland	190	91	...	...
Bridalveil	Bridalveil	United States	189	189	...	...
Khone	Mekong	Kampuchea–Laos	14	...	11,600	1
Niagara (Horseshoe)	Niagara	Canada–U.S.	49	...	5,525	...
Paulo Afonso	São Francisco	Brazil	84	...	2,800	3-C
Urubupungá	Paraná	Brazil	12	...	2,750	1
Iguaçu	Iguaçu–Paraná	Argentina–Brazil	82	...	1,750	C
Victoria	Zambezi	Zambia–Zimbabwe	108	108	1,080	1
Churchill (Grand)	Churchill (Hamilton)	Canada	75	...	990	...
Cauvery	Cauvery	India	98	...	935	...
Rhine	Rhine	Switzerland	24	...	700	C
Kaieteur	Potaro	Guyana	251	226	660	1
Detti	Jokulsá	Iceland	44	...	200	...

waterfalls result. The water in limestone caves also is rich in calcium carbonate, and where ponds occur in the path of small subterranean streams there is preferential precipitation at the spillage rims. The barriers that are raised are self-perpetuating, can attain heights of about 15 metres under certain circumstances, and have been called rimstone dams and falls.

Relationship of volcanic activity to waterfall development

Volcanic activity, principally in the form of basaltic lava flows, is related to waterfall development in many parts of the world. The flows compose the bulk of such great plateau areas as the Columbia River region of the United States and the Deccan Plateau in India and often serve as cap rock. The association of falls with plateaus in general and with cap-rock arrangements was noted previously, but, in addition, some falls result from drainage diversion and the ponding of streams and rivers by lava dams. This has occurred in some parts of New Zealand, Iceland, and Hawaii and, in general, in regions where volcanic activity is a prominent aspect of the landscape.

Ice dams can produce similar effects. One of the most interesting examples is Dry Falls, a "fossil waterfall" in the Columbia River Plateau, Washington, which formed in late Pleistocene time. A large ice sheet blocked and diverted the then-westward-flowing Columbia River and formed a vast glacial lake. The lake drained to the south when permitted to do so by periodically occurring ice dams, and torrents of water were released during these breakouts. The water flowed through the Grand Coulee channel and eroded a canyon nearly 300 metres deep. Dry Falls occurs along this flow path; it is about 120 metres high and five kilometres wide. The Columbia River has reestablished its path to the sea since the disappearance of the ice sheet, and so the falls are dry today.

The magnitudes of flow that must have occurred during the Pleistocene, however, can be appreciated from data on some of the great glacier outburst floods (*jøkulhlaups*) of modern history. The breaching of an ice dam at Grímsvötn, Ice., in 1922, for example, released about 7.1 cubic kilometres of water, and the discharge attained a value of 57,000 cubic metres per second.

There are other depositional features that may pond and dam streams, notably glacial moraines—which attain heights as great as 250 metres in the formerly glaciated valleys of the Alps—and landslides, avalanches, and other downslope movements of earth materials into valleys. The associated falls tend to be rather ephemeral, however, because all such unconsolidated material is cut through relatively swiftly, and smooth stream gradients are reestablished. The damming action of lava flows and glacier ice is far more important in nature; the lava flows consist of more durable material, and ice damming leads to outburst floods and great attendant erosion.

Development of waterfalls. With the passage of time a particular waterfall must either migrate upstream, as in the case of a cap-rock falls, or serve as the locus for general downcutting along the reach of river containing the falls. In either case, the process depends on the height of the falls, the volume of flow, and the nature and arrangement of the rocks involved. Any discussion of waterfall development requires knowledge of these three factors and, more importantly, knowledge of the former locations and configurations of any particular waterfall under consideration. If the changes of location through time are lacking, then rates of waterfall recession are basically indeterminate.

Rates of waterfall recession

The available data on the recession of the Horseshoe Falls of the Niagara River are little short of astonishing in comparison to the general paucity of such information elsewhere. Instrumental surveys of the configuration and position of Horseshoe Falls were made in 1842, 1875, 1886, 1890, 1905, 1927, and 1950. Still earlier delineations of position were provided by visual observations as long ago as 1678. For this reason, general waterfall development must be considered in terms of the Horseshoe Falls example. It should be noted, however, that the recession rates pertaining to this cap-rock-type falls are not necessarily average rates for all falls of this kind; they certainly do not apply to non-cap-rock falls in crystalline rocks, for example, where much slower rates generally prevail.

The average rate of recession of any falls can be determined from knowledge of the total upstream distance of migration and the time period during which the migration occurred. In the case of Horseshoe Falls, the total distance involved is about 12 kilometres, and retreat of the falls has been accomplished in approximately 12,500 years, since the disappearance of the most recent ice sheet from the area. The average rate of recession is therefore about one metre per year. The several instrumental surveys, however, suggest that a rate of 1.2 metres per year occurred during the 1842–75 period and two metres per year during the 1875–1905 period.

By way of comparison, the average recession rate for the American Falls, which occur downstream and to one side of Horseshoe Falls because of branching by the Niagara River, is only 0.08 metre per year. And, in a comparable vein, upstream migration of the Gullfoss in Iceland during the last 10,000 years is estimated to have occurred at an average recession rate of 0.25 metre per year. This is, again, a far slower rate of falls recession than has occurred at Horseshoe Falls.

To some extent the various recession rates are related to differential resistance of the rocks to erosion. Indeed, the discrepancy between the 1842–75 and 1875–1905 rates for Horseshoe Falls have been attributed in the past not only to possible surveying errors but also to the relative abundance of joints (fractures) in different parts of the dolomite cap rock. One study of Horseshoe Falls suggests, however, that another factor is of still greater importance—namely, the configuration of the crest of the falls and the relative stability of differing kinds of configurations. (L.K.L./Ed.)

Streamflow and sediment yield

PEAK DISCHARGE AND FLOODING

Rapid variations of water-surface level in river channels through time, in combination with the occurrence from time to time of overbank flow in flat-bottomed valleys, have promoted intensive study of the discharge relationships and the probability characteristics of peak flow.

Stage-discharge relations

Stage (depth or height of flow) measurements treat water level; discharge measurements require determinations of velocity through the cross section. Although records of stage respond to frequency analysis, the analysis of magnitude and frequency is preferable wherever stage is affected by progressive scour or fill, and also where channels have been artificially embanked or enlarged or both. The velocity determinations needed to calculate discharge range from those obtained with portable Venturi flumes on very small streams, through observations with gaging staff or fixed Venturi flumes on streams of modest size, to soundings with current meters at intervals of width and depth at cross sections of large rivers. Frequent velocity observations on large rivers are impracticable. It is standard practice to establish a rating formula, expressed graphically by a rating curve. Such a curve relates height of water surface to the area of and velocity through the cross section and thus to discharge. Secular changes in rating occur where a stream tends progressively to raise or lower its bed elevation. Short-term changes are common where the bed is mobile and especially where the bed elevation–discharge relation, and thus the stage–discharge relation, differs between the rising and the falling limb of a single peak discharge curve. In such cases the rating curve describes a hysteresis loop. Rating curves for sand-bed streams can include discontinuities, chiefly during rising discharge, that relate to behavioral jumps on the part of the bed.

Floods in hydrology are any peak discharges, regardless of whether or not the valley floor (if present) is inundated. The time–discharge or time–stage characteristics of a given flood peak are graphed in the hydrograph, which tends to assume a set form for a given station in response to a given input of water. The peak flow produced by a single storm is superimposed on the base flow, the water already in the channel and being supplied from the groundwater reservoir. Rise to peak discharge is relatively swift and is absolutely swift in small basins and on torrents where the duration of the momentary peak is also short. On very large streams, by contrast, peak discharge can be sustained

for lengths of days. Recession from peak discharge is usually exponential. The form of the hydrograph for any one station is affected by characteristics of the channel and the drainage net, as well as by basin geometry, all of which can be taken as permanent in this context.

As noted above, flood-flow prediction that is based on permanent characteristics has hitherto achieved but partial success. Transient influences, also highly and at times overwhelmingly important, include the storage capacity of bedrock and soil, the interrelationships of infiltration, evaporation, and interception and detention (especially by vegetation), plus storm characteristics, which vary widely with respect to amount, duration, intensity, and location of rainfall with respect to the catchment.

Flood-frequency analysis

In the longer term, flood-frequency analysis based on recorded past events can nevertheless supply useful predictions of future probabilities and risks. Flood-frequency analysis deals with the incidence of peak discharges, whereas frequency analysis generally provides the statistical basis of hydraulic geometry. Percentage frequency analysis has been much used in engineering: here, the 1 percent and 90 percent discharges, for instance, are those that are equalled or exceeded 1 and 90 percent of time, respectively. General observations of the flashy character of floods in headwater streams, in contrast to the long durations of flood waves far downstream, combine with analytical studies to suggest, however, that percentage frequency is in some respects an unsuitable measure. Magnitude–frequency analysis, setting discharge against time, is directly applicable in studies of hydraulic geometry and flood-probability forecasting.

Regional graphs of magnitude–frequency can be developed, given adequate records, for floods of any desired frequency or magnitude. Predictions for great magnitudes and low frequencies, however, demand records longer than those usually available. Twelve years of record are needed to define the mean annual flood within 25 percent, with an expectation of correct results for 95 percent of time; and in general, a record should be at least twice as long as the greatest recurrence interval for which magnitude is desired.

Predictions of overbank flow, whether or not affected by artificial works, are relevant to floodplain risk and floodplain management. Notably in the conterminous United States, floodplain zoning is causing risks to be reduced by the withdrawal of installations from the most flood-liable portions of floodplains or risks to be totally accepted by occupiers.

In the long geomorphic term the transmission of sediment through floodplain storage systems and through stream channels seems to result mainly from the operation of processes of modest magnitude and high frequency. Specifically, analyses suggest that total sediment transport by rivers is normally affected by flows approximating bankfull over durations ranging down from 25 to 1 percent of total time. Infrequent discharges of great magnitude, which can be expected on grounds of the probabilities of precipitation, snowmelt, and streamflow, range widely in destructive effect. Severe flooding is normally accompanied by great loss of life and property damage, the mean annual floods along the Huang Ho themselves affecting some 29,800 square kilometres of floodplain, but geomorphic effects may be minimal, even with very large floods. The approximately 100-year floods of eastern England in the spring of 1947, fed by unusually great and deferred snowmelt, scarcely affected either channels or floodplains. The 1955 floods in Connecticut, fed by rains amounting to 58 centimetres in places, produced only spotty effects of erosion and deposition, even where floodplains were inundated to a depth of six metres. For a given valley, there could be a threshold of inundation, river velocity, and sediment load, beyond which drastic changes occur. This is suggested, for example, by the catastrophic alluviation of valleys in eastern Australia and New Zealand during the last 4,000 or 2,000 years.

Catastrophic events

Sudden catastrophes in historical and geomorphic records are related to special events, mainly nonrecurrent: the 1841 Indus flood, which destroyed an army; the Gohna Lake flood of 1894 on the Ganges; and the 1925 Gros Ventre flood in Wyoming, accompanied the breaching of natural landslide barriers. The Lake Issyk-Kul (Kyrgyzstan) flood of 1963, which caused widespread erosion and deposition, followed the overtopping of a landslide barrier by waves produced by a mudflow. The Vaiont Dam (Italy), although itself holding, was overtopped in 1963 by 91-metre-high waves raised by a landslide: the floods downstream took more than 2,500 lives in 15 minutes. On the Huang Ho the floods of 1887 took an estimated 900,000 lives. In late Pleistocene time the overtopping of an erodible natural dam by the then-existing Lake Bonneville eventually released nearly 1,666 cubic kilometres of water; the maximum discharge of about 280,000 cubic metres per second is comparable to the flow of the Amazon, but velocities were very high, perhaps ranging to 7.6 metres per second. The greatest flood peak so far identified is that of the ice-dammed Lake Missoula in Montana, which, on release, discharged 2,085 cubic kilometres of water at an estimated peak flow of 8,500,000 cubic metres per second. Iceland is notable for glacier bursts, which are nonrecurrent where they result from subglacial eruptions but recurrent where they involve the sudden failure of ice dams, as with Grímsvötn, which periodically releases 8.3 or more cubic kilometres of water in floods that peak at 57,000 cubic metres per second. Deposition by glacier-burst floods is illustrated by Iceland's Sandur plains.

Ice-dam flooding

Peak discharges that close the range between natural floods of great magnitude and low frequency on noncatastrophic streams and natural catastrophic floods of great magnitude and perceptible frequency include stormwater discharges from expanding urban areas. Because of the progressive spread of impermeable catchment and efficient runoff systems, such floods tend to increase both in frequency and in magnitude. (G.H.D./Ed.)

SEDIMENT YIELD AND SEDIMENT LOAD

All of the water that reaches a stream and its tributaries carries sediment eroded from the entire area drained by it. The total amount of erosional debris exported from such a drainage basin is its sediment yield. Sediment yield is generally expressed in two ways: either as a volume or as a weight—*i.e.*, as acre-feet (one-foot depth of material over one acre) or as tons. In order to adjust for the very different sizes of drainage basins, the yield frequently is expressed as a volume or weight per unit area of drainage basin—*e.g.*, as acre-feet per square mile or as tons per square mile or per square kilometre. The conversion between the two forms of expression is made by obtaining an average weight for the sediment and calculating the total weight from the measured volume of sediment. Further, sediment yield is usually measured during a period of years, and the results are thus expressed as an annual average.

The sediment delivered to and transported by a stream is its sediment load. This can be classified into three types, depending on sediment size and the competence of the river. The coarsest sediment, consisting of boulders and cobbles as well as sand, moves on or near the bed of the stream and is the bed load of the river. The finer particles, silts and clays, are carried in suspension by the turbulent action of flowing water; and these fine particles, which are moved long distances at the velocity of the flowing water, constitute the suspended load of the river. The remaining component of the total sediment load is the dissolved load, which is composed of chemical compounds taken into solution by the water moving on or in the soils of the drainage basin. These three types of sediment constitute the total sediment load of the stream and, of course, the sediment yield of the drainage basin.

Measurement of the load

The sediment load can be measured in different ways. Collection of water samples from a river and measurement of the sediment contained in each unit of water will, when sufficient samples have been taken and the water discharge from the system is known, permit calculation of annual sediment yield. Because sediment in a stream channel is transported in suspension, in solution, and as material rolling or moving very near the bed, the water samples will contain suspended and dissolved load and perhaps some bed load. Much of the bed load, however, cannot be sampled by existing techniques, as it moves too

near the bed of a stream. It is fortunate, therefore, that the greatest part of the total sediment load is in the form of suspended load.

When a dam is constructed, the sediment transported by a stream is deposited in the still waters of the reservoir. In this case, both bed load and suspended load are deposited, but the dissolved load eventually moves out with the water released from the reservoir. Frequent, precise surveys of the configuration of the reservoir provide data on the volume of sediment that accumulates in the reservoir (Figure 8). Water samples can be taken to provide data on the dissolved load transported into the reservoir; and when this quantity is added to the measurements of suspended and bed load, a reasonably accurate measure of sediment yield from the drainage basin above the reservoir can be obtained.

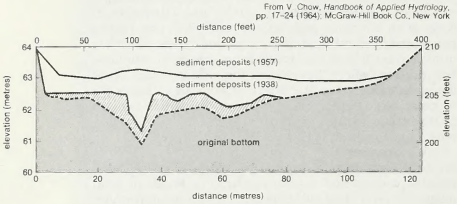


Figure 8: Cross section of sediment deposits, Lake Barcroft, Virginia.

In areas where information on sediment yield is required but the necessary samples have not been taken (perhaps because of the infrequent occurrence of flow in ephemeral streams), estimates of sediment yields may be obtained from measurements of hillslope and channel erosion within the basin or by the evaluation of erosion conditions. Certain characteristics of the drainage basin, such as the average slope of the basin or the number and spacing of drainage channels, may be used to provide an estimate of sediment yield (see below).

All of the techniques utilized to measure sediment yield are subject to considerable error, but data sufficiently accurate for the design of water-regulatory structures can be obtained by sampling or by reconnaissance surveys of the drainage systems.

Sources of sediment and nature of deposition. *Erosion in drainage basins.* The ultimate source of the sediment that is measured as sediment yield is the rock underlying the drainage basins. Until the rock is broken or weathered into fragments of a size that can be transported from the basin, the sediment yield will be low. The diverse mechanisms, both chemical and physical, that produce sediment and soil from rock are termed weathering processes. Depending on type of rock and type of weathering process, the result may be readily transported silts, clays, and sands or less easily transported cobbles and boulders.

Effects of weathering and breakdown of rocks

Most rocks have been fractured during the vicissitudes of geologic history, thereby permitting penetration of water and roots. Wedging by ice and growing roots produces blocks of rock that are then subject to further disintegration and decomposition by chemical and physical agencies. These rocks, if exposed on a hillslope, move slowly down the slope to the stream channel—the rate of movement depending on slope inclination; density of vegetation; frequency of freeze and thaw events; and the size, shape, and density of the materials involved. In addition, water moving through rocks and soil can dissolve soluble portions of rock or weathering products. This is especially important in limestone regions and in regions of warm, humid climate, where chemical decomposition of rocks is rapid and where the dissolved load of streams is at a maximum.

Sediment in storage

When sediment eroded from the hillslopes is not delivered directly to a channel, it may accumulate at the base of the slope to form a colluvial deposit. The sediment derived directly from the hillslope may be stored temporarily at the slope base; therefore, sediment once set in motion does not necessarily move directly through the stream system. It is more likely, in fact, that a given particle of sediment will be stored as colluvium before moving into the stream. Even then, it may be stored as alluvium in the floodplain, bed, or bank of the stream for some time before eventually moving out of the drainage system. Thus there is a steady export of sediment from a drainage basin, but an individual grain of sediment may be deposited and eroded many times before it leaves the system.

The preceding suggests that, over a period of time, the total erosion within a drainage basin is greater than the sediment yield of the system. Proof of this statement is the fact that the quantity of sediment per unit area that leaves a drainage system decreases as the size of the drainage basin increases. This is partly explained by the decrease in stream gradient and basin relief in a downstream direction. That is to say, much sediment is produced in the steeper areas near drainage divides, and sediment production decreases downstream. Moreover, the increasing width of valleys and floodplains downstream and the decreasing gradient of the streams provide an increasing number of opportunities for sediment to be deposited and temporarily stored within the system.

Factors related to quantity of sediment

Each of the components of the drainage system—hillslopes and channels—produces sediment. The quantity provided by each, however, will vary during the erosional development of the basin and during changes of the vegetational, climatic, and hydrologic character of the drainage system. Most rivers flow on the upper surface of an alluvial deposit, and considerable sediment is thus stored in most river valleys. During great floods or when floodplain vegetation does not stabilize this sediment, large quantities may be flushed from the system as the channel widens and deepens. At these times, the sediment produced by stream-channel erosion is far greater than that produced by the hillslopes, and sediment yields will be far in excess of rates of hillslope erosion. Such cycles of rapid channel erosion or gullying and subsequent healing and deposition are common in arid and semiarid regions.

Environments of deposition. It is clear that a great range of sediment sizes may be transported by a river. Sediment of small size (*e.g.*, suspended load), when set in motion by erosive agents, may be transported through a river system to the sea, where it may be deposited as a deep-sea clay. Most sedimentary particles, however, have a more eventful journey to their final resting place. (In a geologic context, this may be a temporary resting place; sediment, for example, when it reaches the coast, may be incorporated in a delta at the river mouth or be acted upon by tides, currents, and waves to become a beach deposit.)

Effects of climate on deposition

If sediment is moved downstream into a progressively more arid environment, the probability of deposition is high. Thousands of metres of alluvial-fan deposits flank the mountains of the western United States, the basin-and-range terrain of Iran and Pakistan, and similar desert regions (see below). In the arid climates of these areas the sediment cannot be moved far, because the transporting medium—water—diminishes in a downstream direction as it infiltrates into the dry alluvium. In extremely arid regions, wind action may be important: the transport of sand-size and smaller sediment by wind may be the only significant mechanism for the transport within and out of some drainage systems in deserts.

The impact of human activity on river flow has come to play a major role in determining the site of sediment deposition. The many dams that have been constructed for flood control, recreation, and power generation hold much of the sediment load of rivers in reservoirs. Furthermore, the contribution of sediment from the small upstream drainage systems has been decreased by the construction of stock-water reservoirs and various erosion-control techniques aimed at retaining both water and sediment in the headwater areas. Diversion of water for irrigation also decreases the supply of water available to transport sediment; and in many cases, the diversion actually moves sediment out of the streams into irrigation canals and back onto the land.

Factors that influence sediment yield. Of greatest concern to the human community are the factors that cause

rapid rates of erosion and high sediment yields. The quantity and type of sediment moving through a stream channel are intimately related to the geology, topographic character, climate, vegetational type and density, and land use within the drainage basin. The geologic and topographic variables are fixed, but short-term changes in climatic conditions, vegetation, and land use produce abrupt alterations in the intensity of erosion processes and in sediment yields.

The sediment yield from any drainage system is calculated by averaging the data collected over a period of years. It is, therefore, an average of the results of many different hydrologic events. The sediment yield for each storm or flood will vary, depending on the meteorologic character of the storm event and the resulting hydrologic character of the floods. High-intensity storms may produce sediment yields well above the norm, whereas an equal amount of precipitation occurring over a longer period of time may yield relatively little sediment. During short spans of time (days or years), sediment yields may fluctuate greatly because of natural or man-induced accidents (*e.g.*, floods and fires), but over longer periods of time, the average sediment yield will be typical of the geologic and climatic character of a region.

Short-term variations. An example of a short-term change in sediment yield is provided by data on the sediment transported by the Colorado River in Arizona for the years 1926–54 (Figure 9). It is evident that sediment yield varied widely from year to year. It is greatest for years of highest runoff, but for a given amount of runoff, the maximum sediment yield may be twice the minimum sediment yield. These variations reflect the frequency of storms and their duration and intensity during the years of record.

Another interesting aspect of the relation is that in each of the years after 1940 the annual sediment load at the Grand Canyon was 50,000,000 to 100,000,000 tons less than would be expected on the basis of the curve fitted to the data for the period 1926–40 (Figure 9). This major decrease in sediment yield reflects some significant change in the hydrology of the Colorado River drainage basin. A study of the precipitation patterns for the years 1926–54 suggests that the change in the sediment yield–runoff relation beginning in 1941 is the result of a drought in the southwestern United States. The high-sediment-producing, weak-rock areas of the Colorado plateaus were affected by the drought, but the low-sediment-producing, hard-rock areas of the Rocky Mountains were not. Thus, during the years 1941–50 the amount of water delivered from the main runoff-producing areas in Colorado, Wyoming, and northern Utah was normal. Runoff was much reduced from the high-sediment-producing areas in southern Utah and Arizona, however. The result was essentially normal runoff but greatly reduced sediment yield. From 1950 the drought encompassed the entire Colorado River Basin, and low runoff was recorded for the years 1950, 1951, 1953, and 1954; yet, the proportion of runoff produced by the high-sediment-producing areas remained low, as did the sediment yield.

Drought-induced variations

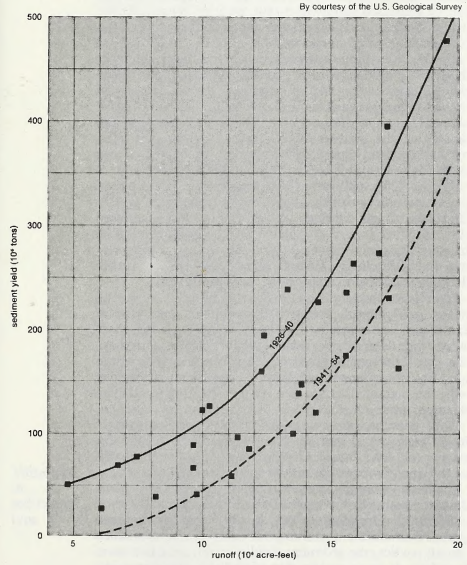


Figure 9: Relation of sediment yield and runoff for the 1926–40 and 1941–54 periods on the Colorado River at the Grand Canyon.

It can be expected that sediment-yield rates will fluctuate with climatic variations. It is possible, therefore, that an average value of sediment yield obtained for a short period of record may not provide a valid measure of the characteristic sediment yield that would be expected over a longer period of years.

A further example of short-term variation of sediment yield, in this case the result of human activity on the landscape, is provided by data illustrating the change from natural conditions to conditions produced by upland farming and from farming conditions to urban conditions in the Piedmont region of the eastern United States. Sediment yields for forested regions normally are about 37 tons per square kilometre (100 tons per square mile), and this was the case during the early part of the 19th century in this region. A significant increase in sediment yield occurred after 1820 as the land was occupied and farmed. During the period of intense farming, 1850–1930, the sediment yield reached almost 310 tons per square kilometre, but a decrease occurred between 1930 and 1960, as much land was permitted to revert to forest or grazing land. With the onset of construction and real estate development, however, vegetation was destroyed, and large quantities of sediment were eroded. The sediment yields for some small areas reached about 770 tons per square kilometre during urbanization, but with the paving of streets, completion of sewage systems, and planting of lawns, the sediment yields decreased markedly. This example demonstrates very clearly both the long-term and short-term effects of human activity on sediment yield rates.

Variations induced by man and urbanization

In any drainage basin, even one not affected unduly by human action, short periods of high sediment yield will alternate with periods of little export of sediment. Prime examples are small drainage basins in arid or semiarid regions, where sediment yield occurs only during and following precipitation. Runoff and sediment yield can be zero between storms but high during and immediately following precipitation.

Even temperature variations have been demonstrated to influence sediment transport and sediment yields. Cooler water is more viscous, and this decreases the fall velocity of sediment particles and enables the stream to transport a larger amount of sediment. Thus, the sediment load of the Colorado River is greater during winter months.

Seasonal and accidental variations

The disastrous effect of fire on sediment yields may be seen in the example of the conditions that followed a major storm and flood in the steep drainage basins of the San Gabriel and San Bernardino mountains of California in 1938. Maximum vegetational cover on these drainage basins is only 65 percent at best, and they are notoriously high sediment producers under the most favourable conditions. Sediment-yield rates were established for several drainage basins that had been subjected to fires as recently as one year before the storm and as long as 15 years before the storm. The results shown (Figure 10) further demonstrate the great effect of vegetational disturbance on sediment yields; for example, a drainage area with only 40 percent of the area burned had a 340 percent increase in sediment yield if the fire occurred one year before the storm. According to the information provided, the burned

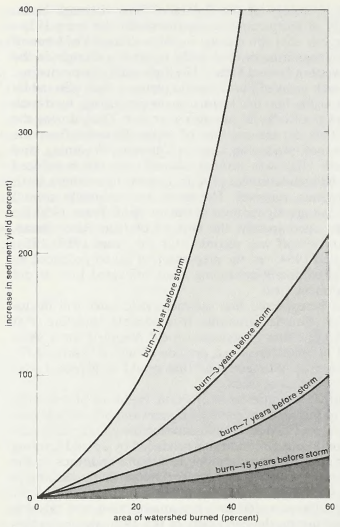


Figure 10: Increase in the sediment yields of southern California drainage basins following the 1938 storm. The several curves show the percentage increase in sediment yield in relation to the area burned and the date of burning (see text).

From Anderson and Trobitz, *Journal of Forestry*, vol. 47, p. 347 (1949)

area one year after the fire had a 10 percent vegetational cover. Obviously, a storm immediately following the fire would have had even more disastrous consequences. Three years after the burn, a 35 percent vegetation cover had been established on the burned area, and sediment yields decreased markedly to only twice the yield preceding the fire. After seven years a 45 percent cover had been established on the burned area, and sediment yields were only 50 percent greater than pre-burn values. After 15 years a 55 percent vegetal cover had been established, and sediment yields were almost normal. The decrease in sediment yield with increased plant cover is apparent. It is also obvious that an average value of sediment yield from a burned drainage basin for a 15-year period would be meaningless; with progressive reestablishment of vegetation, sediment-yield rates progressively decrease with time.

Long-term or average sediment yield. It has been estimated that modern sediment loads of the rivers draining to the Atlantic Ocean may be four to five times greater than the prehistoric rates because of the effects of human activity. Even where human impact is large, however, it is possible to recognize several other independent variables that exert a major influence on long-term sediment yield. These variables can be grouped into three main classes: geologic, geomorphic, and climatic-vegetational.

Influence of rock type and soils

The major geologic influence on sediment yield is through lithology, or the composition and physical properties of rocks and their resistance to weathering and erosion. An easily weathered and eroded shale, siltstone, or poorly cemented sandstone will provide relatively large quantities of sediment, whereas a lava flow, a well-cemented sandstone, or metamorphic and igneous rocks produce negligible quantities of transportable sediment. The highest known sediment yields that have been recorded are produced by the erosion of unconsolidated silts (loess). Loess is readily eroded, especially when the protecting vegetational cover is disturbed, as has happened in the high-sediment-producing areas of western Iowa in the United States and the Huang Ho Basin in China.

In general, sediment yield from drainage systems underlain by granitic rocks is from one-fourth to one-half that of drainage basins underlain by sedimentary rocks. There are exceptions. Limestone, which may be a massive rock, is highly resistant to erosion in arid regions, where mechanical or physical weathering is dominant. It is, however, highly susceptible to chemical weathering, especially solution, in humid regions. Most of the earth material removed from a limestone terrain will be transported as dissolved load, with some suspended load derived from erosion of the residual soil.

Another factor of importance in determining erosion rates is the permeability of earth materials. When soils are permeable, much of the water delivered to the surface infiltrates and does not produce surface runoff, thereby inhibiting surface erosion. This condition is characteristic of very sandy soils. On the other hand, when soil materials are of low permeability (*e.g.*, clayey soils), a greater part of the precipitation runs off on the surface, thereby causing greater erosion and higher sediment yields.

Most drainage areas are composed of more than one rock type. In some areas the sedimentary rocks have been folded, and rocks of different resistance are exposed, with hard rocks forming ridges and mountains and weak rocks forming valleys. The erosional development of such a terrain is complex, and the sediment produced by a drainage basin of this kind will reflect the complex geologic situation, the greater part of the sediment yield being derived from the areas underlain by the rocks that are most susceptible to erosion.

Geomorphic variables

The character of the topography of a drainage basin significantly influences the quantity and type of runoff and sediment yield. The steeper a slope, the greater is the gravitational force acting to remove earth materials from the slope. In fact, the rate of movement of rocks and soil particles is directly related to the sine of the angle of slope inclination.

Steep slopes are readily eroded, and it follows that drainage basins with a great range of relief or steep average slope will produce not only higher sediment yields but coarser sediment. The average slope of a drainage basin can be expressed simply as a ratio of basin relief to basin length. Sediment yields increase exponentially with an increase in this relief–length ratio.

Another important characteristic of a drainage system is the spacing and distribution of drainage channels within the drainage basin. This is referred to as the texture of the topography, and it can be described by a ratio of total channel length to drainage area. This ratio is the drainage density of the system. High drainage density indicates numerous, closely spaced channels that provide an escape route for both runoff and its entrained sediment load.

When relief–length ratio (r), expressing the role of gravity, is combined with drainage density (d), expressing the efficiency of the drainage system, this yields a texture-slope product (rd), a parameter that describes the gross morphology of a drainage system. Hence it is not surprising that it is closely related to sediment yield of small drainage basins of similar geology and land use.

The relation between the texture-slope product and sediment yield is such that a high sediment yield can be expected from basins with a large drainage density and steep slope. For basins with similar relief–length ratios, those with the highest drainage density produce the greater quantity of sediment. In general, however, the basins with the highest drainage density are also those with the steepest slope.

Many geomorphic characteristics can be related to sediment yield, but it can be stated with assurance that the steeper and the better drained the system, the greater will be the quantity of sediment produced per unit area.

The relationship may be used to estimate yield from other drainage basins in the region from which these data were obtained. Similar relations may be developed for other regions when sufficient geomorphic data become available.

In all studies, the sediment yield per unit area has been found to decrease as the size of the drainage basin increases. This reflects the previously discussed downstream decrease in gradient and slope and the increase in area

available for temporary storage of sediment. Therefore total sediment yield per unit area invariably is related inversely to drainage area. Several other factors are involved, of course, but the largest drainage basins do not produce the largest quantity of sediment per unit area of drainage basin.

The morphology of a drainage basin is significantly related to sediment yield, but scientists have not yet done sufficient research to enable the prediction of sediment yields from drainage basins in the diverse regions of the world.

Relationship between climatic and vegetational influences

It is difficult to separate the influences of climate and vegetation on erosion and sediment yield, because the primary effect of climate on sediment yield is determined by the interaction between vegetation and runoff. This effect is displayed by the contrast between the dissolved load and the suspended load and bed load transported by streams. Dissolved load increases from a negligible amount in arid regions to 60 tons per square kilometre in humid regions, where chemical weathering and groundwater contribution to river flow is greatest. The dense vegetational cover of humid regions retards runoff and aids infiltration, thereby enhancing the effects of chemical decomposition of the rocks and soils to produce soluble material. The available data also show a sharp increase in sediment yield (suspended and bed load) as precipitation increases from low to moderate amounts. In semiarid regions, however, the increase of vegetation density with increased precipitation exerts a significant influence on erosion; and sediment transport and sediment yield decrease as the climate becomes increasingly humid. This relationship can, of course, be significantly modified by human activities. As the previously mentioned effect of urbanization demonstrates, removal of vegetation from the land in humid regions greatly accelerates erosion, and it may increase sediment yield to the maximum expected in semiarid regions.

Average temperature also affects sediment yields. The hotter the climate, the more water is lost to evapotranspiration, and the critical zone where vegetation becomes dominant consequently shifts to areas of higher precipitation.

The effect of vegetational cover on sediment yield has been discussed previously for areas where fires have destroyed much of the cover and catastrophically increased erosion and export of sediment from the system (Figure 10). Additional data on the effect of vegetation on erosion rates reveal that with a 65 percent plant cover little erosion will occur, but as plant cover decreases erosion increases significantly.

Although erosion increases greatly with a decrease in plant cover between 20 and 15 percent, it cannot continue to increase at this high rate. At some point, the maximum rate of erosion of the soil will be achieved. At some value of low-plant-cover density, the influence of vegetation must be negligible; erosion then will be determined only by soil erodibility.

Although average precipitation significantly influences vegetation type and density and the sediment yield, it has been demonstrated that, for a given quantity of annual precipitation, sediment yields will be greatest where highly seasonal (*e.g.,* monsoonal) climatic conditions prevail. Precipitation, when concentrated during a few months of the year, produces large quantities of sediment because of the higher intensity of the precipitation events and the long dry season when vegetational cover is severely weakened by drought.

Climate also plays a role in determining the type of sediment produced by a drainage basin. A study of the type of sediment deposited on the inner continental shelf reveals that the type of sediment (mud, sand, or gravel) is indeed influenced by climate. Mud, for example, is most abundant off shores of high temperature and rainfall, where chemical weathering is important. Gravel is common off areas of both low temperature and rainfall, where mechanical weathering is dominant. Sand is found everywhere, but it is most abundant in areas of moderate climate and in arid areas. Average temperature also may be important where the annual temperature is below the freezing point. (S.A.S./Ed.)

Variations in sediment type

Rivers as agents of landscape evolution

Every landform at the Earth's surface reflects a particular accommodation between properties of the underlying geologic materials, the type of processes affecting those materials, and the amount of time the processes have been operating. Because landforms are the building blocks of regional landscapes, the character of the local surroundings is ultimately controlled by those factors of geology, process, and time—a conclusion reached in the late 19th century by the noted American geologist and geographer William Morris Davis. In some regions, severe climatic controls cause a particular process agent to become preeminent. Deserts, for example, are often subjected to severe wind action, and the resulting landscape consists of landforms that reflect the dominance of erosional or depositional processes accomplished by the wind. Other landscapes may be related to processes operating beneath the surface. Regions such as Japan or the Cascade Range in the northwestern part of the United States clearly have major topographic components that were produced by repeated volcanic activity. Nevertheless, rivers are by far the most important agents in molding landscapes because their ubiquity ensures that no region of the Earth can be totally devoid of landforms developed by fluvial processes.

The key role of rivers in molding landscapes

Rivers are much more than sluiceways that simply transport water and sediment. They also change a nondescript geologic setting into distinct topographic forms. This happens primarily because movement of sediment-laden water is capable of pronounced erosion, and when transporting energy decreases, landforms are created by the deposition of fluvial sediment. Some fluvial features are entirely erosional, and the form is clearly unrelated to the transportation and deposition of sediment. Other features may be entirely depositional. In these cases, topography is constructed of sediment that buries some underlying surface that existed prior to the introduction of the covering sediment. Realistically, many fluvial features result from some combination of both erosion and deposition, and the pure situations probably represent end members of a continuum of fluvial forms.

VALLEYS AND CANYONS

River valleys constitute a major portion of the natural surroundings. In rare cases, spectacular valleys are created by tectonic activity. The Jordan River and the Dead Sea, for example, occupy a valley that developed as a fault-bounded trough known as a rift valley. The distinct property of these and other tectonically controlled valleys is that the low topographic zone (valley) existed before the river. Notwithstanding tectonic exceptions, the overwhelming majority of valleys, including canyons and gorges, share a common genetic bond in that their characteristics are the result of river erosion—*i.e.,* rivers create the valleys in which they flow. In most cases, erosion was accomplished by the same river that occupies the valley bottom, although sometimes rivers are diverted from one valley into another by a process known as stream piracy, or stream capture. Piracy of a large river into another valley often creates a situation where the original expansive valley is later occupied by a river that is too small to have created such a large valley. The opposite case also may occur. The implication here is that valley size is directly related to river size, an observation that generally holds true. Exceptions to this rule arise because of capture events during the evolution of a valley and because valley morphology is strongly influenced by variations in the bedrock into which the valley is carved.

Stream piracy

A genuine bedrock valley is usually covered by valley-fill deposits that obscure the actual configuration of the valley floor. Therefore, little is known about valley morphology unless drill holes or geophysical techniques are employed to document the buried bedrock-alluvium contact. Where information is available, it suggests that the deepest part of most valleys is not directly beneath the river. Commonly, the influx of load at a tributary junction forces the river to the opposite side of the valley, a phenomenon demonstrated clearly in the upper Mississippi River Valley between St. Louis, Mo., and St. Paul, Minn.

Where a valley is devoid of thick deposits and is completely occupied by a river, the bedrock valley floor often develops an asymmetrical configuration such that the deepest part of the valley occurs on the inside of bends. This general rule is not inviolate because the position of incision depends on the amount of load entrained by the river. When sediment load is totally entrained and velocity is high, entrenchment will most likely occur on the inside of the bend. If deposition occurs or sediment cannot be entrained, however, incision will normally be on the outside of the bend. In straight reaches the deepest part of the valley floor is normally associated with an inner channel cut into bedrock. Its position is determined by where the river was at the time that it flowed at the level of the valley floor. Inner channels form as the culmination of a progressive change in erosional features during the initial phase of incision. Scour features gradually coalesce until a distinct channel appears that is able to contain the entire river flow. Inner channels are rarely seen except when exposed during excavation associated with dam construction. Where observed, such channels commonly have a narrow, deep gorgelike shape. For example, at the site of the Prineville Dam in the state of Oregon, the inner channel averages 21 metres wide and as much as 18 metres deep.

Valley evolution. The ultimate form assumed by any valley reflects events that occurred during its developmental history and the characteristics of the underlying geology. During initial valley development in areas well above regional baselevel, valley relief tends to increase as rivers expend most of their energy in vertical entrenchment. Valleys are generally narrow and deep, especially in areas where they are cut into unfractured rocks with lithologic properties that resist erosion (most igneous rocks, well-indurated sedimentary rocks such as quartzites, and high-rank, silica-rich metamorphic rocks). Abrupt changes in river and valley bottom gradients, such as knickpoints and waterfalls, are common in the initial developmental phase. As downcutting continues, however, rivers gradually smooth out the longitudinal profile of the valley floor. Eventually most, if not all, waterfalls are eliminated, and rivers reach an elevation close to their baselevel (see above). In this condition, more energy is expended laterally than vertically, and a river progressively broadens its valley floor. As a result, most river valleys change over time from narrow forms to broader ones, the shape at any time being dependent on baselevel, rock type, and rock structures.

Initial stage of valley development

In areas where pronounced macrostructures such as major folds or faults exist in the geologic framework, the position and character of valleys are controlled by those structures. For example, the folds in the Appalachian Mountains in the eastern United States exert a very strong control on the orientation and form of many valleys developed in the region.

Formation of canyons and gorges. The most spectacular valley forms are canyons and gorges that result from accelerated entrenchment prompted by recent tectonic activity, especially vertical uplift. Canyons and gorges are still in the initial phase of valley development. They range in size from narrow slits in resistant bedrock to enormous trenches (Figure 11). Where underlying bedrock is composed of flat-lying sedimentary rocks, regional uplift creates high-standing plateaus and simultaneously reinvigorates the erosive power of existing rivers, a phenomenon known as rejuvenation. Vertical entrenchment produces different valley styles depending on the size of the river and the magnitude and rate of uplift. The Grand Canyon of the Colorado River, located in the southwestern United States and formed in response to uplift of the Colorado Plateau, has entrenched about 1,800 metres and widened its walls six to 29 kilometres during the past 10,000,000 years. The Grand Canyon is only one of many spectacular canyons that developed in response to uplift of the Colorado Plateau. Uplift of the Allegheny Plateau in the eastern United States has led to the creation of the narrow, deep valleys that are so prominent in West Virginia and western Pennsylvania.

Rejuvenation

Canyons and gorges frequently develop across the trends of underlying macrostructures. In normal situations, valleys should follow the orientation of the major folds and faults; however, the geologic setting prior to uplift and the processes associated with tectonic activity permit the development of transverse canyons. Transverse canyons, gorges, or water gaps are most easily explained in terms of accelerated headward erosion of rivers along faults cutting across the trend of resistant ridges. In such cases, the fault zone allows rivers to preferentially expand through an already existing ridge of resistant rocks, thereby creating a canyon.



Figure 11: Canyon of the Yellowstone River in Yellowstone National Park, northwestern Wyoming.
By courtesy of Robert Cole

Most transverse canyons, however, are not associated with faults. When faults are absent, transverse canyons are usually interpreted as developing in one of two ways. First, valleys may have been eroded into the landscape before the tectonic features (folds and faults) were developed. Such macrostructures rise across the trend of these valleys, and if the rate of river downcutting can keep pace with the rate at which the structures rise, gorges or canyons will be developed transverse to the structural trend. Because the valleys are older than the tectonic displacement, they are called antecedent. Antecedent canyons have been identified in the Alps, the Himalayas, the Andes, the Pacific coastal ranges of the United States, and every other region of the world that has experienced recent or ongoing tectonism. Second, complexly folded and faulted terranes are sometimes buried by a variable thickness of younger sediment. Drainage patterns develop on the sedimentary cover in a manner similar to those formed in any basin where there is no structural control. If the region is vertically uplifted, the rejuvenated rivers begin to entrench and will eventually be let down across the trends of resistant rocks in the underlying complex of folds and faults. Canyons and their formative rivers following this evolutionary path are said to be superimposed. The concept of superimposition was first used to explain water gaps in the Appalachians, but superimposition has since been employed as a model for drainage evolution in most areas of the world that have experienced uplift during the Cenozoic Era (the past 65,000,000 years).

Concept of superimposition

In light of the above, it is well to note that detailed studies of physiography are indeed rare in mountain belts where the initial topography created by deformation is

still preserved. One area that has been investigated is the Zagros Mountain system near the borderlands of Iraq and Iran from eastern Turkey to the Gulf of Oman. In this region, none of the accepted models for the creation of transverse canyons is totally acceptable, even though all of them may be involved to a certain degree. Instead, it seems likely that drainage development associated with normal processes of denudation can produce canyons transverse to a fold belt (given some heterogeneity in the geologic framework) without requiring some unique preexisting condition in the system.

FLOODPLAINS

Floodplains are perhaps the most common of fluvial features in that they are usually found along every major river and in most large tributary valleys. Floodplains can be defined topographically as relatively flat surfaces that stand adjacent to river channels and occupy much of the area constituting valley bottoms. The surface of a floodplain is underlain by alluvium deposited by the associated river and is partially or totally inundated during periods of flooding. Thus, a floodplain is not only constructed by but also serves as an integral part of the modern fluvial system, indicating that the surface and alluvium must be related to the activity of the present river.

The hydrologic role of floodplains

The above definition suggests that, in addition to being a distinct geomorphic feature, a floodplain has a significant hydrologic role. A floodplain directly influences the magnitude of peak discharge in the downstream reaches of a river during episodes of flooding. In extreme precipitation events, runoff from the watershed enters the trunk river faster than it can be removed from the system. Eventually water overtops the channel banks and is stored on the floodplain surface until the flood crest passes a given locality farther downstream. As a consequence, the flood crest on a major river would be significantly greater if its floodplain did not store water long enough to prevent it from becoming part of the downstream peak discharge. The capacity of a floodplain system to store water can be enormous. The volume of water stored during the 1937 flood of the Ohio River in the east-central United States, for example, was roughly 2.3 times the volume of Lake Mead, the largest artificial reservoir in North America. The natural storage in the Ohio River watershed during this particular event represented approximately 57 percent of the direct runoff.

Because a floodplain is so intimately related to floods, it also can be defined in terms of the water level attained during some particular flow condition of a river. In that sense a floodplain is commonly recognized as the surface corresponding to the bank-full stage of a river—*i.e.,* the water level at which the channel is completely filled. Numerous studies have shown that the average recurrence interval of the bank-full stage is 1.5 years, though this value might vary from river to river. Nonetheless, this suggests that most floodplain surfaces will be covered by water twice every three years. It should be noted, however, that the water level having a recurrence interval of 1.5 years will cover only a portion of the relatively flat valley bottom surface that was defined as the topographic floodplain. Clearly parts of the topographic floodplain will be inundated only during river stages that are considerably higher than bankfull and occur less frequently. Thus, it seems that the definition of a hydrologic floodplain is different from that of the topographic floodplain, and how one ultimately studies a floodplain surface depends on which point of view concerning the feature is considered of greatest significance.

Floodplain deposits, origins, and features. Although valley-bottom deposits result from processes operating in diverse sub-environments, including valley-side sheetwash, the most important deposits in the floodplain framework are those developed by processes that function in and near the river channel. These deposits are normally referred to as (1) lateral accretion deposits, which develop within the channel itself as the river migrates back and forth across the valley bottom, and (2) vertical accretion deposits, which accumulate on the floodplain surface when the river overflows its channel banks.

Lateral accretion deposits

In any valley where the river tends to meander, maximum erosion will occur on the outside bank just downstream from the axis of the meander bend. Detailed studies have shown, however, that deposition occurs simultaneously on the inside of the bend, the volume of deposition being essentially equal to the volume of bank erosion. Thus, a meandering river can shift its position laterally during any interval of time without changing its channel shape or size. Deposition on the inside of the meander bend creates a channel feature known as a point bar (see above *River channel patterns*), which represents the most common type of lateral accretion. Over a period of years point bars expand laterally as the opposite bank is continually eroded backward. The bars progressively spread across the valley bottom, usually as a thin sheet of sand or gravel containing layers that dip into the channel bottom. Point bars tend to increase in height until they reach the level of older parts of the floodplain surface, and the maximum thickness of laterally accreted deposits is controlled by how deeply a river can scour its bottom during recurrent floods. A general rule of thumb is that river channels are probably scoured to a depth 1.75 to two times the depth of flow attained during a flood. Because bank-full depth increases in the downstream direction, the thickness of lateral accretion deposits should increase gradually down the valley.

Vertical accretion deposits

Vertical accretion (also called overbank deposition) occurs when rivers leave their channel confines during periodic flooding and deposit sediment on top of the floodplain surface. The floodplain, therefore, increases in elevation during a flood event. Overbank deposition is usually minor during any given flood event. Table 3 shows measured increments of vertical accretion of floodplain surfaces during a few major floods in the United States. The insignificant deposition reflects the documented phenomenon that maximum concentration of suspended load occurs during the rising phase of any flood. Thus, much of the potential overbank sediment is removed from the system before a river rises to bank-full stage.

Table 3: Increment Rates of Overbank Deposition in Major Floods

river basin	flood	average thickness of deposition (metres)
Ohio River	January–Febuary 1937	0.0024
Connecticut River	March 1936	0.0347
Connecticut River	September 1938	0.0223
Kansas River	July 1951	0.0299

Source: Based on M.G. Wolman and L.B. Leopold, "River Flood Plains: Some Observations on Their Formation," U.S. Geological Survey professional paper no. 282-C, 1957

Because lateral and vertical accretionary processes occur during the same time interval, alluvium beneath a floodplain surface usually consists of both type of deposits. The two types often differ in their particle-size characteristics, with lateral accretion deposits having larger grain sizes. These textural differences, however, are not always present. In fact, suspended-load rivers that transport mostly silt and clay develop point bars composed of fine-grained sediment. Conversely, mixed-load rivers with cohesive banks may deposit sand and gravel on a floodplain surface as vertical accretion deposits.

Floodplains also are developed by braided rivers, but the fluvial processes are more dynamic and less regular. Bars and bank erosion, for example, are not confined to one particular side of the channel, and the river often changes its position without laterally eroding the intervening material. Channels and islands associated with the braided-stream pattern become abandoned, and these eventually coalesce into a continuous floodplain surface when old channels become filled with overbank sediment. The result is that floodplain sediments in a braided system are often irregular in thickness, and recognition of the true floodplain sequence may be complicated because braided streams are often associated with long-term valley aggradation. In this case, the total deposit might appear to be very thick, but the actual floodplain sediment relates only to the present river hydrology. The true floodplain

deposit, therefore, is merely a thin cap on top of a thick, continuous valley fill.

Topography developed on a floodplain surface is directly related to depositional and erosional processes. The dominant feature of lateral accretion, a point bar, is subjected to erosion during high discharge when small channels called chutes are eroded across the back portion of the point bar. As the river shifts laterally and chutes continue to form, point bars are molded into alternating ridges and swales that characterize a distinct topography known as meander scrolls. As the river changes its position, meander-scroll topography becomes preserved as part of the floodplain surface itself. Overbank processes also create microtopography. The latter includes natural levees, which are elongate narrow ridges that form adjacent to channels when the largest particles of the suspended load are deposited as soon as the river leaves the confines of its channel. Natural levees build vertically faster than the area away from the channel, which is known as a backswamp. For example, during the 1973 flood on the Mississippi River, 53 centimetres of sediment were deposited on natural levees, while only 1.1 centimetres accumulated in the backswamp area. The backswamp area of a floodplain is usually much more regular, and its flatness is disrupted only by oxbow channels (abandoned river channels) or by ridgelike deposits known as splay deposits that have broken through natural levees and spread onto the backswamp surface. Oxbows, or oxbow lakes, gradually fill in with silts and clays during normal overbank deposition, leaving that surface more regular than might be expected.

Meander scrolls

Time and the floodplain system. The variety of floodplain deposits and features raises the question as to which process, lateral river migration or overbank flow, is the most important in floodplain development. There is probably no universal answer to this question, but rates of the depositional processes suggest that most floodplains should result primarily from the processes and deposition associated with lateral migration. Assuming that vertical accretion proceeds according to the increments shown in Table 3, the level of a floodplain constructed entirely by overbank deposition should rise at a progressively decreasing rate. This follows because as the floodplain surface is elevated relative to the channel floor, the river stage needed to overtop the banks is also increasing. The floodplain surface, therefore, is inundated less frequently, and the growth rate necessarily decreases. Indeed, studies have shown that the initial phase of floodplain elevation by vertical accretion is quite rapid because flooding occurs frequently. It is generally accepted that 80 to 90 percent of floodplain construction by vertical accretion would take place in the first 50 years of the process. A three-metre thick overbank deposit would probably take several thousand years to accumulate.

Primacy of lateral river migration

Given the above, it seems certain that the total thickness of vertically accumulated sediment will depend primarily on the rate at which the river migrates laterally. In fact, the total thickness of overbank deposition will be controlled by the amount of time it takes a river to migrate across the entire width of the valley. For example, if a floodplain is one kilometre wide and the river shifts laterally at a rate of two metres a year, it will take approximately 500 years for the river to migrate completely across the valley bottom. At any given point in the valley bottom, several metres of overbank sediment may accumulate in that 500-year interval, but the entire deposit will be reworked by lateral erosion when the river once again reoccupies that particular position. Thus the lateral migration rate of the river becomes a limiting factor on the thickness of vertical accretion deposits.

Given the rapidity of lateral migration of sediments shown in many river valleys, it is doubtful that the minor rates of vertical accretion shown in Table 3 could create floodplain surfaces that are predominantly formed by overbank deposition. This conclusion, however, cannot be considered as an inviolate rule. Many rivers have extremely slow rates of lateral migration when geologic conditions prevent riverbank erosion. In these cases, vertical accretion may be the dominant process of floodplain development.

DELTAS

The most important landform produced where a river enters a body of standing water is known as a delta. The term is normally applied to a depositional plain formed by a river at its mouth, with the implication that sediment accumulation at this position results in an irregular progradation of the shoreline. This surface feature was first recognized and named by the ancient Greek historian Herodotus, who noted that sediment accumulated at the mouth of the Nile River resembled the Greek letter Δ (delta). Even though a large number of modern deltas have this triangular form, many display a variety of sizes and shapes that depend on a number of environmental factors. Thus, the term now has little, if any, shape connotation. Deltas, in fact, exhibit tremendous variation in their morphological and sedimentologic characteristics and also in their mode of origin. Most of the variation results from (1) characteristics within the drainage basin that provides the sediment (*e.g.,* climate, lithology, tectonic stability, and basin size); (2) properties of the transporting agent, such as river slope, velocity, discharge, and sediment size; and (3) energy that exists along the shoreline, including such factors as wave characteristics, longitudinal currents, and tidal range. The shoreline zone, therefore, becomes the battleground between variable amounts and sizes of sediment delivered to the river mouth and the energy of the ocean waters at that site. The balance between these two factors determines whether accumulation of the river-borne sediment will occur or whether ocean processes will disperse the sediment or prevent its deposition. The combination of these numerous variables tends to create deltas that occur in a complete spectrum of form and depositional style.

Causes of variations in morphological and sedimentologic characteristics

Deltas are distributed over all portions of the Earth's surface. They form along the coasts of every landmass and occur in all climatic regimes and geologic settings. The largest deltas of the world are those created by major river systems draining regions that are subcontinental in size and yield abundant sediment from the watershed.

Classification of deltas. Deltas come in a multitude of plan-view shapes, as their characteristics are determined by the balance between the energy and sediment load of a fluvial system and the dynamics of the ocean. Various ways of classifying deltas have been devised. One of the more widely used schemes is based on deltaic form as it reflects controlling energy factors. This scheme divides deltas into two principal classes: high-constructive and high-destructive.

High-constructive deltas develop when fluvial action and depositional process dominate the system. These deltas usually occur in either of two forms. One type, known as elongate, is represented most clearly by the modern bird-foot delta of the Mississippi River. The other, called lobate, is exemplified by the older Holocene deltas of the Mississippi River system (see Figure 12). Both of these high-constructive types have a large sediment supply relative to the marine processes that tend to disperse sediment along the shoreline. Normally, elongate deltas have a higher mud content than lobate deltas and tend to subside rather rapidly when they become inactive.

High-destructive deltas form where the shoreline energy is high and much of the sediment delivered by the river is reworked by wave action or longshore currents before it is finally deposited. Deltas formed by rivers such as the Nile and the Rhône have been classified as wave-dominated. In this class of high-destructive delta, sediment is finally deposited as arcuate sand barriers near the mouth of the river. In another subtype, called tide-dominated, tidal currents mold the sediment into sandy units that tend to radiate in a linear pattern from the river mouth. In such a delta, muds and silts are deposited inland of the linear sands, and extensive tidal flats or mangrove swamps characteristically develop in that zone.

Considerable attention has been given to deltas that are composed of very coarse deposits—those of sand and gravel. Deltas developing from this type of material are commonly classified as either fan deltas or braid deltas. A fan delta is a depositional feature that is formed where an alluvial fan develops directly in a body of standing water from some adjacent highland. A braid delta is a coarse-

Special delta types

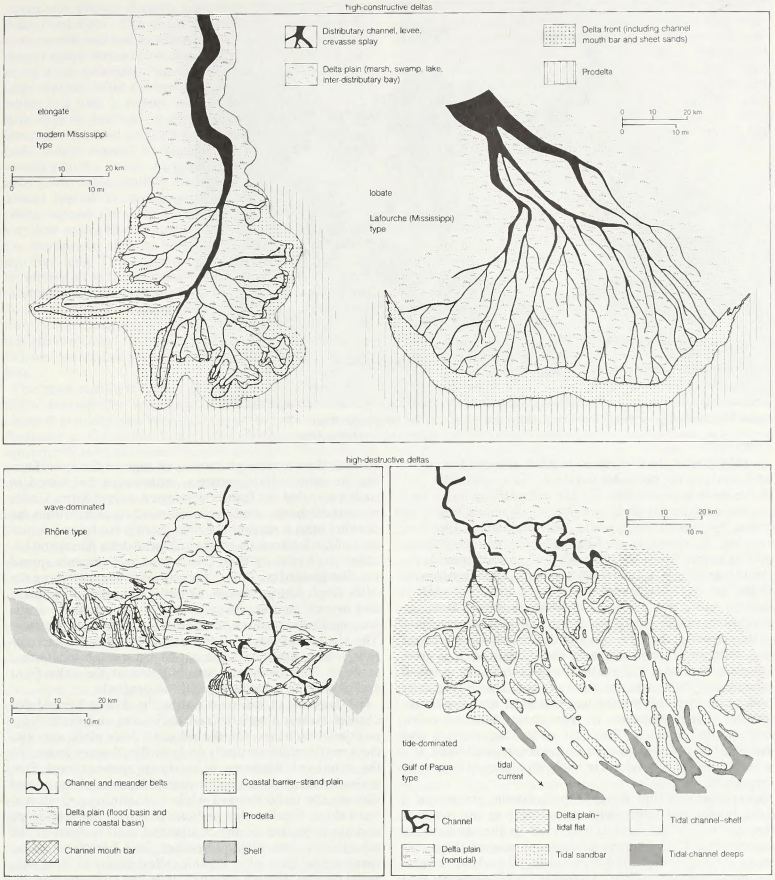


Figure 12: Basic delta types.
From W.L. Fisher *et al., Delta Systems in the Exploration for Oil and Gas* (1969), Bureau of Economic Geology, the University of Texas at Austin

grained delta that develops by progradation of a braided fluvial system into a body of standing water. The two are related by the fact that they are composed primarily of very coarse sediment; however, they differ in that braid deltas result from well-defined, highly channelized braided rivers that are deeper and have more sustained flow than streams which develop alluvial fans. In addition, the braided system that ultimately forms the braid delta may have its source far removed from the body of standing water and may in fact consist of large alluvial plains rather than the restricted areal and longitudinal extent associated with alluvial fans.

Morphology of deltas. Deltas consist of three physiographic parts called the upper delta plain, the lower delta plain, and the subaqueous delta. The upper delta plain begins as the river leaves the zone where its alluvial plain is confined laterally by valley walls. When the valley wall constraint ends, the river breaks into a multitude of channels, and the depositional plain widens. This point source of the upper delta plain can be thought of as the apex of the entire delta, which is analogous to the same reach of an alluvial fan. The entire upper delta plain is fluvial in origin except for marshes, swamps, and freshwater lakes that exist in areas between the many river channels. The surface of the upper delta plain is above the highest tidal level and thus is not affected by marine processes. In contrast, the lower delta plain is occasionally covered by tidal water. For this reason, the boundary between the upper delta plain and the lower delta plain is determined by the maximum tidal elevation. Features and deposits in the lower delta plain are the result of both fluvial and marine processes. Tidal flats, mangrove swamps, beach ridges, and brackish-water bays and marshes are common in this zone.

Deltas affected by high tidal ranges, such as those constructed by the Niger River and the Ganges–Brahmaputra system, are dominated by marine incursions and expansive lower delta plains. For example, the Ganges–Brahmaputra system in Bangladesh has a lower delta plain that occupies more than half of its total surface area of 60,000 square kilometres and is characterized by enormous mangrove swamps. Low tidal ranges result in deltas having better developed upper delta plains (*e.g.*, the Nile of Egypt).

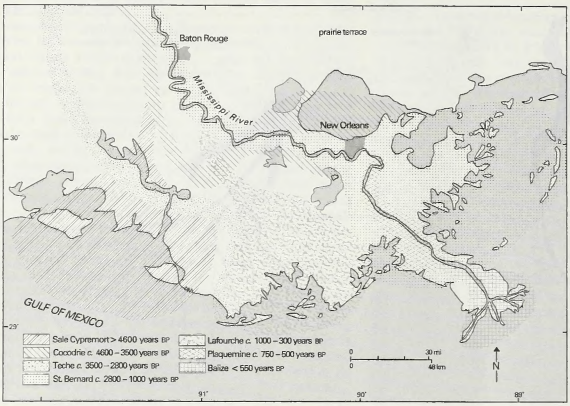


Figure 13: Chronology of deltas that make up the deltaic plain of the Mississippi River.
From J.P. Morgan, "Deltas—a Resume," *Journal of Geological Education* (1970); used with permission of the National Association of Geology Teachers

Subaqueous delta plain

A subaqueous delta plain is located entirely below sea level, and marine processes dominate the system. This part of the delta is responsible for the topographic bulge seen on the continental shelf seaward of channels that flow across the exposed delta plains. Sediment-laden river flow entering the ocean in well-defined channels loses transporting power where the channels end, and sediment is deposited as the subaqueous delta plain. Large subaqueous plains are best developed where the continental shelf is shallow and gently sloping and where sediment loads derived from source basins are great. The subaqueous deltas of the Amazon, the Orinoco, and the Huang Ho are broad and widespread in response to these controls. It is true, however, that even if these ideal conditions exist, a broad subaqueous delta does not always result. This is especially true where large submarine canyons exist near the terminations of river channels. In these cases, sediment delivered to the ocean is funneled down the canyon and deposited beyond the margin of the continental shelf. If a subaqueous delta develops in such situations, it is usually very small.

Channel patterns

River channels that traverse the subaerial portion of a delta (upper and lower delta plain) serve as the conduits through which sediment is delivered to the subaqueous component. The channels assume any one of three patterns: (1) long, straight single channels, (2) braided or anastomotic (veinlike) multiple channels, or (3) channels that bifurcate (branch) in a downstream direction. In general, the channel pattern is controlled both by source basin characteristics (sediment size and volume, flood-discharge features, etc.) and marine properties (tidal range and wave energy, for example). Rivers transporting fine-grained sediment tend to develop either single channels or downstream bifurcating patterns. The single-channel pattern results where offshore wave energy is high (*e.g.,* the Mekong and Congo deltas). Braided or anastomotic channels develop best where rivers carry a large volume of coarse-grained bed load. Branching distributaries form most commonly where tidal range and wave energy is low (*e.g.,* the Mississippi and Volga deltas).

Most delta channels are bordered by natural levees that resemble those found on floodplains. These features are best developed by rivers that flood frequently and transport large volumes of suspended load, as, for example, the Mississippi. Interfluve areas (those between adjacent streams flowing in the same direction) are variable in character, depending on climate, tidal range, and offshore wave energy.

Deposits and stratigraphy. Delta growth indicates that a river delivers sediment to the shore faster and in greater volume than marine processes can remove the load. During the delta-building process, sediment is distributed in such a way that the feature develops a unique form. Under normal discharge conditions, sediment remains within the channel until it reaches the river mouth. No lateral dispersion of the load occurs on the subaerial delta plain, and because river velocity is so low, waves and currents spread the fine-grained portion of the sediment laterally along the delta front. During floods, however, suspended sediment and organic matter are deposited in the interfluve areas, causing those portions of the subaerial delta to aggrade. The high river velocity at the mouth offsets wave and current action, allowing sediment to be transported farther seaward. This facilitates accumulation at the delta front and causes the subaqueous delta to prograde.

The dispersal of sediment during floods and normal discharges creates a well-defined horizontal and vertical depositional sequence. On the subaerial delta plain, silts and clays accumulate vertically in inter-distributary zones. At the mouths of deltaic rivers, marine processes rework fine-grained sediment, but more coarse deposits of sands and silts usually build forward while maintaining a steep seaward slope. Smaller clay particles pass over the delta slope and are deposited on the continental shelf in front of the subaqueous delta plain. Therefore, in a horizontal sense, many deltas have silty, organic-rich deposits in their subaerial portion, though channel sands and levee deposits interrupt the fine-grained interfluve sequence. More coarse sediment is deposited at the river mouth, and very fine-grained materials (clays) accumulate beyond the delta front. The vertical sequence is essentially the same with marine clays at the lowest elevation (greatest depth), silts and sands at nearshore depths, and silts, clays, and organics—along with associated channel and levee sands—at the highest (subaerial) elevations. This model of alluviation does not accommodate very coarse (gravel and sand) deposition on the subaerial delta plain, which provides the special deltaic types known as fan deltas or braid deltas (see above), but it is representative of most of the major deltas of the world.

Topset, foreset, and bottomset beds

Deposits found in the deltaic stratigraphic sequence were named topset, foreset, and bottomset by the American geologist Grove K. Gilbert in his 1890 report on Lake Bonneville, the vast Pleistocene ancestor of what is now the Great Salt Lake of Utah. Although Gilbert examined small deltas along the margins of the ancient lake, the stratigraphic sequence he observed is similar to that found in large marine deltas. Topset beds are a complex of lithologic units deposited in various sub-environments of the subaerial delta plain. Layers in the topset unit are almost

horizontal. Foreset deposits accumulate in the subaqueous delta front zone. The deposits are usually coarser at the river mouth and become finer as they radiate seaward into deeper water. Strata in the foreset unit are inclined seaward at an angle reflecting that of the delta slope or front. In large marine deltas the beds rarely dip more than 1°, but where bed load is coarse, such as in braid deltas, foreset beds may be inclined at angles greater than 20°. Foreset layers are beveled at their landward positions by topset beds, which expand horizontally as the entire delta advances into the ocean. At their seaward extremity, foreset beds grade imperceptibly into the bottomset strata. Bottomset deposits are composed primarily of clays that were swept beyond the delta front. These beds usually dip at very low angles that are consistent with the topography of the continental shelf or lake bottom in front of the subaqueous delta. This depositional environment is commonly referred to as the prodelta zone.

Deltas and time. One of the most important perceptions needed to understand deltas is how their depositional framework changes with time. Because delta characteristics are controlled by factors that are subject to change, it follows that deltaic growth patterns are dynamic and variable.

The most significant effect is that the site of deposition shifts dramatically with time. This occurs because the channel gradient and transporting power of a delta river decreases as the deltaic lobe extends farther seaward and shorter routes to the ocean become available. These shorter pathways may begin far inland, usually being occupied when the river is diverted through breaches in levees called crevasses. This process effectively shifts the locus of deposition and initiates the development of a new deltaic lobe. For example, the Mississippi Delta actually consists of the coalescence of seven major lobes constructed at different times and positions during the last 5,000 years (Figure 13). In fact, the modern bird-foot delta of the Mississippi River is only a small part of the entire deltaic system, and there is good reason to believe that another major shift in the depositional position is imminent. The Atchafalaya River, a major distributary, branches from the Mississippi upstream from Baton Rouge, La., and its route to the ocean is approximately 300 kilometres shorter than the present course of the Mississippi. This channel carries 30 percent of the Mississippi flow, and sediment reaching Atchafalaya Bay (160 kilometres west of New Orleans) is actively building a new delta lobe. Complete diversion of the Mississippi discharge into the Atchafalaya will accelerate growth of the new delta. The present bird-foot delta will be abandoned and, starved of any incoming sediment, will become severely eroded by the unopposed attack of marine processes.

Formation of subdeltas

Even within a modern delta, water and sediment, funneled through crevasses, build smaller subdeltas, which are ephemeral in space and time. What emerges is a picture of a dynamic system in which depositional sites change over different time scales. On a short-term basis (years to decades), a limited area (subdelta) may receive sediment, but the position of accumulation shifts rapidly. On a longer time scale (hundreds to thousands of years), the position of an entire active delta may migrate over a considerable distance.

ESTUARIES

Estuaries are partially enclosed bodies of water located along coastal regions where flow in downstream reaches of rivers is mixed with and diluted by seawater. The landward limit of an estuary is defined in terms of salinity, often where chlorinity is 0.01 parts per thousand. The inland extent of this chemical marker, however, varies according to numerous physical and chemical controls, especially the tidal range and the chemistry of river water. Actually, the term estuary is derived from the Latin words *aestus* ("the tide") and *aestuo* ("boil"), indicating the effect generated when tidal flow and river flow meet. Nonetheless, if estuaries are defined on the basis of salinity, many coastal features such as bays, tidal marshes, and lagoons can be regarded as estuaries.

Estuaries have always been extremely important to humankind. From early times, they have served as centres of shipping and commerce. In fact, many seaports were originally founded at the seaward margin of major river systems. Concomitantly, some of the oldest civilizations developed in estuarine environments. In addition to shipping, much of the world's fishing industry is dependent on the estuarine environment. Many species of fish and shelled bottom dwellers spend much of their life cycle there. In most cases, these animals have a tolerance for wide ranges in salinity and temperature. Pollutants introduced by humans, however, can affect such forms of marine life significantly if large enough amounts of the contaminants accumulate among bottom sediments.

Origin and classification. Most modern estuaries formed as the result of a worldwide rise in sea level, which began approximately 18,000 years ago during the waning phase of the Wisconsin Glacial Stage. When glaciation was at its maximum, sea level was significantly lower than it is today because much of the precipitation falling on the continents was locked up in massive ice bodies rather than returning to the ocean. In response, rivers entrenched their downstream reaches as baselevel (sea level) declined. As the ice began to dissipate, sea level rose, and marine waters invaded the entrenched valleys and inundated other portions of the coastal zone, such as deltas and coastal plains. It is known that the subsidence of a coast produces the same effect as a rise in sea level; thus tectonic activity sometimes creates estuaries.

Models of formation

In general, estuaries develop in one of three ways. First, estuaries represent drowned valleys. The valleys may have been formed by normal river entrenchment (*e.g.,* Chesapeake Bay in the eastern United States) or as the result of glacial erosion. The latter type, called fjords, are deep, narrow gorges cut into bedrock by tongues of glacial ice advancing down a former stream valley (see CONTINENTAL LANDFORMS: *Glacial landforms*). Fjords are most common in Norway and the coastal margins of British Columbia, Can. Both valley types (river and glacial) became estuarine environments with the postglacial rise in sea level. Second, some estuaries develop when barrier islands and/or spits enclose large areas of brackish water between the open ocean and the continental margin. These depositional features restrict free exchange between river and marine water and create lagoons or partially enclosed bays that develop the chemical characteristics of an estuarine environment. Such settings are best exemplified in the Gulf Coast region of the United States (*e.g.,* Galveston Bay), the Vadehavet tidal area of Denmark, the Swan Estuary of Western Australia, and the Waddenzee of The Netherlands. Third, some estuaries are clearly submerged in response to tectonic activity, such as down-faulted coastal zones or isostatically controlled subsidence (*e.g.,* San Francisco Bay).

Physical oceanographers commonly classify valley-type estuaries by the process and extent of mixing between fresh water and seawater. A salt-wedge estuary is dominated by river discharge, and tidal effects are negligible. In this situation, fresh water floats on top of seawater as a distinct layer, which thins toward the ocean. A wedge-shaped body of seawater underlies the freshwater layer and thins toward the continent. The interface between the two water types is well defined, and very little mass transfer or mixing of the two waters occurs. Partially mixed estuaries are characterized by an increased tidal effect to a condition where river discharge does not dominate the system. Mixing of the two water types is prominent in this system and is caused by increased turbulence. Mass transfer of water involves movement in both directions across a boundary that becomes less distinct than the one found in the salt-wedge estuaries. In vertically homogeneous estuaries, the velocity of tidal currents is large enough to produce total mixing and eliminate the fresh/salt water boundary. The water salinity is constant in the vertical sense and tends to decrease toward the continent. In general, the classification of estuaries by mixing indicates that the more substantial the river discharge, the weaker is the mixing. In addition, the dominance of river flow causes a greater salinity gradient. This indicates that sizable fluvial activity tends to block the entrance of seawater.

Sedimentation in estuaries. The bedrock floor near the mouth of most estuaries is usually buried by a thick accumulation of sediment. The texture and composition of sediment in estuaries in the United States is known to be a function of river-basin geology, bathymetry, and hydrologic setting. Where sediment supply is inadequate to fill drowned valleys, clay and silt are usually deposited in the central part of bays and grade shoreward and seaward into bodies of sand. Where sediment supply and tidal range are both large, such as in Oregon and northern California in the western United States, the clay and silt are commonly swept from the channels and deposited on the marginal flats. In the Gulf Coast region, small tides and abundant fine-grained sediment tend to create very shallow estuaries. Silt and clay are usually deposited in lagoons behind barrier bars, although these grade into sands around the lagoonal margins.

The character and distribution of estuarine sediment are influenced by many physical, chemical, and biologic processes, such as tidal currents, flocculation, bioturbation (the reworking and alteration of sediment by organisms), storms, morphology of the estuary, and human activities. The sediment type that is deposited, therefore, depends on the dynamics of the system, which in turn are controlled by an equilibrium between river and tidal flow. River discharge develops inertia, which results in the collision of river and ocean waters in the estuary itself. Most sediment is derived from the river system, and whether or not it is deposited within the estuary depends on how quickly the velocity is diminished by the effect of tidal currents and by the extent of the tidal range. Notwithstanding the above, it has been long recognized that net sediment transport in many open estuaries can be from the sea toward the land.

(D.F.R.)

The river system through time

Natural river systems can be assumed to have operated throughout the period of geologic record, ever since continental masses first received sufficient precipitation to sustain external surface runoff. The Precambrian portion of the record, prior to 570,000,000 years ago, is complicated by the widely metamorphosed character of the surviving rocks, although even here the typical cross-bedding of shallow-water sands can be recognized in many places. The Cambrian and post-Cambrian succession of the last 570,000,000 years contains multiple instances of deposition of deltaic sandstones, which record intermittent deposition by rivers in many areas at many intervals of past time. The span since the Precambrian is long enough, at present rates of erosion, for rivers to have shifted the equivalent of 25 to 30 times the bulk of the existing continental masses, but the rate of erosion and sedimentation is estimated to have increased with time. Of necessity, river systems now in existence date from times not earlier than the latest emergence of their basins above sea level, but this limitation allows numbers of them to have histories of 100,000,000 years or more in length.

Evidence of ancient rivers

DRAINAGE DIVERSION BY STREAM CAPTURE

A river system of appreciable size is likely to have undergone considerable changes in drainage area, network pattern, and profile and channel geometry. Adjoining streams compete with one another for territory. Although competition is effectively nil where divides consist of expanses of plateau or where opposing low-order streams of similar slope flow down the sides of ridges, it frequently happens that fluvial erosion is shifting a divide away from some more powerful trunk stream and toward a weaker competing trunk. In extreme cases, the height difference is so marked that a tributary head from one system can invade, and divert, a channel in the adjoining system: such diversion, termed stream capture, has already been noted as a principal mechanism in the adjustment of network patterns to structural patterns. Close general adjustment to structure implies multiple individual adjustments, unless the stream network has developed solely by the headward extension of tributaries along lines of structural and lithologic weakness: the network predicated on a single regional slope is dendritic in pattern. By encroachment and capture a successfully competing stream becomes yet more powerful, the headward extension of its basin increasing the discharge of the trunk channel and permitting reduction of slope; *i.e.,* additional downcutting. Seaward extensions of basins occur where deltas lead the outbuilding of alluvial plains and where crustal uplift (and also at times strandline movements) result in emergence. Conversely, basin area is reduced along the seaward edge by submergence, in response to crustal depression or rise in sea level. The potential limits to basin size are fixed by available areas of continent with surface moisture surplus, in combination with theoretical optimum shape of basin; however, actual basin shapes, for all large rivers, are to some extent affected by crustal deformation.

NON-FLUVIAL INVASION AND DEPOSITION

Derangements other than the captures effected in stream competition include those due to non-fluvial invasion and deposition. Regional flooding by basalts, as during the Tertiary Period (from 65,000,000 to 2,500,000 years ago) in the Deccan of India and the northwestern part of the United States, obliterates the former landscape and provides a new surface on which new drainage networks form. Major invasions by continental ice displaces fluvial systems for the time being. Glacial deposits, especially till sheets, can conceal the preglacial topography and provide initial slope systems for postglacial streams. Individual diversions occur at and near ice fronts, also where preglacial divides in mountain country are breached by the ice of caps or impounded mountain glaciers. The full history of drainage derangement by continental ice is often complex, depending on the particular combinations of preglacial outlet directions, extent of glacial invasion, relationship of regional slope to direction of ice advance, thickness of glacial sedimentation, amount and speed of postglacial isostatic rebound, and self-selection of postglacial outlet directions and drainage lines. The North American Great Lakes and Midwest areas, the Thames Basin in England, and the Eurasiatic plain all record intricate histories of damming during glacial maxima, with postglacial networks and outlets differing markedly from those of preglacial times. Glacial breaching of divides requires the passage of thick ice through a preglacial notch or gap, with erosion severe enough to provide a new drainage line when the ice melts. The spinal divide of Scandinavia was breached by the ice cap centred over the Gulf of Bothnia, just as the highland rim of Greenland is being breached by effluent glaciers today. After deglaciation, areas of divide breaching display streams with anomalous courses through gaps in major relief barriers. Morphologically related to glacial breaching, especially with respect to indeterminate present-day divides, are the disordered drainage nets of formerly glaciated terrains where bedrock is widely exposed and where relief is subdued.

Role of continental ice

EFFECTS OF CLIMATIC CHANGE

Climatic shifts are known to be capable of effecting fill or clearance of channels and valleys: they can also change channel habit. In addition to the alternation in some near-glacial areas between braiding during maximum cold and meandering during interglacial warmth, the record includes conversions of channel width and meander pattern. On numerous mid-latitude streams, existing channels have been much reduced from their earlier dimensions; and on many, but by no means all streams, existing floodplains are contained in the floors of meandering valleys where the wavelength is determined by the plan of the floodplain as opposed to the existing channel. Valley meanders were cut by streams 20 to 100 times as voluminous as existing streams, at the bank-full stage. They illustrate only one variant, although a widespread one, of the underfit stream, which combines a former large with an existing reduced channel. Reduction to the underfit state is commonly, although not invariably, accompanied by the infilling of former large channels both laterally and from below, so that existing floodplains are contained in valley-bottom fills. Accidents of capture and glacial diversion apart, the underfit condition results generally from climatic shift. The

last major shift responsible for channel shrinkage appears to have occurred in the interval 12,000–9,000 BP, or later in areas that were still ice-covered 9,000 years ago. Involving a reduction of bed width to as much as one-tenth of earlier values, and in meander wavelength by similar proportions, channel shrinkage is known to have been succeeded in well-studied areas by lesser fluctuations that are recorded in episodes of partial clearance followed by renewed fill. Significant alternations between cut and fill during the last 10,000 to 20,000 years have perhaps averaged a periodicity of 1,000 to 2,000 years. There is no a priori reason to suppose that the corresponding periodicity differed from this value during the whole Pleistocene, 2,000,000 to 3,000,000 years in duration so far. Inferences about pre-Pleistocene fluctuations await detailed analysis of rates of deposition of graded beds, coral growth, and the like.

River channels and networks as open systems

On account of the temporal–dynamic qualities that have been discussed, river channels and networks are to be regarded as open systems (those open to additions or subtractions of materials or energy through time), whether in relation to short-term adjustments to individual peak discharges, in relation to accommodation to the constraints of climate, vegetal cover, characteristics of infiltration and overland flow, or in relation to the long-term influences of crustal movement, interbasin competition, and land wastage. Channels and networks experience inputs and outputs of matter and energy. Some of them, but probably a small minority at any one time and for a minor duration of total time in any one channel or network, act as open systems in disequilibrium. The general tendency seems to be for channel and river systems to attain steady-state conditions, wherein negative feedback tends to counter individual disequilibrium tendencies, and counteracting effects ensure variations about recurrent norms of form and behaviour. (G.H.D.)

BIBLIOGRAPHY

General works: Discussions of all aspects of rivers are found in LUNA B. LEOPOLD, M. GORDON WOLMAN, and JOHN P. MILLER, *Fluvial Processes in Geomorphology* (1964); G.H. DURY (ed.), *Rivers and River Terraces* (1970); RICHARD J. CHORLEY, STANLEY A. SCHUMM, and DAVID E. SUGDEN, *Geomorphology* (1984); ARTHUR L. BLOOM, *Geomorphology: A Systematic Analysis of Late Cenozoic Landforms* (1978); DALE F. RITTER, *Process Geomorphology,* 2nd ed. (1986); and MARIE MORISAWA, *Rivers: Form and Process* (1985). See also LAURENCE PRINGLE, *Rivers and Lakes* (1985); and the *Rand McNally Encyclopedia of World Rivers* (1980). (G.H.D./D.F.R./Ed.)

Environmental problems attendant on river use are discussed in M.J. STIFF (ed.), *River Pollution Control* (1980); *Environmental Effects of Cooling Systems* (1980), a report from the International Atomic Energy Agency on cooling systems and thermal discharges from nuclear power stations; *Cooling Water Discharges from Coal Fired Power Plants: Water Pollution Problems* (1983), proceedings of an international conference; R.G. TOMS, "River Pollution-Control Since 1974," *Water Pollution Control,* 84(2):178–186 (1985); and M. CHEVREUIL, A. CHESTERIKOFF, and R. LETOLLE, "PCB Pollution Behavior in the River Seine," *Water Research,* 21(4): 427–434 (April 1987). (Ed.)

River channels and waterfalls: Works on the formation and change of river channels include WALTER B. LANGBEIN and LUNA B. LEOPOLD, *River Meanders, Theory of Minimum Variance* (1966), U.S. Geological Survey professional paper no. 422-H; MARK A. MELTON, "Methods for Measuring the Effect of Environmental Factors on Channel Properties," *Journal of Geophysical Research,* 67(4):1485–90 (April 1962); and N.A. RZHANITSYN, *Morphological and Hydrological Regularities of the River Net* (1964; originally published in Russian, 1960).

A dated but still useful source on waterfalls is the article by THEODORE W. NOYES, "The World's Greatest Waterfalls," *National Geographic Magazine,* 50:29–59 (July 1926), on the Niagara, Victoria, and Iguaçu falls. Modern treatments of waterfalls are rare; the interested reader might best consult the following references: H.F. GARNER, "Derangement of the Rio Caroni, Venezuela," *Revue de Géomorphologie Dynamique,* 16:54–83 (1966), describing the occurrence of Angel Falls; MARTIN VON SCHWARZBACH, "Isländische Wasserfälle und eine genetische Systematik der Wasserfälle überhaupt," *Zeitschrift für Geomorphologie,* 11:377–417 (Dec. 1967), one of the best general surveys of the several kinds and occurrences of waterfalls, with specific reference to Icelandic examples; SHAILER S. PHILBRICK, "Horizontal Configuration and the Rate of Erosion of Niagara Falls," *Geological Society of America Bulletin,* 81(2):3723–31 (Dec. 1970), providing a summary of information on the history of Horseshoe Falls and on the general recession of cap-rock-type falls; EBERHARD CZAYA, "Waterfalls and Rapids," ch. 4 in his *Rivers of the World* (1981, reprinted 1983; originally published in German, 1981), pp. 121–137, which includes a list of famous waterfalls classified by location and height; and R.W. YOUNG, "Waterfalls: Form and Process," *Zeitschrift für Geomorphologie Supplementband,* 55:81–95 (1981). (G.H.D./L.K.L.)

Rivers as agents of landscape evolution: Literature concerning the evolution of valleys and the origin of transverse canyons is found mostly in older classic treatments of the topic, such as WILLIAM MORRIS DAVIS, *Geographical Essays* (1909, reprinted 1954); and two articles from the *Bulletin of the Geological Society of America*: ARTHUR N. STRAHLER, "Hypotheses of Stream Development in the Folded Appalachians of Pennsylvania," 56 (1):45–87 (Jan. 1945); and J. HOOVER MACKIN, "Erosional History of the Big Horn Basin, Wyoming," 48(6):813–893 (June 1, 1937). An excellent discussion of the initial development of valleys and canyons in an area of recent tectonism is given in THEODORE OBERLANDER, *The Zagros Streams: A New Interpretation of Transverse Drainage in an Orogenic Zone* (1965). Discussions and additional references about valley morphology in natural and experimental settings can be found in STANLEY A. SCHUMM, *The Fluvial System* (1977); and STANLEY A. SCHUMM, M. PAUL MOSLEY, and WILLIAM E. WEAVER, *Experimental Fluvial Geomorphology* (1987).

Detailed analyses of floodplains are provided by EDWARD J. HICKIN and GERALD C. NANSON, "The Character of Channel Migration on the Beatton River, Northeast British Columbia," *Geological Society of America Bulletin,* 86(4):487–494 (April 1975); and by two U.S. Geological Survey professional papers: M. GORDON WOLMAN and LUNA B. LEOPOLD, *River Flood Plains: Some Observations on Their Formation* (1957), no. 282-C; and STANLEY A. SCHUMM and R.W. LICHTY, *Channel Widening and Floodplain Construction Along Cimarron River in Southwestern Kansas* (1963), no. 352-D.

Extensive reviews of delta formation can be found in MARTHA L. SHIRLEY (ed.), *Deltas in Their Geologic Framework* (1966); W. FISHER *et al., Delta Systems in the Exploration for Oil and Gas* (1969, reprinted 1974); and JAMES P. MORGAN, "Deltas—A Résumé," *Journal of Geological Education,* 18(3):107–117 (May 1970). Detailed analyses of specific deltas, including the Mississippi River delta, can be found in FRANCIS P. SHEPARD, FRED B. PHLEGER, and TJEERD H. VAN ANDEL (eds.), *Recent Sediments, Northwest Gulf of Mexico* (1960); and RICHARD J. RUSSELL, "Geomorphology of the Rhône Delta," *Annals of the Association of American Geographers,* 32(2):149–254 (June 1942), and *River and Delta Morphology* (1967).

The characteristics and formative processes of estuaries are discussed in ANDRÉ GUILCHER, *Coastal and Submarine Morphology* (1958); MAURICE L. SCHWARTZ (ed.), *The Encyclopedia of Beaches and Coastal Environments* (1982), with illustrated entries on estuaries and on estuarine coasts, deltas, habitats, and sedimentation; RUSSELL SACKETT, *Edge of the Sea,* rev. ed. (1985); and ERIC C.F. BIRD, *Coasts: An Introduction to Coastal Geomorphology,* 3rd ed. (1984). More technical treatment is presented in GEORGE H. LAUFF (ed.), *Estuaries* (1967); BRUCE W. NELSON (ed.), *Environmental Framework of Coastal Plain Estuaries* (1973); and L. EUGENE CRONIN (ed.), *Estuarine Research,* vol. l, *Chemistry, Biology, and the Estuarine System* (1975). Estuarine sediment is described for 45 estuarine zones of the United States in DAVID W. FOLGER, *Characteristics of Estuarine Sediments of the United States* (1972), U.S. Geological Survey professional paper no. 742. (D.F.R.)

Rock Music

It is certainly arguable that by the end of the 20th century rock was the world's dominant form of popular music. It originated in the United States in the 1950s and spread to other English-speaking countries and across Europe in the '60s, and by the '90s its impact was obvious globally (if in many different local guises). Rock's commercial importance was by then reflected in the organization of the multinational recording industry, in the sales racks of international record retailers, and in the playlist policies of music radio and television. If other kinds of music—classical, jazz, easy listening, country, folk, etc.—are marketed as minority interests, rock defines the musical mainstream. And so over the last half of the 20th century it became the most inclusive of musical labels—everything can be "rocked"—and in consequence the hardest to define. To answer the question What is rock? one first has to understand where it came from and what made it possible. And to understand rock's cultural significance one has to understand how it works socially as well as musically.

What is rock? 870
The difficulty of definition
Crucial rock musicians
Musical eclecticism and the use of technology
Rock and youth culture
Authenticity and commercialism
Rock in the 1950s 872
The development of the new vocal pop star
Rural music in urban settings
Marketing rock and roll
Rock in the 1960s 873
A black and white hybrid
The British reaction
Folk rock, the hippie movement, and "the rock paradox"
Rock in the 1970s 874
Corporate rock
Challenges to mainstream rock
Rock in the 1980s and '90s 874
Digital technology and alternatives to adult-oriented rock
The global market and fragmentation
Rock as a reflection of social and cultural change 875
Bibliography 876

WHAT IS ROCK?

The difficulty of definition. Dictionary definitions of rock are problematic, not least because the term has different resonance in its British and American usages (the latter is considerably broader in compass). There is basic agreement that rock "is a form of music with a strong beat," but it is difficult to be much more explicit. The *Collins Cobuild English Dictionary,* which is based on a vast database of British usage, suggests that "rock is a kind of music with simple tunes and a very strong beat that is played and sung, usually loudly, by a small group of people with electric guitars and drums," but there are so many exceptions to this description that it is practically useless.

Legislators seeking to define rock for regulatory purposes have not done much better. The Canadian government defined "rock and rock-oriented music" as "characterized by a strong beat, the use of blues forms and the presence of rock instruments such as electric guitar, electric bass, electric organ or electric piano." This assumes that rock can be marked off from other sorts of music formally, according to its sounds. In practice, though, the distinctions that matter for rock fans and musicians have been ideological. *Rock* was developed as a term to distinguish certain music-making and listening practices from those associated with pop;

Ron Kuntz—Reuters/© Archive Photos



Chuck Berry (right) and Bruce Springsteen performing at the inauguration of the Rock and Roll Hall of Fame in Cleveland, Ohio, September 1995.

what was at issue was less a sound than an attitude. In 1990 British legislators defined pop music as "all kinds of music characterized by a strong rhythmic element and a reliance on electronic amplification for their performance." This led to strong objections from the music industry that such a definition failed to appreciate the clear sociological difference between pop ("instant singles-based music aimed at teenagers") and rock ("album-based music for adults"). In pursuit of definitional clarity, the lawmakers misunderstood what made rock music matter.

Crucial rock musicians. For lexicographers and legislators alike, the purpose of definition is to grasp a meaning, to hold it in place, so that people can use a word correctly—for example, to assign a track to its proper radio outlet (rock, pop, country, jazz). The trouble is that the term *rock* describes an evolving musical practice informed by a variety of nonmusical arguments (about creativity, sincerity, commerce, and popularity). It makes more sense, then, to approach the definition of rock historically, with examples. The following musicians were crucial to rock's history. What do they have in common?

Elvis Presley

Elvis Presley, from Memphis, Tennessee, personified a new form of American popular music in the mid-1950s. Rock and roll was a guitar-based sound with a strong (if loose) beat that drew equally on African-American and white traditions from the southern United States, on blues, church music, and country music. Presley's rapid rise to national stardom revealed the new cultural and economic power of both teenagers and teen-aimed media—records, radio, television, and motion pictures.

The Beatles, from Liverpool, England (via Hamburg, Germany), personified a new form of British popular music in the 1960s. Merseybeat was a British take on the black and white musical mix of rock and roll: a basic lineup of lead guitar, rhythm guitar, bass guitar, and drums (with shared vocals) provided local live versions of American hit records of all sorts. The Beatles added to this an artistic self-consciousness, soon writing their own songs and using the recording studio to develop their own—rather than a commercial producer's—musical ideas. The group's unprecedented success in the United States ensured that rock would be an Anglo-American phenomenon.

Bob Dylan, from Hibbing, Minnesota (via New York City), personified a new form of American music in the mid-1960s. Dylan brought together the amplified beat of rock and roll, the star imagery of pop, the historical and political sensibility of folk, and—through the wit, ambition, and obscurity of his lyrics—the arrogance of urban bohemia. He gave the emerging rock scene artistic weight (his was album, not Top 40, music) and a new account of youth as an ideological rather than a demographic category.

Jimi Hendrix, from Seattle, Washington (via London), personified the emergence of rock as a specific musical genre in the late 1960s. Learning his trade as a guitarist in rhythm-and-blues bands and possessing a jazzman's commitment to collective improvisation, he came to fame leading a trio in London and exploring the possibilities of the amplifier as a musical instrument in the recording studio and on the concert stage. Hendrix established versatility and technical skill as a norm for rock musicianship and gave shape to a new kind of event: the outdoor festival and stadium concert, in which the noise of the audience became part of the logic of the music.

Bob Marley

Bob Marley, from Kingston, Jamaica (via London), personified a new kind of global popular music in the 1970s. Marley and his group, the Wailers, combined sweet soul vocals inspired by Chicago groups such as the Impressions with rock guitar, a reggae beat, and Rastafarian mysticism. Marley's commercial success established Jamaica as a major source of international talent, leaving a reggae imprint not just on Western rock but also on local music makers in Africa, Asia, and Australia.

Madonna, from suburban Detroit, Michigan (via New York City), personified a new sort of global teen idol in the 1980s. She combined the sounds and technical devices of the New York City disco-club scene with the new sales and image-making opportunities offered by video promotion—primarily by Music Television (MTV), the music-based cable television service. As a star Madonna had it both ways: she was at once a knowing American feminist artist and a global sales icon for the likes of Pepsi-Cola.

Public Enemy, from New York City, personified a new sort of African-American music in the late 1980s. Rap, the competitive use of rhyming lines spoken over an ever-more-challenging rhythmic base, had a long history in African-American culture; however, it came to musical prominence as part of the hip-hop movement. Public Enemy used new digital technology to sample (use excerpts from other recordings) and recast the urban soundscape from the perspective of African-American youth. This was music that was at once sharply attuned to local political conditions and resonant internationally. By the mid-1990s rap had become an expressive medium for minority social groups around the world.

What does this version of rock's history—from Presley to Public Enemy—reveal? First, that rock is so broad a musical category that in practice people organize their tastes around more focused genre labels: the young Presley was a rockabilly, the Beatles a pop group, Dylan a folkie, Madonna a disco diva, Marley and the Wailers a reggae act, and Public Enemy rappers. Even Hendrix, the most straightforward rock star on this list, also has a place in the histories of rhythm and blues and jazz. In short, while all these musicians played a significant part in the development of rock, they did so by using different musical instruments and textures, different melodic and rhythmic principles, different approaches to song words and performing conventions.

Musical eclecticism and the use of technology. Even from a musicological point of view, any account of rock has to start with its eclecticism. Beginning with the mix of country and blues that comprised rock and roll (rock's first incarnation), rock has been essentially a hybrid form. African-American musics were at the centre of this mix, but rock resulted from what white musicians, with their own folk histories and pop conventions, did with African-American music—and with issues of race and race relations.

Rock's musical eclecticism reflects (and is reflected in) the geographic mobility of rock musicians, back and forth across the United States, over the Atlantic Ocean, and throughout Europe. Presley was unique as a rock star who did not move away from his roots; Hendrix was more typical in his restlessness. And if rock and roll had rural origins, the rock audience was from the start urban, an anonymous crowd seeking an idealized sense of community and sociability in dance halls and clubs, on radio stations, and in headphones. Rock's central appeal as a popular music has been its ability to provide globally an intense experience of belonging, whether to a local scene or a subculture. Rock history can thus be organized around both the sound of cities (Philadelphia and Detroit, New York City and San Francisco, Liverpool and Manchester) and the spread of youth cults (rock and roll, heavy metal, punk, and grunge).

Sense of community

Rock is better defined, then, by its eclecticism than by reference to some musical essence, and it is better understood in terms of its general use of technology rather than by its use of particular instruments (such as the guitar). Early rock-and-roll stars such as Presley and Buddy Holly depended for their sound on engineers' trickery in the recording studio as much as they did on their own vocal skills, and the guitar became the central rock instrument because of its amplified rather than acoustic qualities. Rock's history is tied up with technological shifts in the storage, retrieval, and transmission of sounds: multitrack tape recording made possible an experimental composition process that turned the recording studio into an artist's studio; digital recording made possible a manipulation of sound that shifted the boundaries between music and noise. Rock musicians pushed against the technical limits of sound amplification and inspired the development of new electronic instruments, such as the drum machine. Even relatively primitive technologies, such as the double-deck turntable, were tools for new sorts of music making in the hands of the "scratch" deejay, and one way rock

marked itself off from other popular musical forms was in its constant pursuit of new sounds and new sound devices.

Rock and youth culture. This pursuit of the new can be linked to rock's central sociological characteristic, its association with youth. In the 1950s and early 1960s this was a simple market equation: rock and roll was played by young musicians for young audiences and addressed young people's interests (quick sex and puppy love). It was therefore dismissed by many in the music industry as a passing novelty, "bubblegum," akin to the yo-yo or the hula hoop. But by the mid-1960s *youth* had become an ideological category that referred to a particular kind of hedonism, individualism, and modernism. Whereas *youth* once referred to high-school students, it came to include college students. Moreover, rock became multifunctional—dance and party music on the one hand, a matter of serious attention and intimate expression on the other. As rock spread globally this had different implications in different countries, but in general it allowed rock to continue to define itself as youthful even as its performers and listeners grew up and settled down. And it meant that rock's radical claim—the suggestion that the music remained somehow against the establishment even as it became part of it—was sustained by an adolescent irresponsibility, a commitment to the immediate thrills of sex 'n' drugs 'n' outrage and never mind the consequences. The politics of rock fun has its own power structure, and it is not, perhaps, surprising that Madonna was the first woman to make a significant splash in rock history. And she did so by focusing precisely on rock's sexual assumptions.

Neal Preston/Corbis



Madonna, 1987.

Authenticity and commercialism. Madonna can be described as a rock star (and not just a disco performer or teen idol) because she articulated rock culture's defining paradox: the belief that this music—produced, promoted, and sold by extremely successful and sophisticated multinational corporations—is nonetheless somehow noncommercial. It is noncommercial not in its processes of production but in the motivations of its makers and listeners, in terms of what, in rock, makes a piece of music or a musician valuable. The defining term in rock ideology is authenticity. Rock is distinguished from pop as the authentic expression of a performer's or composer's feelings and the authentic representation of a social situation. Rock is at once the mainstream of commercial music and a romantic art form, a voice from the social margins. Presley's first album for RCA in 1956 was just as carefully packaged to present him as an authentic, street-credible musician (plucking an acoustic guitar on the album cover) as was Public Enemy's classic *It Takes a Nation of Millions to Hold Us Back,* issued by the CBS-backed Def Jam in 1988; Madonna was every bit as concerned with revealing her artifice as art in the 1980s as Dylan was in the '60s.

Motivations

Rock, in summary, is not just an eclectic form musically but also a contradictory form ideologically. In making sense of its contradictions, two terms are critical. The first is presence. The effect of rock's musical promiscuity, its use of technology, and its emphasis on the individual voice is a unique sonic presence. Rock has the remarkable power both to dominate the soundscape and to entice the listener into the performers' emotional lives. The second is do-it-yourself (DIY). The credibility of this commercial music's claim to be noncommercial depends on the belief that rock is pushed up from the bottom rather than imposed from the top—hence the importance in rock mythology of independent record companies, local hustlers, managers, and deejays, fanzines, and pirate radio broadcasters. Even as a multimillion-dollar industry, rock is believed to be a music and a culture that people make for themselves. The historical question becomes, What were the circumstances that made such a belief possible?

ROCK IN THE 1950S

The development of the new vocal pop star. If rock music evolved from 1950s rock and roll, then rock and roll itself—which at the time seemed to spring from nowhere—evolved from developments in American popular music that followed the marketing of the new technologies of records, radio, motion pictures, and the electric microphone. By the 1930s their combined effect was an increasing demand for vocal rather than instrumental records and for singing stars such as Bing Crosby and Frank Sinatra. Increasingly, pop songs were written to display a singer's personality rather than a composer's skill; they had to work emotionally through the singer's expressiveness rather than formally as a result of the score (it was Sinatra's feelings that were heard in the songs he sang rather than their writers'). By the early 1950s it was clear that this new kind of vocal pop star needed simpler, more directly emotional songs than those provided by jazz or theatre-based composers, and the big publishers began to take note of the blues and country numbers issued on small record labels in the American South. While the major record companies tried to meet the needs of Hollywood, the national radio networks, and television, a system of independent record companies (such as Atlantic, Sun, and Chess), local radio stations, and traveling deejays emerged to serve the music markets the majors ignored: African-Americans, Southern whites, and, eventually, youth.

Record companies

Rural music in urban settings. Selling rural American musics (blues, folk, country, and gospel) had always been the business of small rather than corporate entrepreneurs, but World War II changed the markets for them—partly because of the hundreds of thousands of Southerners who migrated north for work, bringing their music with them, and partly because of the broadening cultural horizons that resulted from military service. Rural music in urban settings became, necessarily, louder and more aggressive (the same thing had happened to jazz in the early 1920s). Instruments, notably the guitar, had to be amplified to cut through the noise, and, as black dance bands got smaller (for straightforward economic reasons), guitar, bass, and miked-up voice replaced brass and wind sections, while keyboards and saxophone became rhythm instruments used to swell the beat punched out by the drums. Country dance bands, emerging from 1940s jazz-influenced western swing, made similar changes, amplifying guitars and bass, giving the piano a rhythmic role, and playing up the personality of the singer.

Such music—rhythm and blues and honky tonk—was developed in live performance by traveling musicians who made their living by attracting dancers to bars, clubs, and halls. By the late 1940s it was being recorded by independent record companies, always on the lookout for cheap repertoire and aware of these musicians' local pulling power. As the records were played on local radio stations, the appeal of this music—its energy, humour, and suggestiveness—reached white suburban teenagers who

On the radio

otherwise knew nothing about it. Rhythm-and-blues record retailers, radio stations, and deejays (most famously Alan Freed) became aware of a new market—partying teenagers—while the relevant recording studios began to be visited by young white musicians who wanted to make such music for themselves. The result was rock and roll, the adoption of these rural-urban, black and white sounds by an emergent teenage culture that came to international attention with the success of the film *Blackboard Jungle* in 1956.

Marketing rock and roll. Rock and roll's impact in the 1950s reflected the spending power of young people who, as a result of the '50s economic boom (and in contrast to the prewar Great Depression), had unprecedented disposable income. That income was of interest not just to record companies but to an ever-increasing range of advertisers keen to pay for time on teen-oriented, Top 40 radio stations and for the development of teen-aimed television shows such as *American Bandstand*. For the major record companies, Presley's success marked less the appeal of do-it-yourself musical hybrids than the potential of teenage idols: singers with musical material and visual images that could be marketed on radio and television and in motion pictures and magazines. The appeal of live rock and roll (and its predominantly black performers) was subordinated to the manufacture of teenage pop stars (who were almost exclusively white). Creative attention thus swung from the performers to the record makers—that is, to the songwriters (such as those gathered in the Brill Building in New York City) and producers (such as Phil Spector) who could guarantee the teen appeal of a record and ensure that it would stand out on a car radio.

ROCK IN THE 1960S

A black and white hybrid. Whatever the commercial forces at play (and despite the continuing industry belief that this was pop music as transitory novelty), it became clear that the most successful writers and producers of teenage music were themselves young and intrigued by musical hybridity and the technological possibilities of the recording studio. In the early 1960s teenage pop ceased to sound like young adult pop. Youthful crooners such as Frankie Avalon and Fabian were replaced in the charts by vocal groups such as the Shirelles. A new rock-and-roll hybrid of black and white music appeared: Spector derived the mini-dramas of girl groups such as the Crystals and the Ronettes from the vocal rhythm-and-blues style of doo-wop, the Beach Boys rearranged Chuck Berry for barbershop-style close harmonies, and in Detroit Berry Gordy's Motown label drew on gospel music (first secularized for the teenage market by Sam Cooke) for the more rhythmically complex but equally commercial sounds of the Supremes and Martha and the Vandellas. For the new generation of record producer, whether Spector, the Beach Boys' Brian Wilson, or Motown's Smokey Robinson and the team of Holland-Dozier-Holland, the commercial challenge—to make a record that would be heard through all the other noises in teenage lives—was also an artistic challenge. Even in this most commercial of scenes (thanks in part to its emphasis on fashion), success depended on a creative approach to technological DIY.

Motown

The British reaction. Rock historians tend to arrange rock's past into a recurring pattern of emergence, appropriation, and decline. Thus, rock and roll emerged in the mid-1950s only to be appropriated by big business (for example, Presley's move from the Memphis label Sun to the national corporation RCA) and to decline into teen pop; the Beatles then emerged in the mid-1960s at the front of a British Invasion that led young Americans back to rock and roll's roots. But this notion is misleading. One reason for the Beatles' astonishing popularity by the end of the 1960s was precisely that they did not distinguish between the "authenticity" of, say, Chuck Berry and the "artifice" of the Marvelettes.

In Britain, as in the rest of Europe, rock and roll had an immediate youth appeal—each country soon had its own Elvis Presley—but it made little impact on national music media, as broadcasting was still largely under state control. Local rock and rollers had to make the music onstage rather than on record. In the United Kingdom musicians followed the skiffle group model of the folk, jazz, and blues scenes, the only local sources of American music making. The Beatles were only one of many provincial British groups who from the late 1950s played American music for their friends, imitating all kinds of hit sounds—from Berry to the Shirelles, from Carl Perkins to the Isley Brothers—while using the basic skiffle format of rhythm section, guitar, and shouting to be heard in cheap, claustrophobic pubs and youth clubs.

Skiffle

In this context a group's most important instruments were their voices—on the one hand, individual singers (such as John Lennon and Paul McCartney) developed a new harshness and attack; on the other hand, group voices (vocal harmonies) had to do the decorative work that had been provided on the original records by producers in the studio. Either way, it was through their voices that British beat groups, covering the same songs with the same lineup of instruments, marked themselves off from each other, and it was through this emphasis on voice that vocal rhythm and blues made its mark on the tastes of "mod" culture (the "modernist" style-obsessed, consumption-driven youth culture that developed in Britain in the 1960s). Soul singers such as Ray Charles and Sam Cooke were the model for beat group vocals and by the mid-1960s were joined in the British charts by more intense African-American singers such as Aretha Franklin and Otis Redding. British guitarists were equally influenced by this expressive ideal, and the loose rhythm guitar playing of rock and roll and skiffle was gradually replaced by more ornate lead playing on electric guitar as local musicians such as Eric Clapton sought to emulate blues artists such as B.B. King.

Clapton took the ideal of authentic performance from the British jazz scene, but his pursuit of originality—his homage to the blues originals and his search for his own guitar voice—also reflected his art-school education (Clapton was one of many British rock stars who engaged in music seriously while in art school). By the end of the 1960s, it was assumed that British rock groups wrote their own songs. What had once been a matter of necessity—there was a limit to the success of bands that played strictly cover versions, and Britain's professional songwriters had little understanding of these new forms of music—was now a matter of principle: self-expression onstage and in the studio was what distinguished these "rock" acts from pop "puppets" like Cliff Richard. (Groomed as Britain's Elvis Presley in the 1950s—moving with his band, the Shadows, from skiffle clubs to television teen variety shows—Richard was by the end of the 1960s a family entertainer, his performing style and material hardly even marked by rock and roll.)

David Redfern/Retna



The Rolling Stones, mid-1960s. From left, Brian Jones, Bill Wyman, Charlie Watts, Mick Jagger, and Keith Richards.

Folk rock, the hippie movement, and "the rock paradox." The peculiarity of Britain's beat boom—in which would-be pop stars such as the Beatles turned arty while would-be blues musicians such as the Rolling Stones turned pop—had a dramatic effect in the United States, not only on consumers but also on musicians, on the generation who had grown up on rock and roll but grown out of it and into more serious sounds, such as urban folk. The Beatles' success suggested that it was possible to enjoy the commercial, mass-cultural power of rock and roll while remaining an artist. The immediate consequence was folk rock. Folk musicians, led by Bob Dylan, went electric, amplified their instruments, and sharpened their beat. Dylan in particular showed that a pop song could be both a means of social commentary (protest) and a form of self-expression (poetry). On both the East and West coasts, bohemia started to take an interest in youth music again. In San Francisco, for example, folk and blues musicians, artists, and poets came together in loose collectives (most prominently the Grateful Dead and the Jefferson Airplane) to make acid rock as an unfolding psychedelic experience, and rock became the musical soundtrack for a new youth culture, the hippies.

The hippie movement of the late 1960s in the United States—tied up with Vietnam War service and anti-Vietnam War protests, the Civil Rights Movement, and sexual liberation—fed back into the British rock scene. British beat groups also defined their music as art, not commerce, and felt themselves to be constrained by technology rather than markets. The Beatles made the move from pop to rock on their 1967 album, *Sgt. Pepper's Lonely Hearts Club Band,* symbolically identifying with the new hippie era, while bands such as Pink Floyd and Cream (Clapton's band) set new standards of musical skill and technical imagination. This was the setting in which Hendrix became the rock musician's rock musician. He was a model not just in his virtuosity and inventiveness as a musician but also in his stardom and his commercial charisma. By the end of the 1960s the great paradox of rock had become apparent: rock musicians' commitment to artistic integrity—their disdain for chart popularity—was bringing them unprecedented wealth. Sales of rock albums and concert tickets reached levels never before seen in popular music. And, as the new musical ideology was being articulated in magazines such as *Rolling Stone,* so it was being commercially packaged by emergent record companies such as Warner Brothers in the United States and Island in Britain.

Woodstock

Rock fed both off and into hippie rebellion (as celebrated by the Woodstock festival of 1969), and it fed both off and into a buoyant new music business (also celebrated by Woodstock). This music and audience were now where the money lay; the Woodstock musicians seemed to have tapped into an insatiable demand, whether for "progressive" rock and formal experiment, heavy metal and a bass-driven blast of high-volume blues, or singer-songwriters and sensitive self-exploration.

ROCK IN THE 1970S

Corporate rock. The 1970s began as the decade of the rock superstar. Excess became the norm for bands such as the Rolling Stones, not just in terms of their private wealth and well-publicized decadence but in terms of stage and studio effects and costs. The sheer scale of rock album sales gave musicians—and their growing entourage of managers, lawyers, and accountants—the upper hand in negotiations with record companies, and for a time it seemed that the greater the artistic self-indulgence the bigger the payoff. By the end of the decade, though, the 25-year growth in record sales halted, and a combination of economic recession and increasing competition for young people's leisure spending (notably from the makers of video games) brought the music industry, by this point based on rock, its first real crisis. The Anglo-American music market was consolidated into a shape that has not changed much since, while new sales opportunities beyond the established transatlantic route began to be pursued more intently.

Challenges to mainstream rock. The 1970s, in short, was the decade in which a pattern of rock formats and functions was settled. The excesses of rock superstardom elicited both a return to DIY rock and roll (in the roots sounds of performers such as Bruce Springsteen and in the punk movement of British youth) and a self-consciously camp take on rock stardom itself (in the glam rock of the likes of Roxy Music, David Bowie, and Queen). The continuing needs of dancers were met by the disco movement (originally shaped by the twist phenomenon in the 1960s), which was briefly seized by the music industry as a new pop mainstream following the success of the film *Saturday Night Fever* in 1977. By the early 1980s, however, disco settled back into its own world of clubs, deejays, and recording studios and its own crosscurrents from African-American, Latin-American, and gay subcultures. African-American music developed in parallel to rock, drawing on rock technology sometimes to bridge black and white markets (as with Stevie Wonder) and sometimes to sharpen their differences (as in the case of funk).

Rock, in other words, was routinized, as both a money-making and a music-making practice. This had two consequences that were to become clearer in the 1980s. First, the musical tension between the mainstream and the margins, which had originally given rock and roll its cultural dynamism, was now contained within rock itself. The new mainstream was personified by Elton John, who developed a style of soul-inflected rock ballad that over the next two decades became the dominant sound of global pop music. But the 1970s also gave rise to a clearly "alternative" rock ideology (most militantly articulated by British punk musicians), a music scene self-consciously developed on independent labels using "underground" media and committed to protecting the "essence" of rock and roll from commercial degradation. The alternative-mainstream, authentic-fake distinction crossed all rock genres and indicated how rock culture had come to be defined by its own contradictions.

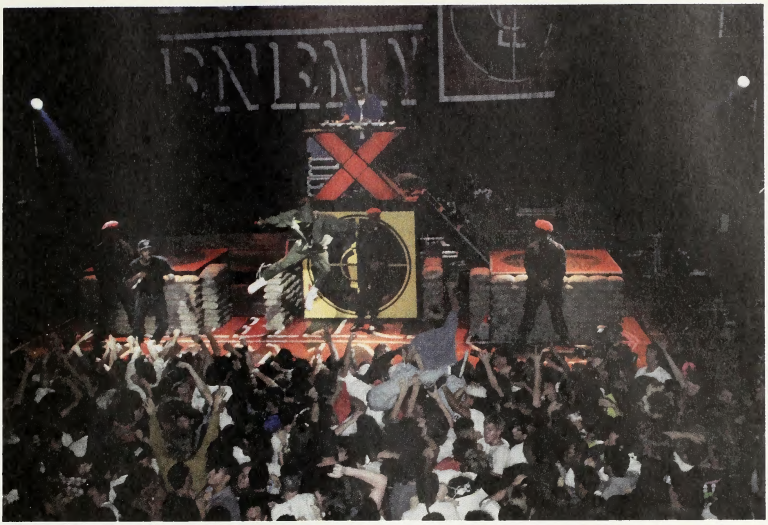
Punk

Second, sounds from outside the Anglo-American rock nexus began to make their mark on it (and in unexpected ways). In the 1970s, for example, Europop began to have an impact on the New York City dance scene via the clean, catchy Swedish sound of Abba, the electronic machine music of Kraftwerk, and the American-Italian collaboration (primarily in West Germany) of Donna Summer and Giorgio Moroder. At the same time, Marley's success in applying a Jamaican sensibility to rock conventions meant that reggae became a new tool for rock musicians, whether established stars such as Clapton and the Rolling Stones' Keith Richards or young punks like the Clash, and played a significant role (via New York City's Jamaican sound-system deejays) in the emergence of hip-hop.

ROCK IN THE 1980S AND '90S

Digital technology and alternatives to adult-oriented rock. The music industry was rescued from its economic crisis by the development in the 1980s of a new technology, digital recording. Vinyl records were replaced by the compact disc (CD), a technological revolution that immediately had a conservative effect. By this point the most affluent record buyers had grown up on rock; they were encouraged to replace their records, to listen to the same music on a superior sound system. Rock became adult music; youthful fads continued to appear and disappear, but these were no longer seen as central to the rock process, and, if rock's 1970s superstars could no longer match the sales of their old records with their new releases, they continued to sell out stadium concerts that became nostalgic rituals (most unexpectedly for the Grateful Dead). For new white acts the industry had to turn to alternative rock. A new pattern emerged—most successfully in the 1980s for R.E.M. and in the '90s for Nirvana—in which independent labels, college radio stations, and local retailers developed a cult audience for acts that were then signed and mass-marketed by a major label. Local record companies became, in effect, research and development divisions of the multinationals.

The radical development of digital technology occurred elsewhere, in the new devices for sampling and manipulating sound used by dance music engineers who had already been exploring the rhythmic and sonic possibilities of electronic instruments and blurring the distinctions between



Public Enemy performing, *c.* 1991.
© Lynn Goldsmith/Corbis

live and recorded music. Over the next decade the uses of digital equipment pioneered on the dance scene fed into all forms of rock music making. For a hip-hop act, such as Public Enemy, what mattered was not just a new palette of "pure" sound but also a means of putting reality—the actual voices of the powerful and powerless—into the music. Hip-hop, as was quickly understood by young disaffected groups around the world, made it possible to talk back to the media.

The global market and fragmentation. The regeneration of DIY paralleled the development of new means of global music marketing. The 1985 Live Aid event, in which live television broadcasts of charity concerts taking place on both sides of the Atlantic were shown worldwide, not only put on public display the rock establishment and its variety of sounds but also made clear television's potential as a marketing tool. MTV, the American cable company that had adopted the Top 40 radio format and made video clips as vital a promotional tool as singles, looked to satellite technology to spread its message: "One world, one music." And the most successful acts of the 1980s, Madonna and Michael Jackson (whose 1982 album, *Thriller,* became the best-selling album of all time by crossing rock's internal divides), were the first video acts, using MTV brilliantly to sell themselves as stars while being used, in turn, as global icons in the advertising strategies of companies such as Pepsi-Cola.

MTV

The problem with this pursuit of a single market for a single music was that rock culture was fragmenting. The 1990s had no unifying stars (the biggest sensation, the Spice Girls, were never really taken seriously). The attempt to market a global music was met by the rise of world music, an ever-increasing number of voices drawing on local traditions and local concerns to absorb rock rather than be absorbed by it. Tellingly, the biggest corporate star of the 1990s, the Quebecois Céline Dion, started out in the French-language market. By the end of the 20th century, hybridity meant musicians playing up divisions within rock rather than forging new alliances. In Britain the rave scene (fueled by dance music such as house and techno, which arrived from Chicago and Detroit via Ibiza, Spain) converged with "indie" guitar rock in a nostalgic pursuit of the rock community past that ultimately was a fantasy. Although groups such as Primal Scream and the Prodigy seemed to contain, in themselves, 30 years of rock history, they remained on the fringes of most people's listening. Rock had come to describe too broad a range of sounds and expectations to be unified by anyone.

ROCK AS A REFLECTION OF SOCIAL AND CULTURAL CHANGE

How, then, should rock's contribution to music history be judged? One way to answer this is to trace rock's influences on other musics; another is to attempt a kind of cultural audit (What is the ratio of rock masterworks to rock dross?). But such approaches come up against the problem of definition. Rock does not so much influence other musics as colonize them, blurring musical boundaries. Any attempt to establish an objective rock canon is equally doomed to failure—rock is not this sort of autonomous, rule-bound aesthetic form.

Its cultural value must be approached from a different perspective. The question is not How has rock influenced society? but rather How has it reflected society? From the musician's point of view, for example, the most important change since the 1950s has been in the division of music-making labour. When Elvis Presley became a star, there were clear distinctions between the work of the performer, writer, arranger, session musician, record producer, and sound engineer. By the time Public Enemy was recording, such distinctions had broken down from both ends: performers wrote, arranged, and produced their own material; engineers made as significant a musical contribution as anyone else to the creation of a recorded sound. Technological developments—multitrack tape recorders, amplifiers, synthesizers, and digital equipment—had changed the meaning of musical instruments; there was no longer a clear distinction between producing a sound and reproducing it.

From a listener's point of view, too, the distinction between music and noise changed dramatically in the second half of the 20th century. Music became ubiquitous, whether in public places (an accompaniment to every sort of activity), in the home (with a radio, CD player, or cassette player in every room), or in blurring the distinction between public and private use of music (a Walkman,

boom box, or karaoke machine). The development of the compact disc only accelerated the process that makes music from any place and any time permanently available. Listening to music no longer refers to a special place or occasion but, rather, a special attention—a decision to focus on a given sound at a given moment.

Rock is the music that has directly addressed these new conditions and kept faith with the belief that music is a form of human conversation, even as it is mediated by television and radio and by filmmakers and advertisers. The rock commitment to access—to doing mass music for oneself—has survived despite the centralization of production and the ever-increasing costs of manufacture, promotion, and distribution. Rock remains the most democratic of mass media—the only one in which voices from the margins of society can still be heard out loud. Yet, at the beginning of the 21st century, rock and the music industry faced a new crisis. The development of digital technology meant that music could now be stored on easy-to-use digital files, which could, in turn, be transferred from personal computer to personal computer via the Internet. The resulting legal and corporate disputes about new digital formats such as MP3 and services such as Napster reflected both new commercial opportunities (musical rights holders had visions of making money every time a song was downloaded) and fears (they could also see a situation in which their songs would be exchanged without any money changing hands at all).

Beginning in late 1999, the Recording Industry Association of America, Bertelsmann AG, and some artists sued Napster, an Internet start-up company whose "peer-to-peer" file-sharing program allowed users to download music for free. Other artists lined up on either side of the issue. By the end of the century, however, Bertelsmann had become the majority owner of Napster, anxious to provide a fee-based service.

While the issues here were new, the story line was not. Again, an emergent technology meant new commercial opportunities that were explored and developed by fledgling entrepreneurs before being absorbed and reordered by the major music companies; again, new ways of making and using music became subject to new laws and licenses. But what was most striking about the battle over Napster was the assumption that what was at issue was rock music. (Napster was not seen as a threat to classical or country music.) Significantly, the first widely successful use of MP3 technology, Napster, involved a global network of home "tapers" and drew on the rock ideology of DIY, community, and anti-commerce. Whatever Napster's fate, it ensured that rock music would be central to 21st-century ways of doing things. Rock, in short, not only reflects (and reflects on) social and cultural change; it is also a social force in its own right.

BIBLIOGRAPHY

General: There is an extensive literature on rock that ranges from academic musicology and sociology through every kind of journalism to disposable gossip and poster books. PETER VAN DER MERWE, *Origins of the Popular Style* (1989, reissued 1992), a scholarly study of pre-20th-century popular music, helps explain why a music first appearing at the margins of Western culture so quickly became the mainstream. CHARLIE GILLETT, *The Sound of the City: The Rise of Rock and Roll,* 2nd ed., newly illustrated and expanded (1996), is still the best account of how rock and roll was first shaped in a variety of local American settings. Rock and roll's roots in black and white music are covered in *Country: The Music and the Musicians: From the Beginnings to the '90s,* 2nd ed. (1994), an informative overview of country music history published by the Country Music Foundation; and CHARLES KEIL, *Urban Blues* (1966, reissued 1991), an illuminating anthropological study of African-American musical culture in the late 1950s and early 1960s.

The development of rock out of rock and roll was as much an ideological as a musical process, and the classic description of that ideology—of why and how rock drew from and came to articulate the contradictory impulses of American popular culture—is GREIL MARCUS, *Mystery Train,* 4th rev. ed. (1997), which, in its studies of particular musicians, was the first work to reveal the possibilities of rock criticism; GREIL MARCUS, *Invisible Republic: Bob Dylan's Basement Tapes* (1997), fills the biggest gap in *Mystery Train.* SIMON FRITH and HOWARD HORNE, *Art into Pop* (1987), studies how British rock sensibility was shaped by art-school ideas and practices.

SIMON FRITH and ANDREW GOODWIN (eds.), *On Record: Rock, Pop, and the Written Word* (1990), is a useful anthology of 30 years of scholarly writing on rock, from a variety of disciplinary perspectives. The best studies of the rock music industry are GEOFFREY STOKES, *Star-Making Machinery* (1976), a fine and undated piece of reportage on the making and marketing of a Commander Cody LP; ANDREW GOODWIN, *Dancing in the Distraction Factory: Music Television and Popular Culture* (1992), a lucid and thoughtful analysis of MTV's impact on rock culture; and PAUL THÉBERGE, *Any Sound You Can Imagine: Making Music/Consuming Technology* (1997), a comprehensive history of the effects of technology on music making, paying particular attention to digital technology.

Biographies: PETER GURALNICK, *Last Train to Memphis* (1994), is the definitive work on the young Elvis Presley and his influences; PETER GURALNICK, *Careless Love* (1999), provides all one needs to know about Presley's subsequent career—its triumphs and tragedies. Good accounts of the ways in which musicians have tried to make sense of rock's confusion of art, commerce, and politics can be found in the biographies of four musicians who died young: MARC ELIOT, *Death of a Rebel* (1979, reissued 1995), on the muddled life of folk-rock singer-songwriter Phil Ochs; CHARLES SHAAR MURRAY, *Crosstown Traffic* (1989), a biography of Jimi Hendrix focusing on issues of race and identity; DR. LICKS, *Standing in the Shadows of Motown: The Life and Music of Legendary Bassist James Jamerson* (1989), a loving account of the origins and influence of one of rock's most significant rhythmic stylists; and ARMOND WHITE, *Rebel for the Hell of It* (1997), on rap star Tupac (2pak) Shakur, an important reflection on music and the state of the American nation at the end of the 20th century.

Genres: The most enlightening books on particular musical genres are ANDREW HOLLERAN, *Dancer from the Dance* (1978, reissued 1990), a disco novel that captures the disco experience better than any other writing; DICK HEBDIGE, *Cut 'n' Mix: Culture, Identity, and Caribbean Music* (1987, reissued 1990), a suggestive application of cultural theory to the remarkable mobility of reggae music; JON SAVAGE, *England's Dreaming: Sex Pistols and Punk Rock* (1991; also published as *England's Dreaming: Anarchy, Sex Pistols, Punk Rock, and Beyond,* 1992), on music, suburbia, and boredom; DAVID TOOP, *Rap Attack 2: African Rap to Global Hip Hop,* rev. ed. (1991), a well-informed history of hip-hop; ROBERT WALSER, *Running with the Devil: Power, Gender, and Madness in Heavy Metal Music* (1993), the most convincing of all the musicological rock studies; SARAH THORNTON, *Club Cultures: Music, Media, and Subcultural Capital* (1995), an intelligent sociology of British dance clubs in the early 1990s; and SIMON REYNOLDS, *Generation Ecstasy: Into the World of Techno and Rave Culture* (1998), a helpful map of a confused music scene. Finally, EVELYN MCDONNELL and ANN POWERS (eds.), *Rock She Wrote* (1995), is an instructive anthology of rock writing from a female perspective; MARK SLOBIN, *Subcultural Sounds: Micromusics of the West* (1993), is an ethnomusicological study which makes clear that all popular musics, rock included, remain local even as they become global, just as in the first days of rock and roll. (S.F.)

Roman Catholicism

As both its critics and its champions would probably agree, Roman Catholicism has been the decisive spiritual force in the history of Western civilization. There are more Roman Catholics in the world than there are believers of any other religious tradition—not merely more Roman Catholics than all other Christians combined, but more Roman Catholics than all Muslims or Buddhists or Hindus. The papacy is the oldest continuing absolute monarchy in the world. To millions the pope is the infallible interpreter of divine revelation and the vicar of Christ; to others he is the fulfillment of the biblical prophecies about the coming of the Antichrist.

These incontestable statistical and historical facts suggest that some understanding of Roman Catholicism—its history, its institutional structures, its beliefs and practices, and its place in the world—is an indispensable component of cultural literacy, regardless of how one may individually answer the ultimate questions of life and death and faith. Without a grasp of what Roman Catholicism stands for, it is difficult to make political sense of the settlement of the Germanic tribes in Europe at the end of the Roman Empire, or intellectual sense of Thomas Aquinas, or literary sense of *The Divine Comedy* of Dante Alighieri, or artistic sense of the Gothic cathedrals, or musical sense of many of the compositions of Haydn or Mozart.

At one level, of course, the interpretation of Roman Catholicism is closely related to the interpretation of Christianity as such. For by its own reading of history, Roman Catholicism began with the very beginnings of the Christian movement. An essential component of the definition of any one of the other branches of Christendom, moreover, is the examination of its relation to Roman Catholicism: How did Eastern Orthodoxy and Roman Catholicism come into schism? Was the break between the Church of England and Rome inevitable? Conversely, such questions are essential to the definition of Roman Catholicism itself, even to a definition that adheres strictly to the official view, according to which the Roman Catholic Church has maintained an unbroken continuity since the days of the Apostles, while all other denominations, from the ancient Copts to the latest storefront church, are deviations from it.

Like any intricate and ancient phenomenon, Roman Catholicism can be described and interpreted from a variety of perspectives and by one or more of several methodologies. Thus the Roman Catholic Church is itself a complex institution, for which the usual diagram of a pyramid, extending from the pope at the apex to the believers in the pew, is vastly oversimplified; within that institution, moreover, sacred congregations, archdioceses and dioceses, provinces, religious orders and societies, seminaries and colleges, parishes and confraternities, and countless other institutions all invite the social scientist to the consideration of power relations, leadership roles, social dynamics, and other sociological connections that it uniquely represents. As a world religion among world religions, Roman Catholicism in its belief and practice manifests, somewhere within the range of its multicoloured life, some of the features of every religion of the human race; thus only the methodology of comparative religion can encompass them all. Furthermore, because of the normative role of Scholasticism in the formulation of Roman Catholic dogma, a philosophical analysis of its system of doctrine is indispensable even for grasping its theological vocabulary. Nevertheless, the historical method is especially appropriate to this task, not only because two millennia of history are represented in the Roman Catholic Church, but because the heart of its understanding of itself is the hypothesis of continuity and because the centre of its definition of authority is the embodiment of divine truth in that historical continuity.

The history of the early church during the first three centuries of its existence is recounted at greater length in the article CHRISTIANITY, and its general outlines must be presupposed here. The present article concentrates on identifying those historical forces that worked to transform the primitive Christian movement into a church that was recognizably "catholic," namely, a church that had begun to possess identifiable norms of doctrine and life, fixed structures of church authority, and, at least in principle, a universality (which is what "catholic" meant) that extended to all of humanity.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 812 and 827, and the *Index.*

The article is divided into the following sections:

History of Roman Catholicism 878
  The emergence of catholic Christianity 878
  The emergence of Roman Catholicism 878
    Internal factors
    External factors
    The early medieval papacy
  The church of the early and High Middle Ages 880
    The concept of Christendom
    A period of decadence
    Popular Christianity *c.* 1000
    The first reformers: Leo IX and Nicholas II
    The reign of Gregory VII
    The Investiture Conflict (1085–1122)
    The Crusades
  The church of the late Middle Ages 882
    The Proto-Renaissance
    Reformed monasticism
    The papacy at its height: the 12th and 13th centuries
    The age of faith
    The rise of heresy
    The golden age of Scholasticism
    Ecclesiastical life in the 13th century
    Troubles of the church *c.* 1300
    The "Babylonian Captivity"
  From the late Middle Ages to the Reformation 885
    Late medieval reform: the Great Western Schism and conciliarism
    Roman Catholicism on the eve of the Reformation
The Age of Reformation and Counter-Reformation 887
    Roman Catholicism and the Protestant Reformation
    The Roman Catholic Reformation
    The Counter-Reformation
    Post-Reformation conditions
    Developments in France
    Controversies involving the Jesuits
    Religious life in the 17th and 18th centuries
  The church in the modern period 891
    Catholicism in Revolutionary France
    Napoleon I—exportation of the Revolution
    The reign of Pius IX (1846–78)
    The reign of Leo XIII (1878–1903)
    The period of the world wars
    Vatican II
  Roman Catholicism outside Europe 894
    The New World: the Spanish and Portuguese empires
    Roman Catholicism in the United States and Canada
    The spread of Roman Catholicism in Africa and Asia
Structure of the church 896
  Doctrinal basis 896
    The nature of the church
    Apostolic succession
  The papacy 897
    The papal office
    Historical conceptions of papal authority within the church
    Historical conceptions of the relationship of the papacy to the world

Contemporary teaching on papal authority
The offices of the clergy 899
The Roman Curia and the College of Cardinals
The college of bishops
Ecumenical councils
The priesthood
Religious communities 901
Hermits and monks
Mendicant friars and clerks regular
Nuns and brothers
The laity 902
Canon law 902
Beliefs and practices 903
Faith 903
Concepts of faith
Preambles and motivation of faith
Heresy
Revelation 904
The concept of revelation
The content of revelation
Tradition and scripture 904
The teaching authority of the church (the magisterium) 905
The concept of teaching authority
Organs of teaching authority
Object and response
Major dogmas and doctrines 906
The liturgy 906
The eucharistic assembly or mass
The divine office
The cycle and the language of the liturgy
The sacraments 907
The sacraments in general
Baptism
Confirmation
The Eucharist
Penance
The anointing of the sick
Marriage
Holy orders
Paraliturgical devotions 910
Eucharistic devotions
Cult of the saints
Mysticism
The role of the church in society 910
Missions
Education
Eleemosynary activities
Church and state relations
Economic views and practice
The family
Roman Catholicism following the second Vatican Council 912
Bibliography 912

History of Roman Catholicism

THE EMERGENCE OF CATHOLIC CHRISTIANITY

At least in an inchoate form all the elements of catholicity—doctrine, authority, universality—are evident in the New Testament. The Acts of the Apostles begins by focusing on the demoralized band of the disciples of Jesus in Jerusalem; but by the time its account of the first decades is finished, the Christian community has developed some nascent criteria for determining the difference between authentic ("apostolic") and inauthentic teaching and behaviour. It has also moved beyond the borders of Judaism, as the dramatic sentence of the closing chapter announces: "And so we came to Rome" (Acts 28:14). The later epistles of the New Testament admonish their readers to "guard what has been entrusted to you" (1 Timothy 6:20) and to "contend for the faith which was once for all delivered to the saints" (Jude 3), and they speak about the Christian community itself in exalted and even cosmic terms as the church, "which is [Christ's] body, the fulness of him who fills all in all" (Ephesians 1:23). It is clear even from the New Testament that the specification of these catholic features was called forth by challenges from within, not only from without; indeed, scholars have concluded that the early church was far more pluralistic from the very beginning than the somewhat idealized pictures in the New Testament might suggest.

As such challenges continued in the 2nd and 3rd centuries, further specification became necessary. The schema of apostolic authority formulated by the bishop of Lyon, Irenaeus (*c.* 130–*c.* 200), may serve to set forth systematically the three main lines of authority for catholic Christianity: the Scriptures of the New Testament (alongside the Christianized "Old Testament") as the writings of the Apostles of Christ; the episcopal centres established by the Apostles as the seats of their identifiable successors in the governance of the church; and the apostolic tradition of normative doctrine as the "rule of faith" and the standard of Christian conduct. Each of the three depended on the other two for validation; one could determine which purportedly scriptural writings were genuinely apostolic by appealing to their conformity with acknowledged apostolic tradition and to the usage of the apostolic churches, and so on. This was not a circular argument but an appeal to a single catholic authority of apostolicity, in which the three elements were inseparable. Inevitably, however, there arose conflicts—of doctrine and jurisdiction, of worship and pastoral practice, and of social and political strategy—among the three sources of authority, as well as between equally "apostolic" bishops. When bilateral means for resolving such conflicts proved insufficient, there could be recourse to either the precedent of convoking an apostolic council (Acts 15) or to what Irenaeus had already called "the preeminent authority of this church [of Rome], with which, as a matter of necessity, every church should agree." Catholicism was on the way to becoming Roman Catholic.

THE EMERGENCE OF ROMAN CATHOLICISM

Internal factors. Several historical factors, some of them more prominent at one time and others at another, help to account for the emergence of Roman Catholicism from the catholic Christianity of the early church. The twin factors that would eventually be regarded as the most decisive, at any rate by the champions of the primacy of Rome in the church, were the primacy of Peter among the 12 Apostles of Christ and the identification of Peter with the church of Rome. In the several enumerations of the Apostles in the New Testament (Matthew 10:2–5; Mark 3:16–19; Luke 6:14–16; Acts 1:13) there are considerable variations, with further variations in the manuscripts; but what they all have in common is that they list (in Matthew's words) "first, Simon, who is called Peter." "But I have prayed for you," Jesus said to Peter, "that your faith may not fail; and when you have turned again, strengthen your brethren" (Luke 22:32); and again: "Feed my lambs. . . . Tend my sheep. . . . Feed my sheep" (John 21:15–17). Above all, when Christ, according to the New Testament, said to the Apostle Peter, "And I tell you, you are Peter, and on this rock [Greek *petra*] I will build my church" (Matthew 16:18), that was, according to Roman Catholic teaching, the charter of the church—*i.e.*, of the Roman Catholic Church.

The identification of this obvious "primacy" of Peter in the New Testament with the "primacy" of the church of Rome is not self-evident, since, for one thing, the same New Testament remains almost silent about a connection of Peter with Rome. The reference at the close of the Acts of the Apostles to the arrival of the Apostle Paul in Rome gives no indication that Peter was there as the bishop or even as a resident, and the epistle that Paul had addressed somewhat earlier to the church at Rome devotes its entire closing chapter to greetings for many believers in the city but fails to mention Peter's name. On the other hand, the first of the two epistles ascribed to Peter does use the phrase (presumably referring to a Christian congregation) "she who is at Babylon" (1 Peter 5:13), which was a code name for Rome. It is, moreover, the unanimous testimony of early Christian tradition that Peter, having been at Jerusalem and then at Antioch, finally came to Rome, where he was crucified (with his head down, according to Christian legend, in deference to the crucifixion of Christ); there was, however, and still is, dispute about the exact location of his grave. Writing around the end of the 2nd century, the North African theologian Tertullian (*c.* 160–

c. 225) spoke of "Rome, from which there comes even into our own hands the very authority of the apostles themselves. How happy is its church, on which apostles poured forth all their doctrine along with their blood! where Peter endures a passion like his Lord's! where Paul wins his crown in a death like that of John [the Baptist]!"

Alongside this apostolic argument for Roman primacy—and often interwoven with it—Rome was honoured because of its position as the capital of the Roman Empire: the church in the prime city ought to be prime among the churches. As the capital Rome drew visitors or tourists or pilgrims from everywhere and eventually became, for church no less than for state, what Jerusalem had originally been called, "the church from which every church took its start, the mother city [*metropolis*] of the citizens of the new covenant." Curiously, the transfer of the capital of the Roman Empire from Rome to Constantinople by the newly converted emperor Constantine in 330, which weakened Rome's civil authority, served only to strengthen its spiritual authority: the title "supreme priest [*pontifex maximus*]," which had been the prerogative of the emperor, now devolved upon the pope. The transfer of the capital also occasioned a dispute between Rome ("Old Rome") and Constantinople ("New Rome") over whether the new capital, as capital, should be entitled to a commensurate ecclesiastical preeminence alongside the see of Peter. The second ecumenical council of the church (at Constantinople in 381) and the fourth (at Chalcedon in 451) both legislated such a position for the see of Constantinople, but Rome refused to acknowledge the legitimacy of that prerogative.

It was also at the Council of Chalcedon, convoked to resolve the doctrinal controversy between Antioch and Alexandria over the person of Christ, that the council fathers accepted the formula proposed by Pope Leo I (reigned 440–461). "Peter," they declared, "has spoken through the mouth of Leo!" That was only one in a long series of occasions when the authority of Rome, sometimes by invitation and sometimes by its own intervention, served as a court of appeal in jurisdictional and dogmatic disputes that had erupted in various parts of Christendom. During the first six centuries of the church the bishop of every major Christian centre was, at one time or another, charged with heresy and convicted—except the bishop of Rome (although his turn was to come later). The titles that the see of Rome gradually assumed and the claims of primacy it made within the internal life and governance of the church were, in many ways, little more than the formalization of what had meanwhile become widely accepted practice during these first four or five centuries of its history.

External factors. In addition to the transfer of the capital from Rome to Constantinople, there were at least two other external factors at the beginning of the Middle Ages that contributed decisively to the development of Roman Catholicism as a distinct form of Christianity. One was the rise of Islām in the 7th century. During the decade following the death of the Prophet Muḥammad in 632 CE his followers captured three of the five "patriarchates" of the early church—Alexandria, Antioch, and Jerusalem—leaving only Rome and Constantinople, located at opposite ends of the Mediterranean and, eventually, also at opposite ends of the East–West Schism. The other force that encouraged the emergence of Roman Catholicism as a distinct entity was the fall of the Roman Empire and the migration into Europe of the Germanic and other tribes that were eventually to constitute its principal population. Some of them, particularly the Goths, had already become Christian before even coming into western Europe. The form of Christianity they had adopted in the 4th century was, however, by the standards of Christian orthodoxy both Eastern and Western, heretical in its doctrine of the Trinity. Therefore the future of medieval Europe belonged not to the Christian tribes but to the pagan tribes, particularly the Franks, once these had become Christian. The Christianity they accepted after their arrival was not only orthodox on the doctrine of the Trinity but it was allied with the authority of the pope. The coronation by the pope of the Frankish king Charles (Charlemagne) as Roman emperor on Christmas Day 800 clearly symbolized that alliance.

The early medieval papacy. During the centuries that marked the transition from the early to the medieval church Roman Catholicism benefited from the leadership of several outstanding popes; at least two of them—both called "the Great" by historians and "Saint" by the Roman Catholic Church—merit special consideration even in a brief article. Pope Leo I was, even for his pagan contemporaries, the embodiment of the ideal of *Romanitas* in his resistance to the barbarian conquerors. Twice in the space of a few years he was instrumental in saving Rome, from the Huns in 452, when he achieved their withdrawal to the banks of the Danube, and from the Vandals in 455, when his intercession mitigated their depredations in the city. His aforementioned intervention in the doctrinal controversy among Eastern theologians over the person of Christ and the role played by his *Tome* of 449 in the formula of the Council of Chalcedon in 451 was part of a concerted campaign to consolidate and extend the jurisdiction of the see of Rome over such remote areas as Gaul, Spain, and North Africa—a jurisdiction officially acknowledged by the Roman emperor. Pope Gregory I (reigned 590–604), more than any pope before or after him, laid the foundations for the Roman Catholicism of the Middle Ages. It was he who selected Augustine of Canterbury to bring about the conversion of England to the Christian faith and the Roman Catholic obedience. He asserted the primacy of his see over the entire church, including the patriarchate of Constantinople, and his diplomatic and political skills secured the independence of the Western Roman Catholic Church both from the Byzantine Empire and from the Germanic tribes occupying Italy. Gregory the Great was also one of the most important patrons of the Benedictine monastic movement, to which he owed a considerable part of his own spiritual upbringing (as his biography of Benedict manifests).

Nevertheless, medieval Roman Catholicism would not have taken the form it did without the conversion of the emperor Constantine in 312. As a consequence of that event Christianity moved in a few decades from an illegal

Kunsthistorisches Museum, Vienna



Pope Gregory the Great, receiving information from the Holy Spirit, represented as a dove, carved ivory book cover, *c.* 980. In the Kunsthistorisches Museum, Vienna.

to a legal to a dominant position in the Roman Empire. Henceforth every branch of Christendom had to deal with rulers who claimed to profess its faith; conversely, the character of every branch of Christendom could in considerable measure be described on the basis of its way of relating church and state. For medieval Roman Catholicism the centralization of church authority in the pope made the relation of church and state a persistent issue in the very understanding of the nature of the church itself. As the church approached the conclusion of the first millennium of its history, it had become the legatee of the spiritual, administrative, and intellectual resources of the early centuries.

Most of the preceding analysis pertains to the whole of Christendom. The Eastern Orthodox Church has almost as large a share in the developments of the early centuries as does the Roman Catholic Church, and even Protestantism looks to these centuries for its authentication. The Middle Ages may be defined as the era in which the distinctively Roman Catholic forms and institutions of the church were set. The following chronological account of medieval developments shows how these forms and institutions emerged from the context of the shared history of the early Christian centuries. (J.J.Pe.)

THE CHURCH OF THE EARLY AND HIGH MIDDLE AGES

The concept of Christendom. By the 10th century the religious and cultural community that is called Christendom had come into being. In every European state the religion of the state was Roman Catholicism. Christendom fought back against Islām in the Crusades (see below), which failed to repossess the lost territories but strengthened the unity of Christendom and rendered it conscious of its power.

The Middle Ages saw the rise of the universities and of a "Catholic" learning, sparked, oddly enough, by the transmission of Aristotle through Arab scholars. Scholasticism, the highly formalized philosophical and theological systems developed by the medieval masters, dominated Roman Catholic thought into the 20th century and contributed to the formation of the European intellectual tradition. With the rise of the universities, the threefold level of the ruling classes of Christendom was established; *imperium* (political authority), *sacerdotium* (ecclesiastical authority), and *studium* (intellectual authority). The principle that each of these three was independent of the other two within its sphere of authority had enduring consequences in Europe.

Rise of monasticism

The same period saw the growth of monasticism. One may see in this withdrawal from the world a response to the essential conflict between Christianity and Roman civilization; those who refused to accept the prevailing compromise between the religious and secular spheres could find no place in the world of the early Middle Ages. Perhaps the most remarkable feature of monasticism was that this withdrawal did not take the form of heresy or schism. Monasticism found a way of refusing the compromise without departing from the church that had made the compromise. (F.C.O./J.L.McK./Ed.)

A period of decadence. This period also revealed the possibilities of corruption within the Roman Catholic Church. Without the accumulated prestige and the precedents established by the 9th-century popes, the claim to primacy would have had difficulty in surviving the subsequent period of papal decadence. In the 870s the imperial government in Italy declined in influence, and the bishopric of Rome, along with other European bishoprics, was increasingly at the mercy of the local nobility, with spasmodic interventions by the 10th-century German emperors.

The Ottonian Empire

German kingship entered upon a new epoch in the 10th century. Under Otto I, the Great, the bishops and greater abbots were drawn into royal service and enriched with estates and counties, for which they did feudal homage. Otto conquered northern Italy and extracted from the pope an imperial coronation (962). Both he and his grandson Otto III regarded the papal territory as part of their realm; they appointed and removed popes and presided at synods. Otto III, an enlightened ruler, appointed as pope his old tutor, Gerbert of Aurillac—who took the name Sylvester II—whose brief reign (999–1003) was a shaft of light between two periods in which Roman factions dominated the papacy.

German "protection," however, had its price. When the emperor Henry III descended into Italy in 1046, deposing three rival claimants to the papacy (Sylvester III, Gregory VI, and Benedict IX) and then appointing his own candidate, Clement II (and later several successors), the Roman Church was in grave danger of becoming an imperial proprietary church, similar to those multitudinous lower churches in Europe whose royal or aristocratic owners regarded them, in accordance with age-old custom, as their own private property to be disposed of at will.

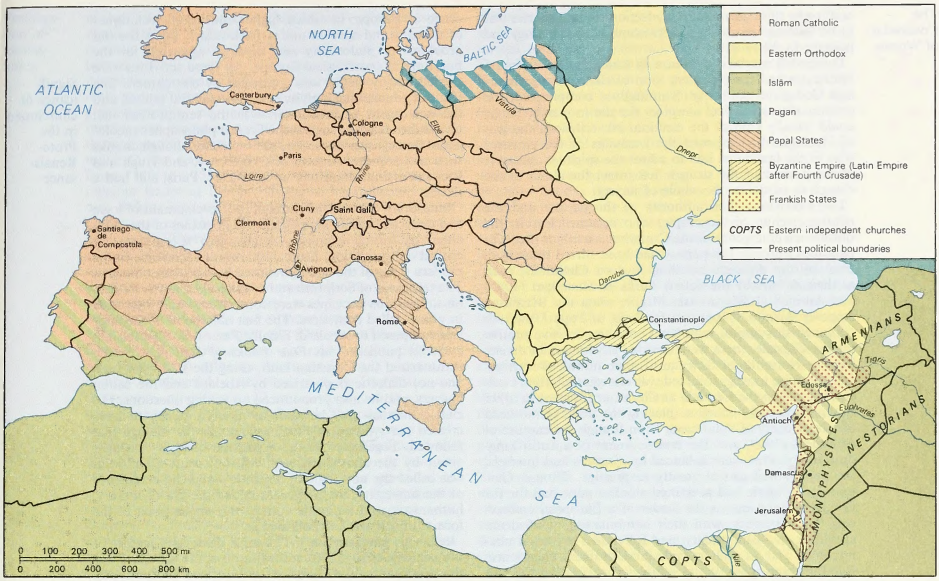
France during this period was fragmented into many feudal domains. This allowed the ecclesiastical hierarchy there a certain independence and cohesion, while the growth of the French reform-oriented monastery at Cluny prepared the country for its message of reform. In England there was a unique intermingling of ecclesiastical and royal administration that, in fact, left the church entirely free. On the fringes of Christendom—Scandinavia, Scotland, Ireland, and northern Spain—there was little hierarchical development.

Popular Christianity c. 1000. The greater part of central Christendom had by the 11th century been divided into bishops' dioceses and individual parishes. But in the northern and western regions the proliferation of small private churches had not yet been wholly absorbed, and the existence of proprietary and exempt enclaves continued to the Reformation and beyond. The priest, in rural districts usually a villein of the lord (subject to the lord but not to others), cultivated his acres of glebe (revenue lands of the parish church), celebrated mass on Sundays and feasts, recited some of the hours (liturgical or devotional services for use at certain hours of the day, according to the monastic daily schedule), and saw that his flock was baptized, anointed, and buried. Lay people normally received communion four times a year—Christmas, Easter, Pentecost, and Assumption (August 15). Auricular (privately heard) confession was widespread but not universal.

Education in the early Middle Ages was at a very low ebb outside the monasteries. Cathedral schools were few, and rural priests who could read Latin easily were rare. Almost all literary work came from the monasteries and in Celtic lands (mainly Ireland) from the half-monastic Culdees (religious recluses). The larger monasteries, such as Cluny or St. Gall (Switzerland), were towns in miniature with a variety of social services; they were also the only reservoirs of learning and artistic skill. On the land, pious practices and beliefs often merged into superstition or "white" magic; and marriage customs, together with the complicated degrees of prohibited relationships, provided endless problems in an epoch when the presence of a priest was not necessary for a valid union. In an age of protective lordship, heavenly patrons were highly valued, and the body or relics of a reputed saint made him the *persona*, a quasi-living protective presence, of a church or abbey. This aspect of belief explains the popularity of pilgrimages to shrines such as that of the Apostles at Rome, St. James at Santiago de Compostela (Spain), the Magi at Cologne (Germany), and countless others. Monastic piety was expressed not only in the liturgy but also in "little offices" (liturgical or devotional services) of the Blessed Virgin, of the cross, of all saints, and of the dead; the primary reason for a monastery's existence was intercessory prayer—hence the numerous monastic foundations by royal and noble families.

Expressions of monastic piety

The first reformers: Leo IX and Nicholas II. Leo IX (reigned 1049–54) was the first pope to impose his authority upon the church in general; he achieved this by a tactic of lengthy tours beyond the Alps, punctuated by synods, in which decrees both dogmatic and disciplinary were passed. He also began the practice of appointing non-Romans to curial (papal administrative) posts and sending legates (papal representatives) to carry out his decrees. A man of great energy and spiritual purpose, he must nevertheless bear the responsibility for a disastrous war that ended in capitulation to the Normans and for



The extent of Christianity during the period of the Crusades.
From F. W. Putzger, *Historischer Weltatlas*

choosing the rigid and violent Humbert for the mission to Constantinople in 1054, the year from which the Schism between the churches of the East and West is dated. In the years of confusion that followed, the papal election decree of Nicholas II in 1059 stands out: it gave the right and duty of papal election to the cardinals, tacitly eliminating the king of Germany. The same pope shortly afterward renewed earlier decrees on simony and clerical celibacy but avoided the issue of pope and empire.

The reign of Gregory VII. Hildebrand, who succeeded in 1073 as Gregory VII (reigned 1073–85), proved to be one of the greatest of his line and had more influence than any other person of his time upon the external fabric of the church. In his long struggle with the German king Henry IV he suspended and excommunicated his opponent, pardoned him as penitent at Canossa, Italy (1077), excommunicated him again (and was himself twice deposed), and was finally driven from Rome by Henry to die in exile at Salerno (1085). In opposition to Henry's claim to be the divinely appointed vice regent of Christ over the activities of the church, Gregory presented himself as heir to the unlimited commission of Christ to Peter over all souls (Matthew 16:18–19). Beneath these lofty claims lay the ruler's resistance to losing his ancestral right of appointing to office his most influential subjects (who often also held the richest fiefs) and the pope's insistence on the authority of ancient canon law and papal decrees. If the king's claims were inconsistent with the current conception of a free church, the pope's claim and actions were without precedent within the memory or records of his age.

Even more directly influential was Gregory's centralization of the church. Through the appointment of plenipotentiary legates (representatives with full power to negotiate), the immediate control of diocesan bishops, canonical elections, and Roman and local synods, and the publication of canonical collections and polemical manifestos a web was spun in which every thread led to Rome. The scattered priests and the distant bishops were gradually becoming a class, the clergy, distinct from others and with a law and a loyalty of their own. Although Gregory died a lonely exile, his principles of reform had found reception all over Europe, and the new generation of bishops was Gregorian in sympathy and obedient in practice to papal commands in a way unknown to their predecessors.

The Investiture Conflict (1085–1122). The efforts of the reformers to make the church independent of lay control inevitably centred upon the appointment of bishops by the ruler of the country or region. In ancient canon law, election of bishops had been by clergy and people; entrance upon office followed lawful consecration. Feudalism and royal claims had transformed election into royal appointment, and admission to office was by means of the bestowal, or investiture, by the lord, of ring and staff (symbols of the episcopal office), preceded by an act of homage. This savoured of simony, both because a layman bestowed a spiritual benefice and because money was often offered or demanded. The conservatives appealed to immemorial practice, accepted and even enjoined by the papacy.

Gregory VII, though asserting the principle of freedom, was in fact tolerant of royal appointments free from simony. Pope Urban II (reigned 1088–99) was equally inconsistent, though in other ways he was a reformer. Pope Paschal II (reigned 1099–1118) at once condemned lay investiture, thus precipitating the crisis in England between Anselm, archbishop of Canterbury, and King Henry I. This and a similar crisis in France were settled by a compromise. Election (by the cathedral chapter) was to be free; lay investiture was waived, but homage before the bestowal of the fief was allowed. Meanwhile Paschal, at odds with the German king Henry V, who was demanding imperial coronation, suddenly offered to renounce all church property held by the king if lay investiture were also abandoned. Henry accepted, but the bishops refused the terms; thereupon the King seized the Pope who, under duress, allowed lay investiture. By this time, however, a large majority of the bishops were Gregorians, and the Pope was persuaded to retract. Eleven years later Pope Gelasius II accepted the Concordat of Worms (1122). Ac-

The Concordat of Worms

cording to this agreement free election by ecclesiastics was to be followed by investiture (without staff and ring) and homage to the king.

This ended a strife of 50 years, in which pamphleteers on both sides had revived every kind of claim to supremacy and God-given authority. Nominally a compromise, the concordat was in effect a victory for the monarch, for he could usually control the election. Nevertheless, the war of ideologies had exposed the weakness of the emperor who in the last resort had to admit the spiritual authority of the pope, and the struggle left intact the claim of the church to moderate the whole of society.

The Crusades. The authority of the papacy and the relative decline of the empire also became clear in the unforeseen emergence of the Crusades as a major preoccupation of Europe. The papacy had been stirred more than once by the disasters befalling Eastern Christians, such as their defeats by the Seljuq Turks at Manzikert (1071) and Antioch (1085) in Asia Minor, when the Byzantine emperor Alexius I appealed for help to Pope Urban II. Although this appeal may have been the decisive motive for the Crusade, there were obvious advantages in diverting the Normans of Sicily and other turbulent warriors from Europe to wage a sacred war elsewhere. Urban's celebrated call to the Crusade at Clermont (France) in 1095 was unexpectedly effective, placing the pope at the head of a large army of volunteers. Even though the capture of Jerusalem (1099) and the establishment of a Latin kingdom in Palestine were balanced by disasters and quarrels, the papacy had gained greatly in prestige. Though Germany as a whole had remained aloof, a pope had for the first time stood out as the leader of a European endeavour. The Crusades, with their combination of idealism, ambition, heroism, cruelty, and folly are a medieval phenomenon and, as such, outside modern man's experience. But they were part of the religious background for two centuries and added greatly to the anxieties, both spiritual and financial, of the papacy.

THE CHURCH OF THE LATE MIDDLE AGES

The Proto-Renaissance. The 12th century, or, more correctly, the century 1050–1150, has been called the first Renaissance. A more accurate title would be the adolescence of Europe, in which higher education, techniques of thought and speech, and a fresh attack upon the old problems of philosophy and theology appeared for the first time in postclassical Europe. All these activities were carried out by clerics and controlled by churchmen. The focus of educational activity was the cathedral school, and the new agent of instruction was the semiprofessional, unattached teacher, such as the French philosopher-theologians Berengarius, Roscelin, and Abelard, though monks such as Lanfranc, Anselm of Canterbury, and Hugh and Richard of the Monastery of St. Victor, Paris, still had a share.

Significance of churchmen in the Proto-Renaissance

Philosophy was revived through the development of logic and dialectic, which were applied to doctrines of the faith, either as formal exercises, Augustinian speculation, or critical reformulation. From 1100 onward theology, in the modern sense of the word (first used by Abelard), emerged. The teachings of Scripture and of the early Church Fathers on the various doctrines were consolidated and organized in works called Sentences. The first handbook of theology was composed by Abelard. Finally, Peter Lombard (bishop *c.* 1159) published his *Four Books of Sentences,* which summarized the Christian faith, using the *sic-et-non* (yes-and-no) dialectic popularized by Abelard and the canon lawyers, and he also pronounced on vexing questions. His classic manual may be said, in modern terms, to have created the syllabus of theological study for the age that followed. Together with the expansion of logic—brought about by the arrival (through Muslim sources) of what was called the new logic of Aristotle—and the emergence of the university, the *Sentences* ended the era of literary, humanistic, and monastic culture and opened that of the formal, impersonal, Scholastic age.

Reformed monasticism. The most distinctive feature of the century 1050–1150, according to some scholars, was the appearance and diffusion of reformed monasticism. Beginning with a few relatively small quasi-hermit orders in Italy, such as the Camaldolese and the Vallombrosans, the movement spread to France with the extreme eremitical Grandmontines (founded in 1077) and the eremitical Carthusians (founded in 1084) and became as wide as Christendom with the multiplication of the daughter monasteries of Cîteaux (founded in 1098). The keynote

SCALA/Art Resource, New York City



Representation of the hierarchical order of society.
The pope enthroned as the supreme authority rules over the worldly powers and the laity (on his left) and the clergy and religious (on his right). The white and black hounds are visual puns on Dominicans—*domini canes* (hounds of the Lord). Detail of "The Church Militant and Triumphant," fresco by Andrea da Firenze, *c.* 1365; in the Spanish Chapel of the church of Santa Maria Novella, Florence.

Proliferation of monasteries

of the Cistercians (based at Cîteaux) was exact observance of the Rule of St. Benedict, with emphasis on simplicity, poverty, and manual work. The addition of lay brothers tapped a large reservoir in an age of economic and demographic expansion, and the organization of the order—with annual visitations and a general chapter—ensured good discipline and enabled the order to accommodate itself to the strain of a vast family of houses scattered throughout the Latin Church. The success of Cîteaux owed much to the genius of St. Bernard, abbot of Clairvaux from 1115 to 1153, who was for 30 years the untitled religious leader of Europe. Owing to his influence, other new orders, such as the Premonstratensians, the English Gilbertines, and the military Knights Templars, accepted or imitated Cistercian practices. All these and others had a popularity that in any other age would have seemed miraculous, since they practiced austerity. By the end of the 12th century the saturation point for monasticism had been reached all over Europe, save in a few peripheral regions, and the golden age of monasticism had passed.

The papacy at its height: the 12th and 13th centuries. Gregory VII has often been portrayed as an innovator who lacked both authentic ancestors and true successors. It must be affirmed, nonetheless, that the later history of the papacy, modern as well as medieval, was shaped by what he and his followers did, while the continuing disabilities characteristic of the medieval papacy owed much to what they left undone. Thus, the assimilation of the biblical notion of church office as grounded in love for others to the political notions of office as grounded in power and law—a development in process since the 4th century and earlier—reached a point of no return with Gregory. He functioned within a unified Christian society in which "state" and "church" were no longer conceived as distinct societal entities and was thus impelled by its very dynamic to assert a claim to jurisdictional supremacy even over the Christian emperor. For the next two centuries papal history was characterized by a deepening involvement, direct and indirect, in matters political. As a result there were, under Alexander III (reigned 1159–81) and Innocent IV (reigned 1243–54), renewed clashes with the German emperors and, under Innocent III (reigned 1198–1216), extensive and damaging papal interference in German internal affairs. What alarmed these popes was the fear that imperial policy, by encroaching upon papal territorial independence, also threatened the autonomy of papal action. But with Innocent IV, at least, such a fear was matched by his wish to vindicate, even in temporal matters, the papal claim to supremacy.

Though much of the drama of papal history in this period focused upon these conflicts, the impact that the thoroughgoing politicization of church office had upon the nature and structure of ecclesiastical government and the pope's place in it was of more enduring significance. Here again Gregory's pontificate was something of a watershed. Any lingering belief that the pope's primacy might be regarded primarily as one of honour was now dispelled, and any hesitation about implementing the jurisdictional primacy that had supplanted it now disappeared. The need for papal leadership was so widely accepted that throughout much of the 12th and 13th centuries the demand for it came from the local churches themselves. The outcome was an acceleration in the process that had led, by the late 13th century, to a papal exercise of judicial authority going far beyond the mere acceptance of appeals from lower courts; to an arrogation of the wide-ranging legislative powers manifest in the *Decretals* of Gregory IX (1234), the first officially promulgated collection of papal laws; and to the system of "papal provisions" (direct papal intervention in the disposal of benefices) that was finally to be completed by Benedict XII in 1335.

The papal monarchy

Papal leadership *in* the church was eventually replaced by papal monarchy *over* the church. Positively, this transformation was evident in the reforming legislation of the fourth Lateran Council (1215). The negative aspect was to become increasingly obvious as the 13th century wore on. It was no accident that what turned out to be the permanent schism between the Latin and Greek churches occurred at a time when Leo IX had embarked upon a more active exercise of the papal primacy. The more his successors succeeded in establishing the fullness of their jurisdictional power (*plenitudo potestatis*) within the Latin Church, the less chance there was of healing the schism. Nor did papal sponsorship of the Crusades, however great the prestige it had brought to Urban II at the time of the First Crusade, ultimately redound to the benefit of the religious life of the church.

Least justified of all was the administrative centralization attendant upon the exercise of the *plenitudo potestatis* when it was finally measured against the price that had to be paid—notably the corruption spawned by the stringent financial measures (*e.g.*, sale of indulgences, benefices, etc.) needed to support the growing army of clerical bureaucrats at Rome. And on this point one of the things left undone by the Gregorian reformers proved to be crucial. Their failure to uproot the notion of the "proprietary church" explains both the willingness of later canonists to classify the laws governing the disposition of ecclesiastical benefices under the heading not of public but of private law (law pertaining to the protection of proprietary right) and also the tendency of medieval persons in general to regard ecclesiastical office less as a focus of duty than as a source of income or an object of proprietary right. When the 13th-century popes found that direct papal taxation did not yield funds sufficient to support their bureaucrats, they adopted the practice of "providing" them to benefices all over Europe, for the law itself encouraged them to think of such benefices as sources of much needed revenue. Thus arose the characteristic abuses of pluralism (holding more than one benefice) and nonresidence against which church reformers from the mid-13th century on railed in vain and the blame for which they were soon to lay at the door of a papacy that had finally come to be regarded as an obstacle rather than a spur to reform.

The age of faith. Below the level of the papacy, however, a spiritual revival had taken place. The 12th century, perhaps more than any other, was an age of faith in the sense that all men, good or bad, pious or worldly, were fundamentally believers, and religious causes and interests (crusades, monastic foundations, building churches, and assisting education and charities) made up much of the life of the literate and administrative classes. Lay religion was, as never before or since, permeated with monastic ideals. Prodigious numbers of the populace became monks, knights (members of military-religious orders), labourers (lay brothers), and lay people who followed monastic rules, and the favourite lay devotions were short versions of monastic offices. Almost every church—whether cathedral, monastic, parochial, or private—was built or rebuilt between 1050 and 1200. Almost all baronial families founded a monastery, and townspeople not only paid for their cathedrals but often supplied materials and labour.

The mendicant orders

The pontificate of Innocent III saw the appearance of a totally new form of religious life, that of the penniless or mendicant friar. Francis of Assisi (1181/82–1226), a personality of magnetic originality who believed that he was called by Christ to preach poverty, had no thought of founding an order; but his message and his genius exactly suited his age, and the vast concourse of his followers gradually changed from a homeless, penniless band of preachers and missionaries in Italy into an international body governed by a single general and devoted to the service of the papacy. Dominic of Spain (*c.* 1170–1221), on the other hand, with a vocation to preach doctrine to heretics and with followers keeping a canonical rule, changed his existing institute into one of friars. Gradually the two groups became similar: international, articulated groups of men bound to an order but not to a community. They took the customary monastic vows of poverty, chastity, and obedience but dropped the vow of *stabilitas* (stability) in favour of mobility, and they were governed by elected superiors under a supreme chapter and general. Unpredictably, first the Dominicans and then the Franciscans entered and soon dominated the theological schools of Paris and Oxford. Two similar bodies joined them, the Carmelites and Austin Friars, and for almost a century the friars were the theologians, the preachers, and the confessors of the Christian people.

The rise of heresy. Before the middle of the 12th century heresy on a large scale was unknown in the West. The early dissenters were often radical reformers such as the Italian canon Arnold of Brescia (d. 1155), an outspoken critic of clerical wealth and corruption. Then there appeared in northern Italy and southern France the sect, Eastern and Manichaean in origin, later known as the Cathari (the "pure," from the ascetic lives of their leaders). This sect had an organization and liturgical life that imitated Christianity; but it overtly denied many key doctrines, such as the incarnation of Christ, and was dualistic in that it regarded matter and the human body as evil and the spirit as good. Its emphasis on poverty and its genuine solidarity of mutual assistance appealed to many by contrast with the luxury and wealth of the Catholic hierarchy. A little later another type of dissent appeared with the Waldenses (founded by a French reformer named Valdes) of the Rhône Valley and Piedmont. These groups, basically and professedly orthodox, together with the reform-minded Humiliati of Lombardy (Italy), practiced poverty, Scripture reading, and preaching. The Cathari were proscribed as heretics by the papacy and were attacked by a crusade and later by the Inquisition, and they gradually disappeared. The Humiliati remained orthodox as a quasi-religious order. The Waldenses, largely through mismanagement by the bishops, drifted away from the church and remained throughout the Middle Ages and after a non-Catholic body. These heretical movements, together with numerous legal disputes between monks and bishops, and bishops and metropolitans (ecclesiastical provincial leaders), imparted a sense of decline and peril to the last decades of the 12th century, which were notably barren of saints and great men. The church was too rich and too set in its hierarchical ways to meet the demands of larger populations and economic stresses, especially in urban conditions. Reformers demanded a spirit of poverty and a fresh wind of spirituality.

The golden age of Scholasticism. The 13th century was an age of fresh endeavour and splendid maturity in the realms of thought, theology, and art. Philosophy, hitherto almost exclusively devoted to logic and dialectic, had stagnated in the later 12th century. It was revived by the gradual arrival from Spain and Sicily of translations of the whole corpus of Aristotle's writings, often accompanied by Arabic and Jewish commentaries and treatises. Aristotle, especially in his *Metaphysics* and *Ethics*, opened the whole field of philosophy to the schools. After a short period of hesitation his works were used by theologians, at first eclectically and then systematically. The great German philosopher and theologian Albert of Cologne (known as Albertus Magnus) and his more famous pupil Thomas Aquinas rethought the system of Aristotle in Christian idiom, pouring into it a fair dose of Neoplatonism from St. Augustine. Aquinas, in some 25 years of work, set theology firmly on a philosophical foundation. The Italian theologian Bonaventure (1217–74), in an even shorter career, renewed the traditional approach of Augustine and the Victorine monks regarding theology as the guide of the soul to the vision of God. At the same time masters in the arts school of Paris used Aristotelian thought to present a naturalistic system that clashed with orthodox teaching. The condemnations that ensued in 1272 and 1277, coinciding with the deaths of Bonaventure and Aquinas (1274), included some Thomist theses. This apparent victory of conservatism ended the long era in which Greek thought was regarded as right reason and foreshadowed the age of individual systems and the divorce of philosophy from theology.

Revival of philosophy

Ecclesiastical life in the 13th century. The coming of the friars and the legislation of the fourth Lateran Council in Rome (1215)—including requirements of annual confession and communion and a reduction in number of the impediments to marriage—saved the lower classes for the church and silenced many of the critics of the establishment. Well-trained and extremely mobile, the friars were able to reach and hold regions and peoples that the static monks and clergy had failed to move. The 13th century in Europe as a whole was a time of pastoral endeavour in which bishops and university-trained clergy perfected the diocesan and parish organization and reformed many abuses. It was an age of active and spiritual bishops, many of them masters in theology and themselves friars. There also were controversies. The early friars served and were welcomed by the bishops and parish clergy, but clashes soon occurred; the papacy gave the friars exemptions and privileges so wide that the basic rights of the secular clergy were threatened. An academic war of pamphlets led to an attack on the vocation and work of the friars. A compromise was finally arranged by Boniface VIII (reigned 1294–1303) that was just and workable; under a revised form it lasted for two centuries. The bishop could refuse friars entry into his diocese, but once they had been admitted, the friars were free from his control.

Troubles of the church c. 1300. The last quarter of the 13th century was a time of growing bitterness and harshness. The golden age of Scholastic theology had come to an abrupt end. The troubles of the Franciscans—divided into those who stood for the absolute poverty prescribed by the rule and testament of Francis (the Spirituals) and those who accepted papal relaxation and exemptions (the Conventuals)—were a running sore for 60 years, vexing the papacy and infecting the whole church. The Inquisition (the ecclesiastical tribunal instituted in 1229 to deal with heretics) and the papal court incurred odium for their inhumane and inequitable treatment of those suspected of heresy.

Another instance of hardening sentiment is seen in the treatment of the Jews. Between 800 and 1200 the Jewish population had increased significantly in Lombardy, Provence, and the towns of the river valleys of the Rhône, the Rhine, and the Danube. They entered England only after the Norman Conquest (1066.) Apart from heretics such as the Cathari they were the only "foreign body" in Western Christendom and as such attracted the special notice of the ignorant and brutal. There were shocking massacres of Jews when the Crusades were preached, especially in the Rhineland, and after various instances of panic on the part of Christians, Jews were accused of sacrilege and child murder. These, however, were all mob movements, resisted by kings and bishops. Later the Jews suffered from suspicions that were aroused by the Cathari. The fourth Lateran Council gave the Jews a distinguishing badge and forbade their employment by governments. This established once and for all the ghetto system in large towns but did not at first impair Jewish prosperity. Later on the growing class of Christian merchants became jealous and hostile, and in 1290 and 1306 the Jews were expelled from England and France. This swelled their numbers in Germany, thenceforward called "the classic land of Jewish martyrdom." Groups remained in Italy, and the Roman colony was never disturbed. In Spain toleration gave way to widespread persecution and conversion under duress, which left a heritage of sorrow for the future.

Persecution of the Jews

The "Babylonian Captivity." In 1303, despite its resounding claims and its complex governmental machinery, the prestige of the papacy had fallen so low that it was possible for mercenaries in French pay and under French leadership to harass and humiliate the pope with impunity; Boniface VIII, at Anagni was arrested in his own family (Caetani) palace. The aftermath of this "outrage of Anagni" was the "Babylonian Captivity"—the desertion of Rome by the popes and their long residence (1309–77) at Avignon, Fr.—so called after the 70 years of Jewish exile in Babylon in the 6th century BC.

The disputes of the Franciscans, which had crystallized finally upon the teaching of the Spiritual Franciscans that their absolute poverty was that of Christ, were harshly settled (1322) by the irascible octogenarian John XXII (reigned 1316–34). A group of Franciscans, however, led by Michael of Cesena, general of the order, and William of Ockham, became bitter and formidable critics of the papacy. With them for a time was the Italian political philosopher Marsilius of Padua, a Paris master who, in his *Defensor pacis* (1324), outlined a secular state in which the church was a government department, the papacy and episcopate human institutions, and the spiritual sanctions of religion relegated to a position of honourable nonentity.

Advocacy of a secular state

Between them, Ockham and Marsilius used almost all the arguments that have ever been devised against the papacy. Condemned more than once, Marsilius had little immediate effect or influence, but during the Great Schism of the papacy (1378–1417) and later, in the 16th century, he and Ockham had their turn.

With the papacy "in captivity" and Nominalism capturing the universities, Europe and the church entered upon an epoch of disasters, of which the Hundred Years' War between England and France (began 1337) and the Black Death (1348–49) were the most clearly seen by contemporaries. For all this, Christian life in the first half of the 14th century changed little. Many of the largest parish churches of Europe date from this time, as do many popular devotions, prayers, hymns, and carols; also, many hospitals and almshouses were founded. Though the relations between the friars and the secular clergy had been canonically settled, friction continued. The friars came under wider criticism for worldliness and immorality, but they remained popular. Though heresy and antisacerdotal (anticlerical) sentiment became almost endemic in the cities of Belgium and the Netherlands, the 14th century produced some of the greatest mystical writers of the church's history: Johann Tauler and Jan van Ruysbroeck in the north, Catherine of Siena in Italy, and the anonymous author of *The Cloud of Unknowing* and Walter Hilton in England.

Reestablishment of Catholic practices in Spain

The missionary enterprise during the period 1000–1350 involved three principal fields of work: Spain, central Europe, and Asia. In Spain the absorption of the Mozarabic Church (the Arabic term for Spanish Christians under Moorish rule) and the reestablishment of Catholic practices was accomplished by Spaniards who followed the crusade ideal and by volunteers, partly monastic, from beyond the Pyrenees. In central Europe, Pope Sylvester II (reigned 999–1003) had founded the ecclesiastical hierarchies of Hungary and Poland. The region between these countries and Germany was gradually conquered and Christianized by neighbouring bishops and German missionaries. The Baltic lands were won by a mixture of preaching and the swords of the Teutonic Knights (a military monastic order) between 1100 and 1400. Purer in motive and magnificent in design were the efforts of the Franciscans and Dominicans in the Middle and Far East. Both orders preached to the Muslims, and early in the 13th century the Franciscans were in Georgia and Persia and the Dominicans in Syria. In mid-century the Franciscans penetrated Mongolia and established a church in China with an archbishop and 10 suffragan bishops, and under John XXII there was a hierarchy in Persia. All this might well have endured, had not the last of the great invasions (1383), under the Turkic conqueror Timur, or Tamerlane, broken all links between Europe and the East.

(M.D.K./F.C.O./Ed.)

FROM THE LATE MIDDLE AGES TO THE REFORMATION

The most decisive—and the most traumatic—era in the entire history of Roman Catholicism was the period from the middle of the 14th to the middle of the 16th century. This was the time when Protestantism, through its definitive break with Roman Catholicism, arose to take its place on the Christian map. It was as well the period during which the Roman Catholic Church, as an entity distinct from other "branches" of Christendom, even of Western Christendom, came into being. There is therefore much to be said for the thesis that Roman Catholicism in the form in which it is known today is, in many fundamental ways, a product of the Reformation.

Late medieval reform: the Great Western Schism and conciliarism. Reformation of the church and the papacy was what the advocates of a return of the papacy from Avignon to Rome had in mind. In the pope's absence, both the ecclesiastical and the territorial authority of the papacy had deteriorated within Italy itself, and the moral and spiritual authority of the papacy was in jeopardy throughout Christian Europe. This condition, so many believed, would continue and even worsen so long as the papacy remained in Avignon. Pope Urban V (reigned 1362–70) attempted to reestablish the papacy in Rome in 1367, but after a stay of only three years he returned to Avignon, only to die soon after his return. It was finally Gregory XI (reigned 1370–78) who, in 1377, permanently moved the papal headquarters back to Rome; but he died only a few months later. The immediate result of the return to Rome was the very opposite of the restoration of confidence and credibility that, for differing reasons, the prophetic voices and the political calculations of the 14th century had predicted would come from it. For not only had the church during its residence in Avignon come under the political and religious domination of France, which resisted the repatriation of the papacy to Italy, but the weakness of the papacy in Avignon had enabled the college of cardinals and the papal bureaucracy to fill the administrative vacuum by developing a pattern of government that can only be described as oligarchic. The powers that the cardinals had succeeded in appropriating were difficult for the centralized authority of the papacy, whether in Avignon or in Rome, to reclaim for itself.

Meeting in Rome for the first time in nearly a century, the college of cardinals elected Pope Urban VI (reigned 1378–89). But his desire to reassert the monarchical powers of the papacy, as well as his evident mental illness, prompted the cardinals to renege on that choice later in the same year. In his stead they elected Clement VII (reigned 1378–94), who soon thereafter took up residence back in Avignon. (This Clement VII is officially listed as an antipope, and the name was later taken by another pope, Clement VII [reigned 1523–34].) The years from 1378 to 1417 count as the time of the Great Western Schism, so identified to distinguish it from the no less great East–West Schism. The Great Western Schism divided the loyalties of Western Christendom between two popes, each of whom excommunicated the other and all of the other's followers. In the conflict between them, kingdoms, dioceses, religious orders, parishes, even families were split; and the pretensions of a church that claimed to be, as the Nicene Creed said, "one, holy, catholic, and apostolic" were seen as a mockery, since the empirical church—whichever it was—was in fact none of these. No one could be absolutely certain about the validity of the sacraments if the integrity and very unity of the church, and therefore of the episcopate, and therefore of the priesthood, were in doubt. Speaking for a broad consensus, the University of Paris proposed three alternatives for resolving the crisis of the institution, which had now become, for laity and clergy alike, a crisis of faith: resignation by both popes, with the election of a single unchallenged successor; adjudication of the dispute between the two popes by some independent tribunal; or appeal to an ecumenical council, which would function as a supreme court with jurisdiction over both claimants.

Council of Constance

The third of these, the summoning of a general church council, seemed to the theologians at Paris and to many others to be the preferable route. The first of several reform councils was held at Pisa in 1409 to deal with the schism and with the many other problems of discipline and doctrine that had arisen. Pisa elected Alexander V (reigned 1409–10) as pope in place of both incumbents. But, because neither of the other two would acknowledge the authority of the council and resign, the immediate result was that for a few years, as one cardinal said, the church was treated to "a simulacrum of the Holy Trinity"—the spectacle of three popes. That spectacle and the Great Western Schism itself came to an end through the work of the Council of Constance (1414–18). In addition to the settlement of the question of papal legitimacy, Constance enacted legislation on a variety of reform issues. Among others it stipulated that thenceforth, as a matter of church law, the church council was not to be seen as an expedient to be resorted to in an emergency but as a standing legislative body, a kind of ecclesiastical senate that should meet at brief and regular intervals. The decree of the Council of Constance justified this provision on the principle that the authority of the ecumenical council as the true representative of the entire church was superior to that of the pope, who could not make a similar claim for himself apart from the council. In oversimplified form, this elevation of conciliar over papal authority may be

taken as the central tenet of the late medieval movement called conciliarism.

This action also helps to account for the ambiguous position of the Council of Constance in the history of later Roman Catholic canon law, with opinions of canonists and historians differing to this day about which sessions of the council are entitled to the status of a true ecumenical council. An ambiguity even more complex attended the next of the reform councils, which used to be known in history books as the Council of Basel-Ferrara-Florence but is now sometimes divided into two councils, that of Basel and that of Ferrara-Florence, with the legitimacy of the Council of Basel contested in whole or at least in part. The council opened at Basel in 1431, was transferred by the pope to Ferrara in 1438 (although a substantial portion of its membership remained in Basel, continued discussing and legislating, and was eventually excommunicated as schismatic), moved to Florence in 1439, and held its closing sessions at Rome in 1443–45. While still at Basel, the council reaffirmed the conciliarist teaching of Constance about the superiority of the council to the pope.

Both the Council of Constance and the Council of Florence have additional importance in the history of late medieval reform in Roman Catholicism: Constance for dealing with the problem of heresy within the Western Church, Florence for addressing itself to the relation of Western Roman Catholicism to Eastern Christendom.

Jan Hus. A major item on the agenda of the Council of Constance was the challenge posed to the authority of both contending parties, council as well as pope, by the teachings of the Czech preacher and reformer Jan Hus (*c.* 1372–1415) in Prague. In every century of the Middle Ages there had been calls for reform in the church, and in times of moral corruption or of administrative chaos such calls inevitably became more intense. But the Hussite movement proved to be more than just another protest. It was animated by a definition of the church, rooted in the Augustinian tradition, that drew a sharp distinction, if not quite a disjunction, between institutional Christendom as headed by the pope and the true church as headed by Christ. The true church consisted only of those who had been predestined for membership by God and who were true believers and saints; no hypocrite, even one in the highest ecclesiastical position, could belong to that true church.

Despite the accusations of his critics, it seems clear that Hus did not draw from this premise the radical conclusion that sacraments administered by a hypocritical priest or bishop or pope were invalid in themselves; the priestly office and the sacraments retained their objective validity. A prominent element of the Hussite demands, however, was a call for the administration of Holy Communion to the laity "under both kinds—bread *and* wine—[*sub utraque specie*]," that is, they demanded the restoration of the chalice; the followers of Hus emblazoned a chalice on their banners. The Hussite program of reform coalesced with the rising nationalism of the Czech people, many of whom saw in the Roman Catholic Church a symbol of Italian and German domination.

In 1411 Hus was excommunicated by Pope John XXIII (reigned 1410–15), now identified as an antipope, but in keeping with the widespread spirit of conciliarism he appealed his case to an ecumenical council of the church. Therefore he was summoned to appear before the Council of Constance and was promised a safe-conduct by Sigismund (1368–1437), the Holy Roman emperor. Once at the council, however, Hus was arrested and incarcerated. He was tried for heresy (particularly because of his doctrine of the church) and condemned, and on July 6, 1415, he was put to death. His main prosecutors were also the leaders of the reform movement at the Council of Constance, notably Jean de Gerson (1363–1429), chancellor of the University of Paris. The death of Hus was not, however, the end of his movement. A principal difference between Hus and most other medieval reformers was that while they and their followers remained (though sometimes just barely) within the boundaries of Roman Catholicism, the outcome of his agitation was in fact the founding of a new church, one that continued to exist outside the structure of Roman Catholicism. In this respect, as well as in various specific doctrinal and moral teachings, he anticipated the development of the Protestant Reformation a century later, and his 16th-century disciples saw that development as a vindication of his and their position.

Efforts to heal the East–West Schism. At Basel, and then especially at Florence, there were extensive negotiations and discussions over the newly revived proposals for effecting a reunion of the Eastern Orthodox Church and Western Roman Catholicism. Earlier attempts at such a reunion, for example at the Council of Lyon in 1274, had failed. But now the time seemed ripe on both sides for a new effort at negotiation and reconciliation. Christian Constantinople was under increasing threat from the Turks and wanted Western support, moral as well as military. Leaders of the West, regardless of party, saw the prospect of achieving a long-sought rapprochement with the East as a means of restoring the prestige of both the papacy and the ecumenical council, which could then be seen as having resolved both of the major schisms of Christian history—the Great Western Schism and the East–West Schism—in the space of one generation. The patriarch of Constantinople, Joseph II (*c.* 1360–1439), and the Byzantine emperor, John VIII Palaeologus (1391–1448), both came in person to the Council of Florence for the theological negotiations pointing toward reunion of the two churches.

In the course of the doctrinal discussions between Greeks and Latins all the major points of difference that had historically separated the two churches received detailed attention. The Greeks acknowledged the primacy of the pope, and the West acknowledged the right of the East to ordain married men into the priesthood. The chief sticking point, as always, was the doctrine of the *Filioque:* Did the Holy Spirit in the Trinity proceed from the Father only, as the East taught, or "from the Father and the Son [*ex Patre Filioque*]," as the Western addition to the text of the Nicene Creed affirmed? At stake here was not only the dogmatic Trinitarian question itself, over which the disputes between the Latins and the Greeks had been raging since the 9th century, but the authority of one part of the church, viz., the Roman Catholic Church, to make an alteration in the text of an ecumenical creed through unilateral action, that is, without the sanction of a truly ecumenical council representing the entire church. Almost all those present at Florence came to an agreement that the dispute over the *Filioque* was chiefly one of words, not of content, since it could be amply documented that both versions of the doctrine of the procession of the Holy Spirit had substantial attestation from the teachings of the Church Fathers in both churches. Agreement on the *Filioque* and on all other points at issue led to the adoption of a document of union, *Laetentur Coeli,* promulgated on July 6, 1439 (and still commemorated in a plaque on the wall of the Duomo in Florence). But the reunion came too late for both sides. It was repudiated in the East, both at Constantinople and in the other Orthodox churches, notably the Church of Russia; and it was soon evident that in the West the internal problems of the church and the papacy had not been laid to rest by this temporary victory. Once again, as so many times throughout Christian history, the reunion of the Eastern and the Western Churches proved to have been a dead letter and an unattainable goal.

Roman Catholicism on the eve of the Reformation. *The decline of Scholastic theology.* The transition from the Middle Ages to the Reformation was a gradual one, but—at least in hindsight—its direction seems to have become clear already in the 14th and especially in the 15th century. One development that was both a cause and a result of that transition was the decline of Scholastic theology. As practiced, albeit with great divergence of opinion on many issues, by its leading expositors, Thomas Aquinas and Bonaventure, Scholasticism had been the systematization of the Roman Catholic understanding of the relation between the claims of human reason and the authority of divine revelation. To that end it had made use of philosophy, particularly of the newly available works of Aristotle, to describe the natural potentialities of human ways to

truth in order then to enthrone Christian theology as "the queen of the sciences."

With good reason have historians seen in that schema of reason and revelation the counterpart in the life of the mind to the schema of church and society set forth, earlier in the century of Aquinas and Bonaventure, by Pope Innocent III (reigned 1198–1216). These historians draw a similar correlation between the waning prestige of the papacy in the late Middle Ages and the shattering of the Scholastic synthesis by the work of such philosophical theologians as William of Ockham. Some of the theological descendants of Bonaventure, less confident of the powers of human reason than he, elevated the primacy of faith and the authority of Scripture to an almost exclusive position as a way to truth, while some of the philosophical descendants of Aquinas appeared, at least to their critics, to be expanding the realm of what was knowable by natural means to the point that the primacy of faith was threatened by an all-engulfing rationalism. All the varieties of Scholastic teaching, moreover, were under attack from those leaders of late medieval Roman Catholic piety who contended that the crisis of faith and of the church called for a return to the authentic religious experience of the primitive church as set forth in the New Testament.

Expressions of spirituality and folk piety. Late medieval spirituality cannot be dismissed as merely a symptom of the general malaise in Roman Catholic Christendom; it must be recognized as a dynamic force. One of its noblest monuments, the devotional manual entitled *The Imitation of Christ,* became, second only to the Bible itself, the most widely circulated book in Christian history. Traditionally attributed to Thomas à Kempis in 1441, the *Imitation,* although impeccably orthodox in its doctrinal emphases, took the reader beyond (or behind) the authoritative structures of both church and dogma to the inner meaning of the gospel and the inner life of the believing heart: the Christ of the creeds was, above all, the Christ of the Gospels, who summoned his followers not only to orthodoxy in their theology but to discipleship in their lives. The author of the *Imitation* participated in the spiritual life and discipline of the Brethren of the Common Life, one of the many lay communities, both female and male, that sprang up during the 15th century as centres for the cultivation of authentic Christianity even in the midst of ecclesiastical corruption and theological sterility.

The Imitation of Christ

Other expressions of folk piety, too, were flourishing on the eve of the Reformation. Partly as a continuing effect of the establishment of the orders of friars in the 12th and 13th centuries, there was a revival of interest in preaching throughout Roman Catholic Europe. Along with it developed a growing attention to the Bible, which for the first time began to circulate widely, also in vernacular translations, as a consequence of the invention of printing. The 15th century is also in many ways the high point in the history of Roman Catholic devotion to the Virgin Mary. At the same time there is also evidence among the common people of a tide of anticlericalism, much of it in reaction to the corruption of the church and the clergy, and of a growing skepticism among intellectuals and secular rulers even about fundamental Roman Catholic teachings.

Roman Catholicism and Renaissance humanism. At least some of that skepticism arose within the intellectual and literary milieu of Renaissance humanism, whose relation to Roman Catholicism was far more complex than has often been supposed. The efforts of 19th-century historians of the Renaissance—many of whom were themselves under the influence of both anticlericalism and skepticism—to interpret humanism as a neopaganism in revolt against traditional Christian beliefs have been fundamentally recast by modern scholarship. Not only were many of the popes during the 15th and 16th centuries themselves devotees and patrons of Renaissance thought and art, but a Renaissance figure like Nicholas of Cusa, arguably the greatest mind in Christendom East or West during the 15th century, was at the same time a metaphysician of astonishing boldness and creativity, an ecumenical theologian looking for points of contact not only with other Christians but even with Islām, and a reform cardinal of the Roman Catholic Church.

Thus in the light of recent study the humanists emerge as Christians who were working simultaneously for the reform of the church and of literary culture. To achieve those ends, they urged a return to the basics of Christian civilization, that is, to the Greek and Latin classics and to the monuments of biblical and patristic literature. Lorenzo Valla in Italy and then Desiderius Erasmus in the North are by no means isolated cases among the humanists for this blending of Christianity and classical culture. Erasmus ridiculed the Scholastics for their philosophical abstractions and for their bad Latin, and in his anonymous satire *Julius exclusus e coelis* he lampooned the effort of Pope Julius II (reigned 1503–13) to get into heaven. Erasmus also edited the writings of most of the major Church Fathers in both Latin and Greek. His edition of the Greek New Testament, the *Novum instrumentum* of 1516, was intended to stimulate a renewal of authentic Christian faith and life, which he himself called "the philosophy of Christ," in a corrupt Roman Catholicism. Significantly, this merciless critic of the current state of Roman Catholicism nevertheless found it impossible to affiliate himself with the Protestant Reformation when it arose, and he died a faithful, if unappreciated, member of the Roman Catholic Church.

Roman Catholicism and the emergence of national consciousness. As it had done since the time of the emperor Constantine, the relation of church and state shaped much of the history of Roman Catholicism on the eve of the Reformation. In most of the states constituting Western Christendom the 15th century was the time of an awakening of national consciousness, whose particularity and regionalism could set it into opposition with the universalism of a world church. In the Protestant Reformation of the 16th century such opposition between nation and church was to lead to a break with Roman Catholicism as such; but it is evident from the examples of 15th-century France and Spain that it could also lead to the alternative of a national Catholicism that remained in communion with Rome. As the seat of the Avignon papacy and the stronghold of the conciliarism represented by Chancellor Jean de Gerson and Cardinal Pierre d'Ailly, 15th-century France stood for just such a definition of Catholicism; and in the Pragmatic Sanction of Bourges of July 7, 1438, the French clergy came out in support of what were taken to be the historical rights of the Gallican Church to administer its own affairs independently of Rome while maintaining its ties of filial loyalty and doctrinal obedience to the Holy See.

National Catholicism

A few decades later, in 1469, the marriage of King Ferdinand of Aragon and Queen Isabella of Castile effected the union of Catholic Spain. In 1482 Ferdinand and Isabella concluded a concordat with the Holy See, under whose terms the Spanish crown retained the right to nominate candidates for the episcopate. Queen Isabella's father confessor, the humanist educator, Roman Catholic primate of Spain, and grand inquisitor, Francisco Jiménez de Cisneros (1436–1517), blended Spanish patriotism, Renaissance scholarship, and a strictly orthodox Roman Catholicism in a form that was to characterize the church in the Hispanic lands of both the Old and the New World for centuries to come.

The Age of Reformation and Counter-Reformation

The spectre of many national churches supplanting a unitary Catholic Church became a grim reality during the age of the Reformation. What neither heresy nor schism had been able to do before—to divide Western Christendom permanently and irreversibly—was done by a movement that confessed a loyalty to the orthodox creeds of Christendom and professed an abhorrence for schism. By the time the Reformation was over, Roman Catholicism had become something different from what it had been in the early centuries or even in the later Middle Ages.

Roman Catholicism and the Protestant Reformation. Whatever its nonreligious causes may have been, the Protestant Reformation arose within Roman Catholicism; there both its positive accomplishments and its negative ef-

fects had their roots. The standing of the church within the political order and the class structure of western Europe had been irrevocably altered in the course of the later Middle Ages. Thus the most extravagant claims put forward for the political authority of the church and the papacy, as formulated by Pope Boniface VIII (reigned 1294–1303), had come just at the time when such authority was in fact rapidly declining. By the time Protestantism arose to challenge the spiritual authority of the papacy, therefore, there was no longer any way to invoke that political authority against the challenge. The medieval class structure, too, had undergone fundamental and drastic changes with the rise of the bourgeoisie throughout western Europe; it is not a coincidence that in northern Europe and Britain the middle class was to become the principal bulwark of the Protestant opposition to Roman Catholicism. The traditional Roman Catholic prohibition of any lending of money at interest as "usury," the monastic glorification of poverty as an ascetic ideal, and the Roman Catholic system of holidays as times when no work was to be done were all seen by the rising merchant class as obstacles to financial development.

Accompanying these sociopolitical forces in the crisis of late medieval Roman Catholicism were spiritual and theological factors that also helped to bring on the Protestant Reformation. By the end of the 15th century there was a widely-held impression that the resources for church reform within Roman Catholicism had been tried and found wanting: the papacy refused to reform itself, the councils had not succeeded in bringing about lasting change, and the professional theologians were more interested in scholastic debates than in the nurture of genuine Christian faith and life. Such sentiments were often oversimplified and exaggerated, but their very currency made them a potent influence even when they were mistaken (and they were not always mistaken). The financial corruption and pagan immorality within Roman Catholicism, even at the highest levels, reminded critics of "the abomination of desolation" spoken of by the prophet Daniel, and nothing short of a thoroughgoing "reformation in head and members [*in capite et membris*]" seemed to be called for.

These demands were in themselves nothing new, but the Protestant Reformation took place when they coincided with, and found dramatic expression in, the highly personal struggle of one medieval Roman Catholic. Martin Luther asked an essentially medieval question: "How do I obtain a God who is merciful to me?" He also tried a medieval answer to that question by becoming a monk and by subjecting himself to fasting and discipline—but all to no avail. The answer that he eventually did find, the conviction that God was merciful not because of anything that the sinner could do but because of a freely given grace that was received by faith alone (the doctrine of justification by faith), was not utterly without precedent in the Roman Catholic theological tradition; but in the form in which Luther stated it there appeared to be a fundamental threat to Catholic teaching and sacramental life. And in his treatise *The Babylonian Captivity of the Church*, issued in 1520, Luther denounced the entire system of medieval Christendom as an unwarranted human invention foisted on the church.

Although Luther in his opposition to the practice of selling indulgences was unsparing in his attacks upon the moral, financial, and administrative abuses within Roman Catholicism, using his mastery of the German language to denounce them, he insisted throughout his life that the primary object of his critique was not the life but the doctrine of the church, not the corruption of the ecclesiastical structure but the distortion of the gospel. The late medieval mass was "a dragon's tail," not because it was liturgically unsound but because the medieval definition of the mass as a sacrifice offered by the church to God—not only, as Luther believed, as a means of grace granted by God to the church—jeopardized the uniqueness of the unrepeatable sacrifice of Christ on Calvary. The cult of the Virgin Mary and of the saints diminished the office of Christ as the sole mediator between God and the human race. Thus the pope was the Antichrist because he represented and enforced a substitute religion in which the true church, the bride of Christ, had been replaced by—and identified with—an external juridical institution that laid claim to the obedience due to God himself. When, after repeated warnings, Luther refused such obedience, he was excommunicated by Pope Leo X in 1521.

Until his excommunication Luther had gone on regarding himself as a loyal Roman Catholic and had appealed "from a poorly informed Pope to a Pope who ought to be better informed." He had, moreover, retained an orthodox Roman Catholic perspective on most of the corpus of Christian doctrine, not only the Trinity and the two natures in the person of Christ but baptismal regeneration and the Real Presence of the body and blood of Christ in the Eucharist. Many of the other Protestant Reformers who arose during the 16th century were considerably less conservative in their doctrinal stance, distancing themselves from Luther's position no less than from the Roman Catholic one. Thus Luther's Swiss opponent, Ulrich Zwingli, lumped Luther's sacramental teaching with the medieval one, and Luther in turn exclaimed: "Better to hold with the papists than with you!" John Calvin was considerably more moderate than Zwingli, but both sacramentally and liturgically he broke with the Roman Catholic tradition. The Anglican Reformation strove to retain the historical episcopate and, particularly under Queen Elizabeth I, steered a middle course, liturgically and even doctrinally, between Roman Catholicism and continental Protestantism.

The polemical Roman Catholic accusation—which the mainline Reformers vigorously denied—that these various species of conservative Protestantism, with their orthodox dogmas and quasi-Catholic forms, were a pretext for the eventual rejection of most of traditional Christianity, seemed to be confirmed with the emergence of the radical Reformation. The Anabaptists, as their name indicated, were known for their practice of "rebaptizing" those who had received the sacrament of baptism as infants; this was, at its foundation, a redefinition of the nature of the church, which they saw not as the institution allied with the state and embracing good and wicked members but as the community of true believers who had accepted the cost of Christian discipleship by a free personal decision. Although the Anabaptists, in their doctrines of God and Christ, retained the historical orthodoxy of the Nicene Creed while rejecting the orthodox doctrines of church and sacraments, those Protestants who went on to repudiate orthodox Trinitarianism as part of their Reformation claimed to be carrying out, more consistently than either Luther and Calvin or the Anabaptists had done, the full implications of the rejection of Roman Catholicism, which they all had in common.

The challenge of the Protestant Reformation became also the occasion for a resurgent Roman Catholicism to clarify and to reaffirm Roman Catholic principles; that endeavour had, in one sense, never been absent from the life and teaching of the church, but it came out now with new force. As the varieties of Protestantism proliferated, the apologists for Roman Catholicism pointed to the Protestant principle of the right of the private interpretation of Scripture as the source of this confusion. Against the Protestant elevation of the Scripture to the position of sole authority, they emphasized that Scripture and church tradition were inseparable and always had been. Pressing that point further, they denounced justification by faith alone and other cherished Protestant teachings as novelties without grounding in authentic church tradition. And they warned that the doctrine of "faith alone, without works" as taught by Luther would sever the moral nerve and remove all incentive for holy living.

Yet these negative reactions to Protestantism were not by any means the only, perhaps not even the primary, form of participation by Roman Catholicism in the history of the Reformation. The emergence of the Protestant phenomenon did not exhaust the reformatory impulse within Roman Catholicism, nor can it be seen as the sole inspiration for Catholic reform. Rather, to a degree that has usually been overlooked by Protestant historians and that has often been ignored even by Roman Catholic historians, there was a distinct historical movement in the 16th

century that can only be identified as the Roman Catholic Reformation.

The Roman Catholic Reformation. *The Council of Trent.* The most important single event in that movement was almost certainly the Council of Trent, which met intermittently in 25 sessions between 1545 and 1563. The bitter experiences of the late medieval papacy with the conciliarism of the 15th century made the popes of the 16th century wary of any so-called reform council, for which many were clamouring. After several false starts, however, the council was finally summoned, and it opened on Dec. 13, 1545. The legislation of the Council of Trent enacted the formal (and apparently final) Roman Catholic reply to the doctrinal challenges of the Protestant Reformation and thus represents the official adjudication of many questions about which there had been continuing ambiguity throughout the early church and the Middle Ages. The either/or doctrines of the Protestant Reformers—justification by faith alone, the authority of Scripture alone—were anathematized, in the name of a both/and doctrine of justification by faith *and* works on the basis of the authority of Scripture *and* tradition; and the privileged standing of the Latin Vulgate was reaffirmed, against Protestant insistence upon the original Hebrew and Greek texts of Scripture.

Prescriptions of the Council of Trent

No less important for the development of modern Roman Catholicism, however, was the legislation of Trent aimed at reforming—and at re-forming—the internal life and discipline of the church. Two of its most far-reaching provisions were the requirement that every diocese provide for the proper education of its future clergy in seminaries under church auspices and the requirement that the clergy and especially the bishops should give more attention to the task of preaching. The financial abuses that had been so flagrant in the church at all levels were brought under control, and stricter rules were set requiring the residency of bishops in their dioceses. In place of the liturgical chaos that had prevailed, the council laid down specific prescriptions about the form of the mass and liturgical music. What emerged from the Council of Trent, therefore, was a chastened but consolidated church and papacy, the Roman Catholicism of modern history.

New religious orders. Some of the outcome, and much of the enforcement, of the Council of Trent was in the hands of the newly established religious orders, above all of the Society of Jesus, the Jesuits. Unlike the Benedictine monks or the Franciscan and Dominican friars, the Jesuits were specifically dedicated to the task of reconstructing church life and teaching in the aftermath of the Protestant Reformation. They thus came to be called the "shock troops of the Counter-Reformation." In pursuit of that mission they became especially active in scholarship and education, above all in the education of the nobility; through their pupils they sometimes wielded as great an influence in the affairs of the state as in those of the church. Although they were by no means the only religious order in the foreign missions of the church, their responsibility for regaining outside of Europe the power and territory that the church had lost in Europe as a consequence of the Protestant Reformation made them the leading force in the Christianization of newly discovered lands in the Western Hemisphere, Asia, and the islands of the sea. At the beginning of the 17th century, for example, they established in Paraguay a virtually autonomous Jesuit colony.

In addition to the Jesuits, other Roman Catholic religious orders, too, owed their origin to the age of the Reformation. The Capuchin friars renewed the ideals of the Franciscan order, and by their missions both within and beyond the historical boundaries of Christendom they furthered the revival of Roman Catholicism. The Theatines were founded by Gaetano da Thiene and the bishop of Chieti (Theate), Gian Pietro Carafa, who went on to become Pope Paul IV (reigned 1555–59); both through the program of the order and in his pontificate, the correction of abuses in the church assumed primary importance. Despite the attacks of the Reformers on the institutions and even the ideals of monasticism, it was in considerable measure a reformed monasticism that carried out the program of the Roman Catholic Reformation.

The Counter-Reformation. Recognition of the scope and success of the indigenous movements for reform within 16th-century Roman Catholicism, therefore, has rendered obsolete the practice of certain earlier historians, who lumped all of these movements under the heading "Counter-Reformation," as though only Protestantism (or, perhaps, only the historian's own version of Protestantism) had the right to the title of "*the* Reformation"; hence the use here of the term Roman Catholic Reformation. Yet that does not deny a proper meaning of "Counter-Reformation" as part of the larger phenomenon, for counteracting the effects of Protestantism was part of the program of the Council of Trent, the Society of Jesus, and the papacy during the second half of the 16th century and beyond.

The Counter-Reformation was launched wherever there had been a Protestant Reformation, but it met with strikingly varied degrees of success. Most of the "German lands" in which Luther had worked remained Protestant after his death in 1546, but major territories, above all Bavaria and Austria, had been regained for Roman Catholicism by the time the 16th century was over. The Huguenot Wars between 1562 and 1598 regained France for the Roman Catholic cause, although the Edict of Nantes of 1598 granted a limited toleration to the Protestants; it was revoked in 1685. Perhaps the most complete victory for the Counter-Reformation was the restoration of Roman Catholic domination in Poland and in Hussite Bohemia.

Victories for the Counter-Reformation

The victory of the Habsburg Counter-Reformation there and the defeat of Czech Protestantism were a consequence of the Battle of White Mountain of 1620 in the early years of the Thirty Years' War. Often called the first modern war, this series of conflicts wrought devastation in the populations of central Europe, Roman Catholic at least as much as Protestant. The conclusion of the war in the Peace of Westphalia of 1648 meant for Roman Catholicism the de facto acceptance of the religious pluralism that had come out of the Reformation: Protestantism, both Lutheran and Calvinist, obtained a legal standing alongside Roman Catholicism in what had previously been regarded as "Catholic Europe." In a war that had presumably begun as a "religious war" aimed at the resolution of the confessional impasse brought about by the Reformation, the formation of a military alliance between Cardinal Richelieu of France and the Lutheran king of Sweden, Gustav II Adolf, was a symbol of a process of the secularization of politics in which the old antitheses, including finally the very antithesis between Roman Catholic and Protestant, no longer seemed as relevant as they had once been.

(J.J.Pe.)

Post-Reformation conditions. The signing of the peace in 1648 may have meant that the era of the Reformation had ended, but for those who remained loyal to the see of Rome it meant that what had been thought of as a temporary disturbance would now be a permanent condition. The church still claimed to be the only true church of Jesus Christ on earth, but, in the affairs of men and of nations, it had to live with the fact of its being one church among several. The Roman Catholic Church was also obliged to deal with the nations and national states of the modern era one by one. To understand the history of modern Roman Catholicism, therefore, it is necessary to identify trends that went beyond geographic boundaries and to consider trends within particular states or regions—such as France, Germany, the New World, or the mission field—only as illustrations of tendencies that permeated the entire life of the church. Most of the development of Roman Catholicism since 1648 makes sense only in the light of this changed situation.

The results of the change became evident in the papacy of the 17th and 18th centuries. On June 6, 1622, Gregory XV (1621–23) created the Congregation for the Propagation of the Faith (Congregatio de Propaganda Fide, hence "propaganda"). Its responsibility was, and still is, the organization and direction of the missions of the church to the non-Christian world as well as the administration of the affairs of the church in areas that do not have an ordinary ecclesiastical government (for example, the United States as late as 1908). It has therefore played an important role

in the efforts to restore Roman Catholicism in Protestant and, to some degree, in Eastern Orthodox territories.

Developments in France. *The Gallican problem.* In many ways it was the relation of the church to individual political powers rather than the leadership of the popes that determined the course of church history. Not only the shrinking authority of the church as a consequence of the Reformation but also the expanding ambition of the state as a consequence of the growth of nationalism put ecclesiastical and secular governments on a collision course throughout Europe. France, "the first daughter of the church," was the national state whose development during the 17th and 18th centuries most strikingly dramatized the collision, so much so that Gallicanism, as the nationalistic ecclesiastical movement was called in France, is still the label put on the efforts of any national church to achieve autonomy.

Nationalistic ecclesiastical movements

Usually the autonomy from Rome implied subjection to the French crown, particularly during the reign of Louis XIV, who sought to extend still further the so-called prerogatives of France when Rome resisted. A conclave of bishops and deputies met on March 19, 1682, in Paris and adopted the Four Gallican Articles, which had been drafted by Jacques-Bénigne Bossuet, a French bishop and historian. These asserted that: (1) In temporal matters rulers are independent of the authority of the church. (2) In spiritual matters the authority of the pope is subject to the authority of a general council, as had been declared at the Council of Constance. (3) The historic rights and usages of the French church cannot be countermanded even by Rome. (4) In matters of faith the judgment of the pope is not irreformable but must be ratified by a general council. The next move was up to the papacy: Innocent XI and Alexander VIII rejected Louis's candidates for bishoprics in France, and only in 1693, when Innocent XII was pope, was this all but schismatic conflict resolved. Gallicanism was in part an expression of the distinctive traditions of French Catholicism and in part a result of the personal power of Louis XIV, the Sun King. But it was also, and perhaps even more fundamentally, a systematic statement of the inevitable opposition between the papacy and a series of rulers from Henry VIII (1491–1547) of England to Joseph II (1741–90) of Austria, who, though remaining basically Catholic in their piety and belief, wanted no papal interference in their royal business but insisted on the right of royal interference in the business of the church.

Jansenism. The church in France was the scene of controversies other than these administrative and political ones. In 1640 there was published, posthumously, a book by the Dutch theologian Cornelius Jansen, entitled *Augustinus,* which was a defense of the theology of Augustine against the dominant theological trends of the time within Roman Catholicism. Its special target was the teachings and practices associated with the Jesuits. Jansen and his followers claimed that the theologians of the Counter-Reformation in their opposition to Luther and Calvin had erred in the other direction in their definition of the doctrine of grace; *i.e.,* emphasizing human responsibility at the expense of the divine initiative and thus relapsing into the Pelagian heresy, against which Augustine had fought in the early 5th century. Over against this emphasis, Jansenism asserted the Augustinian doctrine of original sin, including the teaching that man cannot keep the commandments of God without a special gift of grace and that the converting grace of God is irresistible. Consistent with this anthropology was the rigoristic view on moral issues taken by Jansenism in its condemnation of the tendency, which it claimed to discern in Jesuit ethics, to find loopholes for evading the uncompromising demands of the divine law. When it was espoused in the *Lettres Provinciales* ("Provincial Letters") of Blaise Pascal, a French philosopher, this campaign against Jesuit theology became a cause célèbre. The papacy struck out against Jansenism in 1653, when Innocent X issued his bull *Cum Occasione* ("With Occasion"), and again in 1713, when Clement XI promulgated his constitution *Unigenitus* ("Only-Begotten").

Theologically, Jansenism represented the lingering conviction, even of those who refused to follow the Reformers, that the official teaching of the Roman Catholic Church was Augustinian in form but not in content; morally, it bespoke the ineluctable suspicion of many devout Roman Catholics that the serious call of the Gospel to a devout and holy life was being compromised in the moral theology and penitential practice of the church. Though Jansenism was condemned, it did not remain without effect, and in the 19th and 20th centuries it contributed to an evangelical reawakening not only in France but throughout the church.

Quietism. Quietism, another movement within French Roman Catholicism, was far less strident in its polemics and far less ostentatious in its erudition but no less threatening in its ecclesiastical and theological implications. Quietism was, in many ways, yet another form of the Augustinian opposition to any recrudescence of the Pelagian idea that man's religious activity can make God propitious to him. In Quietism this belief was associated with the development of a technique of prayer in which passive contemplation became the highest form of religious activity. Christian mysticism had always combined, in an uneasy alliance, the techniques of an aggressive prayer that stormed the gates of heaven and a resigned receptivity that awaited the way and will of God, whatever it might be. In the theology of François de Fénelon, a French archbishop and mystical writer, Quietism was combined with a scrupulous orthodoxy of doctrine to articulate the distinction between authentic Catholic mysticism and false spiritualism. Nevertheless, as scholars of medieval mystical movements have suggested, Quietism showed the great gulf between the Roman Catholicism that came out of the Counter-Reformation and the spirituality of the preceding centuries, both Greek and Latin. A devotion such as that of St. Gregory of Nyssa and Evagrius of Pontus, Greek theologians of the 4th century, was completely ruled out by the legalistic theology that condemned Quietism.

Controversies involving the Jesuits. *The Chinese rites controversy.* An analogous judgment would have to be voiced concerning the Chinese rites controversy centring on Matteo Ricci, an Italian Jesuit missionary in China. Decades of scholarly research into Buddhist and Confucian thought had prepared Ricci for a campaign that sought to attach the Roman Catholic understanding of the Christian faith to the deepest spiritual apprehensions of the Chinese religious tradition; the veneration of Confucius, the great Chinese religious and philosophical leader of the 6th century BC, and the religious honours paid to ancestors were to be seen not as elements of paganism to be rejected out of hand, nor yet as pagan anticipations of Christianity, but as rituals of Chinese society that could be adapted to Christian purposes. Ricci's apostolic labours won him many converts in China, but they also won the suspicion of many in the West that the distinctiveness of Christianity was being compromised in syncretistic fashion. The suspicion did not assert itself officially until long after Ricci's death; but, when it did, the outcome was a condemnation of the Chinese rites by Pope Clement XI in 1704 and again in 1715 and by Pope Benedict XIV in 1742. Ancestor worship and Confucian devotion were said to be an inseparable element of traditional Chinese religion and hence incompatible with Christian worship and doctrine. Here again, the embattled situation of the Roman Catholic Church in the 17th and 18th centuries helps to account for an action that seems, in historical perspective, to have been excessively defensive and rigoristic.

Adaptation to non-Western religious traditions

Suppression of the Jesuits. Among the repercussions of the controversy over Chinese rites was an intensification of the resentment directed against the Society of Jesus, to which some of the other movements mentioned above also contributed. The widespread support enjoyed by Jansenism was due in part to its attack on the moral theology associated with the Jesuits. Pascal's *Lettres Provinciales,* although placed on the Index in 1657, voiced an opposition to Jesuit thought and practice that continued to be read throughout the century that followed. The political role played by members of the Society most probably evoked the campaign to suppress it. The Portuguese crown expelled the Jesuits in 1759, France made them illegal in 1764, and in 1767 Spain and the Kingdom of the Two Sicilies also took repressive action against them. But the

opponents of the Society achieved their greatest success when they took their case to Rome. Pope Clement XIII is said to have replied that the Jesuits "should be as they are or not be at all" and refused to act against them. But his successor, Clement XIV (1769–74), whose election was urged by the anti-Jesuit forces, finally did take action. On July 21, 1773, he issued a brief, *Dominus ac Redemptor* ("Lord and Redeemer"), suppressing the Society for the good of the church. Frederick II of Prussia and Empress Catherine II of Russia—one of them Protestant and the other Eastern Orthodox—were the only monarchs who refused to promulgate the order to suppress the Jesuits when it was issued. In these lands and in others the Society of Jesus maintained a shadow existence until, on Aug. 7, 1814, Pope Pius VII restored it to full legal validity. Meanwhile, however, the suppression of the Jesuits had done serious damage to the missions and the educational program of the church, and this at a time when both enterprises were under great pressure.

Religious life in the 17th and 18th centuries. Yet it would be a mistake to allow the narrative of these controversies to monopolize one's attention. Less dramatic but no less important was the continuing life of the Roman Catholic Church during these centuries as "mother and teacher." Bossuet was not only the formulator of Gallican ideology but also one of the finest preachers of Christian history. He addressed king and commoner alike and asserted the will of God with eloquence, if sometimes with undue precision. Together with Jean Mabillon, a Benedictine monk and scholar, Bossuet helped to lay the foundations of modern Roman Catholic historiography. During the 18th century their work was continued and expanded, especially by Mabillon's confreres, the Maurists, a Benedictine group that edited the works of the Greek and Latin fathers. Both Jansenism and Quietism must be seen not only as parties in a controversy but also as symptoms of religious vitality. Engaging as they did considerable segments of the Roman Catholic laity, they expressed "the practice of the presence of God" with a new vigour.

Scholarly and literary activities

The Roman Catholic Church of this period exercised a profound influence on culture and the arts. Indeed, the spirit of Baroque is inseparable from the Counter-Reformation, as is visible, for example, in the church of Il Gesù in Rome and in the sculpture and architecture of Gian Lorenzo Bernini. Pascal and Cervantes are notable literary figures who expressed Roman Catholic thought and piety through their works. The most fateful of the church's conflicts with modern culture in this period took place in the natural sciences. The condemnation of Galileo in 1616 and again in 1633 as "vehemently suspected of heresy" was more important symbolically than intrinsically, as a sign of the alienation between science and theology. This period saw the establishment or further development of several major religious orders, including the Daughters of Charity, founded by Vincent de Paul in 1633, and the Trappists, who take their name from the Cistercian abbey of La Trappe, which in 1664 was transformed into a community of the Strict Observance.

THE CHURCH IN THE MODERN PERIOD

Catholicism in Revolutionary France. The period of the Reformation and the Counter-Reformation was a time of convulsion for the Roman Catholic Church, but the era of revolution that followed it was, if anything, even more traumatic. This was partly because, despite the polemical rancour of Reformation theology, both sides in the controversies of the 16th and 17th centuries still shared much of the Catholic tradition. Politically, too, the assumption on all sides was that rulers, even when they opposed one another or the church, stood in the Catholic tradition. In the 18th century, however, there arose a political system and a philosophical outlook that no longer took Christianity for granted, that in fact explicitly opposed it, compelling the church to redefine its position more radically than it had done since the conversion of the Roman emperor Constantine in the 4th century.

What made the relation of the Roman Catholic Church to the ancien régime, the political and social system before the French Revolution in 1789, so problematic at the time of the Revolution was a subtle but fundamental difference between them. Although the rhetoric of the Revolution spoke as though the church and the old order had been one, no one could study the history of the church under (or over against) Louis XIV and accept so simplistic an interpretation. Conflict there had been, bitter and uncompromising conflict—and yet conflict within the context of given presuppositions. It is significant, for example, that the French aristocracy, soon to become the hated object of revolutionary zeal, constituted the source of almost all the bishops of the church in the ancien régime. This also meant that positions of authority in the church were largely foreclosed to the lower clergy because of their class. The theological and ecclesiastical parties identified with opposition to Rome were frequently those that drew the support of the laity; Jansenism, for example, was identified as the position of the lay lawyers who spoke for the French courts of justice over against the hierarchy. In spite of the hostility between church and state, therefore, the old regime appeared to its critics to be a monolith. Thus, when the French philosopher Voltaire said, "Écrasez l'infâme" ("Crush the infamous one"), he may have meant superstition, ignorance, and tyranny, but what it added up to concretely in the minds of the revolutionaries was the supposed alliance of the monarchy with the Roman Catholic Church. This identification was only confirmed when the defenders of the established order, both lay and clerical, spoke out against the threat of revolution with a greater awareness of its dangers than of its justification.

Underlying causes of anti-ecclesiastical sentiments

Complicating the predicament of the church in the old regime was the corrosive influence of the Enlightenment on the religious beliefs of much of the lay intelligentsia. Enlightenment rationalism took hold among many defenders of the political status quo as well as among clerical scholars, helping to produce the beginnings of critical biblical scholarship and of religious toleration. It would be an oversimplification, therefore, to put the Enlightenment unequivocally on the side of the critics and revolutionaries. Perhaps no one embodied the spirit of the Enlightenment more completely than Frederick II the Great of Prussia. But the confidence in reason and the hostility to "superstition" cultivated by the Enlightenment inevitably clashed with the Christian reliance on revelation and with the belief in supernatural grace as communicated by the sacraments.

The political and social prerogatives of the church were also threatened by the Enlightenment, especially when it was allied with the expanding claims of an autocratic "enlightened despotism." The brotherhood taught by such groups as the Freemasons, members of secret fraternal societies, and the Illuminati, a rationalistic secret society, provided a rival to the Catholic sense of community. In *The Magic Flute,* the Austrian composer Wolfgang Amadeus Mozart (who wrote his *Requiem Mass* in the same year) celebrated the Masonic alternative to the mass of the church.

Although leaders of the state were often more hospitable to the ideas of the Enlightenment than were leaders of the church, the latter proved more accurate in their estimate of the revolutionary implications of these ideas. The "heavenly city of the 18th century philosophers" may originally have been intended as a substitute for the City of God, but it also provided much of the ideological rationale for the attack upon the ancien régime. In the familiar epigram of the Swiss writer Jacques Mallet du Pan, after the French Revolution, "philosophy may boast her reign over the country she has devastated." The action of the French Revolution against the church took many forms, but the most significant was the Civil Constitution of the Clergy of 1790. In it, a Gallicanism originally enunciated in the name of the absolute French monarchy attempted to subject the church to the National Assembly. The entire church in France was reorganized, with the authority of the pope restricted to doctrinal matters. Later in that year a constitutional oath was required of all the French clergy, most of whom refused. Pope Pius VI (1775–99) denounced the Civil Constitution in 1791, and Catholic France was divided between the adherents of the papal system and the proponents of the new order. The closing

decade of the 18th century was dominated by this conflict, and no resolution was provided by either church or state. The ultimate humiliation of the church came when Pius VI was driven out of Rome by the French armies in 1798 and in the following year was taken captive by them and dragged back to France, where he died. Not since the Great Schism and the Babylonian Captivity had the prestige of the papacy sunk so low.

Napoleon I—exportation of the Revolution. As it was obvious that the French Revolution itself had to be carried to some more permanent settlement, so it was recognized on all sides that a more stable arrangement of church-state relations was essential. This was achieved by Napoleon Bonaparte in a concordat concluded with Pope Pius VII on July 15/16, 1801. It recognized that Roman Catholicism was the faith of most Frenchmen and granted freedom of worship. All incumbents of bishoprics were to resign and were to be replaced by bishops whom Napoleon, as first consul, would nominate. The properties of the church that had been secularized during the Revolution were to remain so, but the clergy was to be provided with proper support by the government. Many historians maintain that the Concordat of 1801 was as decisive for modern church history as the conversion of Constantine had been for ancient church history. As Constantine had first recognized and then established Christianity in the Roman Empire, so a series of concordats and other less formal agreements created the modus vivendi between the church and modern secular culture. What this meant for the papacy was the realization that most of the temporal holdings of the church in Europe had to be surrendered. The eventual outcome of this realization was the creation of Vatican City as a distinct political entity, but only after a long conflict over the States of the Church during the unification of Italy in 1869–70. First, however, came the period after the fall of Napoleon, when those who had emerged victorious at the Battle of Waterloo (1815) attempted to restore the previous condition. The Society of Jesus was revived in 1814, and the Congress of Vienna in 1814–15 helped to establish a basis for the recovery of the church during the 19th century. Temporary though these supposed settlements were, they made it clear to those living in the following period that the church would continue to be a force to be reckoned with in the affairs of Europe and America.

The Concordat of 1801

The reign of Pius IX (1846–78). Much of the history of Roman Catholicism in the 19th century is identified with the pontificates of two men: Pius IX, who was pope for a third of a century, and his successor, Leo XIII, who was pope for a quarter of a century (1878–1903).

Few popes of modern times have presided over so momentous a series of decisions and actions as Pius IX. During his reign the development of the modern papacy reached a kind of climax with the promulgation of the dogma of papal infallibility. It had long been taught that the church, as "the pillar and bulwark of the truth," could not fall away from the truth of divine revelation and therefore was "indefectible" or even "infallible." Inerrancy had likewise been claimed for the Bible by both Roman Catholic and Protestant theologians. As the visible head of that church and as the authorized custodian of the Bible, the pope had also been thought to possess a special gift of the Holy Spirit, enabling him to speak definitively on faith and morals. But this gift had not itself been identified in a definitive way. The outward conflicts of the church with modern thought and the inner development of its theology converged in the doctrinal constitution *Pastor Aeternus* ("Eternal Shepherd"), promulgated by the first Vatican Council on July 18, 1870. It asserted that "the Roman Pontiff, when he speaks ex cathedra, that is, when in discharge of the office of pastor and teacher of all Christians, by virtue of his supreme apostolic authority he defines a doctrine regarding faith or morals to be held by the universal Church, by the divine assistance promised to him in blessed Peter, is possessed of that infallibility with which the divine Redeemer willed that his Church should be endowed." The decree was, of course, retroactive, even though there were historical incidents that appeared to contradict the retroactivity, such as the condemnation of Pope Honorius I by the third Council of Constantinople in 680, which were cited by opponents of the decree. This opposition was, however, ineffective, and the dogma of infallibility became the public doctrine of the church. Those who continued to disagree withdrew to form the Old Catholic Church, which was centred in The Netherlands, Germany, and Switzerland.

Promulgation of the dogma of papal infallibility

Even before the promulgation of this dogma, Pope Pius had exercised the authority that it conferred on him. In 1854, acting on his own prerogative and without any council, he defined as official teaching the doctrine of the immaculate conception of the Virgin Mary, "that the Most Blessed Virgin Mary, at the first instant of her conception, was preserved immaculate from all stain of original sin, by the singular grace and privilege of the Omnipotent God, in virtue of the merits of Jesus Christ." This put the church unequivocally on one side of a debate over the doctrine of Mary that had been going on since the Middle Ages. Ten years later, Pius issued a document that was in some ways even more controversial, the *Syllabus of Errors* (Dec. 8, 1864). In it he condemned various "errors" characteristic of modern times, including pantheism, Socialism, civil marriage, secular education, and religious indifferentism. By thus appearing to put the church on the side of reaction against the forces of liberalism, science, democracy, and tolerance, the *Syllabus* seemed to be part of the retreat of Roman Catholicism from the modern world. At the same time, it did seek to clarify the identity of Roman Catholic teaching at a time when it was being threatened on all sides.

This combination of reactions to modern thought and society came to a head in the conflict over "Americanism," which was condemned by Leo XIII in 1899, and even more vigorously in the Kulturkampf (*i.e.*, struggle in Germany with Catholicism). Prince Otto von Bismarck, both because he was a Prussian and because he was a Protestant, resisted the basic trend of the developments just traced. In the Roman Catholic parties of the centre in the German states, he saw an obstacle to the form of German reunion to which he was dedicated, viz., a predominantly Protestant Germany without Roman Catholic Austria. The *Syllabus of Errors* and the dogma of infallibility represented the hostility of Roman Catholicism to the very sort of state he was trying to establish. Much of the theological opposition to papal infallibility came from German thinkers, notably Ignaz von Döllinger, to whose defense Bismarck sprang. The conflict between church and state came in several principal areas. The Kulturkampf began with the exclusion of the Roman Catholic Bureau from the Ministry of Culture and Cultus in the Prussian state. Bismarck asserted the authority of the state over all education in Prussia and had the Society of Jesus expelled. Then, in direct defiance of the *Syllabus of Errors*, he required civil marriage of all, regardless of whether or not they had also exchanged their vows before a clergyman. Laws were passed compelling candidates for the Roman Catholic priesthood to attend a German university for at least three years. Bismarck summarized his defiance of the Pope in an allusion to the conflict between Pope Gregory VII and Emperor Henry IV in the 11th century: "We are not going to Canossa!" When Pius IX died in 1878, the conflict was still unresolved.

The reign of Leo XIII (1878–1903). Although Leo XIII was no less conservative in his theological inclinations than his predecessor, his positive appreciation of the church's opportunities in modern society gave his pontificate a significantly different cast from that of Pius. On issues of church doctrine and discipline his administration was a strict one. It was during his reign that the Modernist movement, which advocated the use of biblical and historical criticism and freedom of conscience, arose within Roman Catholicism; and, although the formal condemnation of its tendencies did not come until 1907, four years after his death, he had made his opposition to this trend clear by the establishment of the Pontifical Biblical Commission as a monitor over the work of scriptural scholars. The positive side of his theology came to voice in the encyclical *Aeterni Patris* ("Eternal Father") of Aug. 4, 1879, which, more than any other single document, provided a char-

Rise of the Modernist movement

ter for the revival of Thomism (the theological system based on the thought of Thomas Aquinas) as the official philosophical and theological system of the Roman Catholic Church. It was to be normative in the training of priests and in the education of the laity at universities. Leo also sponsored the launching of a critical edition of the works of Aquinas. In 1895 he appointed a commission to decide whether, despite the separation from Rome in the 16th century, the priestly ordination of the Anglican Communion was valid, as that of the separated Eastern churches was; in 1896 he issued *Apostolicae Curae* ("Apostolic Concerns"), which denied the validity of Anglican orders and was a setback for ecumenical hopes.

The "pope of peace"

Leo XIII is best remembered for his social and political thought, which earned him the sobriquet the "pope of peace." He mollified the church's position toward the policies of Bismarck, who in turn moved toward a compromise. Diplomatic relations between Germany and the Vatican were restored in 1882, and gradually the restrictive laws were lifted. But the greatest achievements of Leo's work in the relation between the church and modern culture were his social and political encyclicals, which articulated a positive social philosophy. In *Libertas* ("Liberty"), issued on June 20, 1888, he sought to affirm what was good about political liberalism, democracy, and freedom of conscience. Above all, the encyclical *Rerum Novarum* ("Of New Things") of 1891 put the church on the side of the struggle for social justice. Though rejecting 19th-century socialism, the pope was also severe in his condemnation of an exploitative laissez-faire capitalism and in his insistence upon the duty of the state to strive for the welfare of all its citizens. The social thought of Leo XIII helped to stimulate social action among Roman Catholics, such as the Christian Social Movement. When he died, the church seemed to be entering a new era of respect and influence, but the turmoil of war, depression, and revolution in the 20th century intervened.

The period of the world wars. Pope Pius X (1903–14) symbolized the transition from the 19th century to World War I. In his encyclical, *Pascendi Dominici Gregis* ("Feeding the Lord's Flock"), of Sept. 8, 1907, he condemned Modernism as "the résumé of all the heresies," and in 1910 he prescribed that clergy and seminary professors take an oath abjuring Modernism and affirming the correctness of the church's teachings. He sponsored the revision of the code of canon law. More perhaps than any of his immediate predecessors or successors, Pius X gave attention to the reform of the church's liturgy, especially to the Gregorian chant, and advocated early and frequent reception of Holy Communion. Yet hanging like a cloud over his pontificate was the threat of world war. The last major document issued by Pius X was a lament over the outbreak of war.

World War I, often called the real end of the 19th century, was also a major turning point in modern Roman Catholic history. Ever since ancient times the church had been accustomed to order its relations to human society by negotiations with kings and emperors, preferably members of its own fellowship. The war and the revolutions attending it meant the end of the Hohenzollern (Germany), Habsburg (Austria-Hungary), and Romanov (Russia) dynasties, obliging the church to come to terms with the new realities of democratic, communist, and fascist regimes.

The Lateran Treaty

Of special significance was a series of pacts with the Fascist Italy of Benito Mussolini. In 1929 the church and the Italian government signed the Lateran Treaty, which finally regularized relations between them and gave Vatican City independent status. In 1933 the church concluded a concordat with Nazi Germany, hoping to protect its own interests and those of minorities; but this hope proved to be ill founded, and the church's relation with Hitler and his regime deteriorated. Although Pius XI (1922–39) and Pius XII (1939–58) spoke out several times against the excesses of the regime, they did little to restrain it. The papacy spoke out much more often during the Spanish Civil War (1936–39) against the dangers of communism. As a diplomat and former papal secretary of state, Pius XII was obliged, under the pressures of World War II, to clarify the church's teachings on war and peace as well as to work out a strategy of survival. In 1950 he became the first pope since the first Vatican Council to exercise the right of defining doctrine, proclaiming the bodily assumption of the Virgin Mary to be a dogma binding on all members of the church. Earlier in that same year, in the encyclical *Humani Generis* ("Of the Human Race"), he had given a reproof to various theological trends that appeared to be reviving the ideas and methods of Modernism.

Vatican II. From these two papal promulgations of 1950 many observers were ready to conclude that in the second half of the 20th century Roman Catholicism would assume an essentially defensive posture in relation to the modern world. Those who had come to that conclusion were compelled to revise it by the pontificate of John XXIII (1958–63) and by the second Vatican Council (1962–65). During his brief reign Pope John issued several important encyclicals. Of special interest was *Mater et Magistra* ("Mother and Teacher"), published in 1961, which explicitly attached itself to the *Rerum Novarum* of Leo XIII in calling for justice and the common good as the norms of social conduct. Two years later, in *Pacem in Terris* ("Peace on Earth"), the pope addressed himself not only to members of the church but to "all men of good will." In this encyclical he formulated, more completely than any previous pope had done, a social philosophy for peace among men and between nations. This spirit of reform and concern came to expression in the council, which Pope John convoked but which he did not live to see to its conclusion. The council brought about drastic changes in the life and worship of the church. To Eastern Orthodox and Protestant Christians it extended the hand of fraternal understanding. To the Jewish community it addressed words of regret for the anti-Semitism of the Christian past. To the world religions it spoke of the church's admiration for the spiritual values in those traditions. The council also expressed its respect for the integrity and freedom of humanity. In its importance for the development of the church, the second Vatican Council will probably rank with the councils of Nicaea (325), Chalcedon (451), and Trent (1545–63).

Influence of Vatican II

Aftermath of the council. The legacy of Vatican II remained a divided one. For some Catholics, the promise of far-reaching reform remained unfulfilled; for others, the council went too far, undermining the traditional beauty of church teachings and liturgy. This ambiguity was apparent during the reign of Pope Paul VI (1963–1978) when many of the reforms of the council were implemented, most notably in the liturgy. The Latin mass was replaced by the vernacular mass, altars were turned around so that the priest faced the congregation, and greater participation by the laity in the celebration of mass was instituted. Paul improved relations with the Church of England, the Greek Orthodox church, and non-Christian faiths. In an encyclical in 1964, he called for social justice and denounced the excesses of capitalism, which led conservatives to accuse him of being a Marxist. The pope also issued two controversial encyclicals: *Sacerdotalis caelibatus* in 1967, which affirmed clerical celibacy, and *Humanae vitae* ("Of Human Life") in 1968, which forbade the use of artificial birth control. These encyclicals alienated many Catholics and led many priests to renounce their vows, just as the progressive reforms of the pope and council led to the schism of the French archbishop Marcel Lefebvre.

The divided legacy of the council continued during the reign of Pope John Paul II (1978–). An active and charismatic figure, John Paul moved away from the episcopal collegiality stressed at Vatican II in favour of a more centralized papal authority. He opposed admitting women or openly gay men to the priesthood. He was criticized for not halting the decline in church attendance and in the number of priests as well as for his conservative teachings on sexuality. He promoted controversial conservative groups, including Opus Dei (*q.v.*), and promoted stricter adherence to theology as indicated by his opposition to liberal theologian Hans Küng and liberation theology. On the other hand, John Paul noted the error of the condemnation of Galileo and the importance of adjusting theology to modern science. Although he was a staunch opponent of communism, he also criticized the excesses of Western capitalism. He instituted the new Code of Canon Law, but

his most important activity—fully in the spirit of Vatican II—was his outreach to other faiths, Christian and non-Christian. His most remarkable efforts included overtures to Judaism. He was the first pope to visit the synagogue in Rome, and in 2000 he made a historic pilgrimage to Jerusalem, where, in the spirit of brotherhood, he prayed at Al Aqṣā Mosque and the Western Wall. Despite challenges, the church began its third millennium as one of the largest and most significant religious bodies in the world.

(J.J.Pe.)

ROMAN CATHOLICISM OUTSIDE EUROPE

The New World: the Spanish and Portuguese empires. *Colonial period.* The Western Hemisphere was discovered by Europeans immediately before the Protestant Reformation. The discovery at that moment in history and the development of the New World by Roman Catholic empires (*e.g.,* Spain) is of major significance in the religious history of the hemisphere. The only part of it that was to be non-Catholic in its general outlook was the area that was to become the United States and Anglophone Canada. Spain and Portugal were in their prime as sea powers in the late 15th and early 16th centuries, and they were most responsible for colonizing and establishing the Christian faith in the southern two-thirds of the Americas.

The chief Catholic missionaries were the Franciscans, Dominicans, Augustinians, and Jesuits. Well-trained and self-sacrificing, they were able to go wherever Spanish and Portuguese ships went. Sometimes they were accused of serving as religious supporters of anything the crown desired, but, because the missionaries were in quest of souls, there were also clashes between Catholic churchmen and colonizers or traders. Some missionary efforts met with successes among the indigenous peoples, and at times Catholicism was able to temper the inhumanity of the conquerors. Best-known among the humane spokesmen for Indians was the Dominican Bartolomé de Las Casas (1474–1566), "the Apostle of the Indians," who publicized the white atrocities against the Indians and was named bishop of Chiapas (Mexico) in 1543.

Missionaries

In the course of the 16th–19th centuries, European colonists and immigrants from nations other than Spain and Portugal came to Latin America. Even when these movements included Protestant missionaries, they did little to disrupt the generally Catholic cultures. Modern secular forces, however, jostled the Catholic settlements. The case of Mexico is illustrative; its ruling powers repeatedly proscribed Catholic education and embodied anticlerical interests. Still, the Mexican people remained largely Catholic, although they blended some of their native religious values and practices with Catholic forms.

After independence. The inevitable reaction by Catholic and non-Catholic alike arose against the colonial powers. This took the form of movements of independence, anticlerical revolts that were directed against European powers. Some institutions, particularly those devoted to education, were opposed to the practices of Catholicism. Because so many of the clergy came from Europe, anti-European sentiment assured that the American fields were not attractive, and chronic clerical shortages prevailed. As was the case in Europe, the various revolutions were often concurrent with or encouraging to the various versions of Enlightenment thought, and this meant that they were expectably uncongenial to the truth claims of Christianity.

Anticlericalism and revolution

By the middle of the 20th century, wherever Latin-American Catholicism remained strong, it was dismissed by much of the rest of the world as appearing to be uncongenial to the legitimate aspirations of majorities. Because of the cosmopolitan influences of the second Vatican Council (1962–65), however, the self-generated renewal of the church, and the presence of a new socially responsible leadership, there appeared during the 1960s a more radical Catholicism. Dom Helder Câmara of Recife, Braz., exemplified the impulse toward social reform. Camillo Torres, killed in the role of a Colombian guerrilla, typified the association of a Catholic minority with revolution.

Spanish and French missions in North America. Though at the time of its settlement the United States under British and continental Protestant influences became a largely Protestant outpost, Spanish Catholics did establish missions in Florida and elsewhere. Franciscans began work in California in 1514 and in New Mexico in 1581; this work reached its greatest success when the missionary Fra Junípero Serra founded stations all along the California coast after 1769. French explorers and conquerors settled much of eastern Canada and brought with them a Catholic Church that has remained dominant there up to the present. French missionaries penetrated the Great Lakes region and the Mississippi Valley, but their efforts left few traces when the North American interior came to be settled by English-speaking people late in the 18th century.

Roman Catholicism in the United States and Canada. *United States.* Of the 13 original colonies, only Maryland, settled in 1634 and established in 1649, included an appreciable number of Catholics before American independence. Catholics were often excluded from many colonies, where Congregational or Episcopal churches were supported by law. According to some estimates, there were at most 25,000 Catholics in a colonial population of almost 4,500,000 at the time of independence after 1776.

Relationship of Catholics to Protestants

From the first, Catholic leadership enjoyed its place in the free society of the new United States. Bishop John Carroll, a representative of a notable colonial Catholic family, pioneered in exploring positive relations between Catholic religionists and their fellow citizens. Beginning in the 1830s and 1840s, the assurances of religious freedom were added attractions for millions of Catholic immigrants who made their way to the United States for economic reasons. Coming from Ireland or the European continent to a nation of largely British and almost exclusively Protestant provenance, they awakened suspicion and were met by what has been called a nativist Protestant Crusade.

Catholicism endured and built impressive institutions, including parochial schools. These elementary and secondary schools were formed late in the 19th century because Catholic leaders feared Protestant influences in the public schools. Through these Catholic agencies, Catholic leaders were able to help their people combine religious loyalties to Rome and civil loyalties to America. The church was plagued by "trusteeism," a debate over lay versus clerical control of ecclesiastical institutions; "Americanism," the charge that American Catholics were innovating in doctrine and practice; immigration; and the rescue of souls. The church prospered through all these adversities.

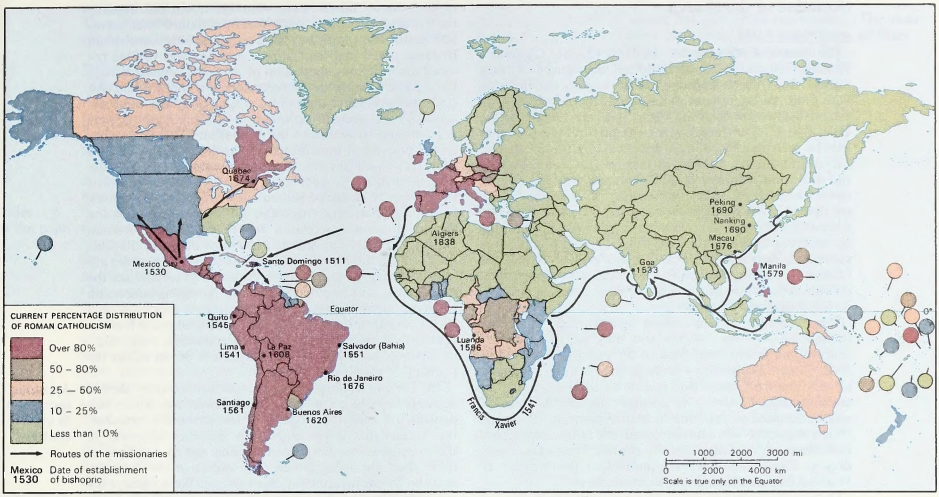
After World War I anti-Catholicism declined. By 1960 a Roman Catholic, John F. Kennedy, had become president—an office previously thought to be out of range for Catholics. Tensions over church–state issues remained, but these were minimized. The ecumenical age also brought about better relations between the various faiths.

Canada. In Canada, England came to dominance in 1713, but the Quebec Act of 1774 guaranteed Catholic rights. The period of new nationalisms after World War II found French Catholics in Quebec nervous about the possible disappearance of their culture. They took steps to assure the perpetuation of the faith, language, and outlook of the French-speaking Catholic millions in an otherwise largely Protestant nation. Militant movements even asked for the formation of a new nation in Quebec.

The spread of Roman Catholicism in Africa and Asia. Though Catholicism had shaped Latin-American and eastern Canadian culture, and though it came to be at home in the United States, it also found itself to be a worldwide presence for the first time in the 19th century. This expansion was the result both of Western nations' imperial presence in Africa and Asia and of the rebirth of a missionary spirit in Christendom.

Catholicism as a worldwide presence

Christian missions in Africa. In Africa almost nothing remained of the early Christian communities in the north, which had become largely Muslim. The Muslim presence there offered more resistance than did the native African religionists in the remaining part of the continent. Christians were not welcomed and were often persecuted. Even in partly Christian Abyssinia (Ethiopia), where the Coptic Church was prominent, Catholics were largely excluded except between 1702 and 1839. An archbishopric was established in Algiers, and in 1868 Archbishop Charles Lavigerie founded the White Fathers,



Routes of missionaries, dates of establishment of dioceses, and current distribution of Roman Catholicism.

who were largely unsuccessful missionaries from that base.

West Africa presented persistent problems for all Christians, because it was from there that European nations had carried on most of the slave trade. Portuguese colonialists did help the Catholic Church establish itself in parts of West Africa, but progress was slow. Catholicism fared better in East Africa. Uganda, Kenya, and Tanganyika (now Tanzania) have thriving churches. The record was less triumphant farther south, in part because of Dutch and British Protestant power. Yet there, as elsewhere, independent missionary societies worked despite hardship.

Christian missions in Asia. In Asia Catholicism was able to profit from Portuguese and Spanish adventures from the 16th century on. In that part of the world different clashes occurred. Asians had not had contact, as Muslims had, with biblical views of history and destiny. Buddhists, Taoists, Confucians, and Hindus were devoted to worldviews uncongenial to Western attitudes toward God and history. Catholicism was itself torn over the permissible degrees of accommodation to Asian traditions.

Problems of accommodation

In India there were traces of missionary extensions from premodern centuries (*e.g.*, the Malabar Syrian Christians), and Catholicism here and there succeeded in finding new bases. But the suppression of the Jesuits in 1773 for reasons of European politics removed the most assertive group from the scene at the most inopportune moment. Catholics flourished under persecution in Indochina, in what is now called Vietnam. The major drama occurred, however, in China and Japan, which were opened to Westerners after centuries of relative isolation. In the 19th century Catholic institutions, such as churches, hospitals, and schools, became familiar sights in China, but the Boxer Rebellion in 1900 revealed Chinese hostility to the Western presence in their country.

In Japan little was left of the 16th-century missions except for an isolated sect of Catholics on an island near Nagasaki. In both China and Japan only a small percentage of the people ever became Catholic. The triumph of communism in China in 1949 brought the end of Catholic missionary activity and proscriptions against native Catholic practices, but an indigenous Catholicism survived, divided between Roman loyalists and the adherents of an autonomous Chinese church. The sacrifices of the church in China were recognized by Pope John Paul II, who canonized 87 Chinese Christians and 33 missionaries to China in 2000. Postwar Japan saw Catholicism engulfed by resurgent religions and a new secular spirit.

The development of indigenous clergy and native institutions. Foreseeing some of the 20th-century difficulties, thoughtful Catholics began during the 19th century to argue that Western religions were not able to be appropriated directly and may not long be permitted in many places. Therefore they began to advocate the development of indigenous clergy. The resultant native institutions often adopted some elements from the local cultures, but seldom were fusions of distinctive elements of Asian or African religion with Christian doctrine consciously permitted.

Conflicts and relations with national governments. If the recent centuries represented much promise for Catholicism's self-definition as a universal church, they also meant setbacks. Christians of the West had often exploited the developing nations, looted their resources, enslaved or demeaned their populations, and extirpated their religions and cultures. As colonial yokes were thrown off, new nations in quest of their own identities encouraged the renewal of the non-Christian religions that had long been part of their cultures. The Western Catholic could serve as a bogey. Overt anti-Christianity of most Marxist or communist parties in these countries meant a rolling back of Catholicism. By the end of the 20th century, however, it was Catholicism, or at least the activities of Pope John Paul II, that contributed to the collapse of communism in eastern Europe.

The new world consciousness of Roman Catholicism. On the other hand, the second Vatican Council also saw the definition of more positive views of non-Catholic high religions, a fact that served somewhat to diminish the impulse to convert the whole world to explicit faith in Christ and obedience to Rome. Catholicism engaged in internal reforms that suggested a new responsiveness to revolutionary social situations. At least minor new local adaptations in Asian and African churches were permitted, and Western imperial pride was specifically condemned by modern popes. Notably, John Paul II apologized for the sins of the Crusaders and others from the history of the church and actively cultivated relationships with the churches of eastern Europe. Following his leadership and that of his predecessors, the world's Catholics have shown themselves more ready than ever before to reach out to adherents of other religions and to have a new regard for secular human development. (M.E.M./J.J.Pe.)

Effects of the second Vatican Council

Structure of the church

DOCTRINAL BASIS

The nature of the church. In 1965 M.-J. le Guillou, a Roman Catholic theologian, defined the church in these terms: "The Church is recognized as a society of fellowship with God, the sacrament of salvation, the people of God established as the body of Christ and the temple of the Holy Spirit." The progress of Roman Catholic theology can be seen in the contrast between this statement and the definition still current as late as 1960, substantially the definition formulated by Robert Bellarmine, a Jesuit controversialist, in 1621: "the society of Christian believers united in the profession of the one Christian faith and the participation in the one sacramental system under the government of the Roman Pontiff." The older definition, created in response to the Protestant claims, defines the church in external and juridical terms. The more recent definition is an attempt to describe the church in terms of its inner and spiritual reality.

From the time of the earliest heresies the church has thought of itself as the one and only worshiping community that traced itself back to the group established by Jesus Christ. Those who withdrew from it were religiously no different from those who had never belonged to it. The ancient adage, "There is no salvation outside the church," was understood as applying to membership in this group. When this adage was combined with the notions contained in Bellarmine's definition of the church, lines were clearly drawn. These lines were maintained in the breakup of Western Christendom in the Reformation.

There were, however, other factors determining the idea of the one true church. The Roman Catholic Church had never excluded the Orthodox Church, which had seceded from the Roman Church in 1054, from the community of Christian believers. Furthermore, the juridical definition of the church did not include such traditional themes as the communion of the saints and the body of Christ, both of which look beyond the visible, juridically constituted church. The theme of the communion of saints refers to the church as a whole, including both the living (the church militant) and the dead (the church suffering in purgatory—a state for those who must be cleansed from lesser sins—and the church triumphant in heaven). The idea of "communion" appears in early church literature to indicate the mutual recognition of union in the one church and the notion of mutual services.

The mystical body of Christ

The theme of the body of Christ appears in the letters of Paul (Romans 12; 1 Corinthians 12; Ephesians 4 and 5; Colossians 1). In modern Roman Catholic theology the term mystical has been added to "body," doubtless with the intention of distinguishing the church as body from the juridical society. Pius XII, in the encyclical *Mystici Corporis* (1943; "The Mystical Body"), identified the mystical body with the Roman Catholic Church. Most Roman Catholic theologians and the second Vatican Council have taken a less rigorous view, trying to find some way of affirming membership in the body for those who are not members of the Roman Catholic Church. The documents of the council described the church as the "People of God" and as a "Pilgrim Church," but no generally accepted statement of membership in this church has yet emerged. The second Vatican Council also departed from established Roman Catholic theology since the Reformation by using the word church in connection with the Protestant churches. This use has caused some confusion, but the trend is now rather to think of one church divided than of one true church and other false churches.

Apostolic succession. The claim of the Roman Catholic Church to be the one legitimate continuation of the community established by Jesus Christ is based on apostolic succession. This does not mean that there are apostles, nor does it mean that individual Apostles transmitted some or all of their commission to others. The officers of the church, the bishops, are a college (organized group or body) that continues the college of the Apostles, and the individual bishop is a successor of the Apostles only through his membership in the college.

The idea of apostolic succession appears in the writings of Irenaeus, a Church Father who died about 202. Against the Gnostics (dualistic sects that maintained that salvation is not from faith but from some esoteric knowledge) Irenaeus urged that the Catholic teaching was verified because a continuous succession of teachers, beginning with the Apostles, could be demonstrated. In the 3rd and 4th centuries problems of schism within churches were resolved by appealing to the power of orders (*i.e.,* the powers a person has by reason of his ordination either as deacon, priest, or bishop) transmitted by the imposition of hands through a chain from the Apostles. Orders in turn empowered the subject to receive the power of jurisdiction (*i.e.,* the powers an ordained person has by reason of his office). In disputes between Rome and the Eastern churches the idea of apostolic succession was centred in the Roman pontiff, the successor of Peter; it will be observed that this goes beyond the idea of collegial succession. Apostolic authority is defined as the power to teach, to administer the sacraments, and to rule the church. Apostolic succession in the Roman Catholic understanding is validated only by the recognition of the Roman pontiff; and the Roman Catholic Church understands the designation "apostolic" in the creed as referring to this threefold power under the primacy of the Roman pontiff.

Apostolic succession in other churches

The Roman Catholic Church has not entirely denied apostolic succession to non-Roman churches. Rome recognizes the validity of orders in the Orthodox churches; this means that it recognizes the sacramental power of the priesthood but does not recognize the government of these churches as legitimate. The orders of the Anglican and the Swedish Lutheran churches, on the contrary, are not recognized by Rome, and the entire threefold quality of apostolic succession is denied them. Oriental churches in union with Rome (Eastern Catholics) are recognized as in full apostolic succession. Luther and Calvin saw clearly that their position could not be maintained if apostolic succession were necessary; they therefore affirmed that apostolic succession had been lost in the Roman Church by doctrinal and moral corruption and that the true church was found only where the gospel was rightly

Copyright Bibliotheque royale Albert 1er, Bruxelles (MS 9428, fol. 153 verso)



Christ giving Peter the keys to the kingdom of heaven, manuscript illumination in an evangelistary from the Benedictine abbey of St. Willibrord at Echternach (in modern Luxembourg), *c.* 1050. In the Bibliothèque Royale Albert 1er, Brussels (MS. 9428, fol. 153 verso).

preached and the sacraments were rightly administered. Thus, Protestant churches generally have not accepted the necessity of apostolic succession. (J.L.McK.)

THE PAPACY

The papal office. The word papacy (Latin *papatia,* derived from *papa,* "pope"; *i.e.,* father) is of medieval origin. In its primary usage it denotes the office of the pope (of Rome) and, hence, the system of ecclesiastical and temporal government over which he directly presides.

The multiplicity and variety of papal titles themselves indicate the complexity of the papal office. In the *Annuario Pontificio,* the official Vatican directory, the pope is described as bishop of Rome, vicar of Jesus Christ, successor of the prince of the Apostles, *pontifex maximus* ("supreme pontiff") of the universal church, patriarch of the West, primate of Italy, archbishop and metropolitan of the Roman province, sovereign of the state of Vatican City, and servant of the servants of God. In his more circumscribed capacities as bishop of Rome, metropolitan of the Roman province, primate of Italy, and patriarch of the West, the pope is the bearer of responsibilities and the wielder of powers that have their counterparts in the other episcopal, metropolitan, primatial, and patriarchal jurisdictions of the Roman Catholic Church. What differentiates his particular jurisdiction from these others and renders his office unique is the Roman Catholic teaching that the bishop of Rome is at the same time successor to St. Peter, prince of the Apostles. As the bearer of the Petrine office, he is raised to a position of lonely eminence as chief bishop or primate of the universal church.

Titles of the pope

Basic to the claim of primacy is the Petrine theory, according to which Christ, during his lifetime, promised the primacy to Peter alone, and, after his Resurrection, actually conferred that role upon him. Thus John 1:42 and, especially, Matthew 16:18 f.: "And I tell you, you are Peter, and on this rock I will build my church, and the powers of death shall not prevail against it. I will give you the keys of the kingdom of heaven, and whatever you bind on earth shall be bound in heaven, and whatever you loose on earth shall be loosed in heaven." Also John 21:15 f.: "Feed my lambs . . . Tend my sheep." Vatican I, in defining the Petrine primacy, cited these three texts, interpreting them to signify that Christ himself directly established St. Peter as prince of the Apostles and visible head of the Church Militant, bestowing on him a primacy not merely of honour but of true jurisdiction. In defining also that the Petrine primacy was, by Christ's establishment, to pass in perpetuity to his successors and that the bishops of Rome were these successors, Vatican I cited no further scriptural texts. In defining further, however, that the Roman pontiffs, as successors in the Petrine primacy, possess the authority to issue infallible pronouncements in matters of faith or morals, the council cited both Matthew 16:18 f. and Christ's promise to Peter at the Last Supper: "But I have prayed for you that your faith may not fail; and when you have turned again, strengthen your brethren" (Luke 22:32).

Primacy of Peter

Historical conceptions of papal authority within the church. Of the Petrine texts, Matthew 16:18 f. is clearly central and has the distinction of being the first scriptural text invoked to support the primatial claims of the Roman bishops. Before the mid-3rd century, however, and even after that date, some Western, as well as Eastern, patristic exegetes (early Church Fathers who in their interpretation of the Bible used critical techniques) understood that by the "rock" Christ meant to refer not to Peter but to himself or to the faith that Peter professed. Nevertheless, in the late 4th and 5th centuries there was an increasing tendency on the part of the Roman bishops to justify scripturally and to formulate in theoretical terms the ill-defined preeminence in the universal church that had long been attached to the Roman Church and to its bishop. Thus, Damasus I, despite the existence of other churches of apostolic foundation, began to call the Roman Church "*the* apostolic see." About the same time the categories of the Roman law were borrowed to explicate and formulate the prerogatives of the Roman bishop. The process of theoretical elaboration reached a culmination in the views of Leo I and Gelasius I, the former understanding himself not simply as Peter's successor but also as his representative, or vicar. He was Peter's "unworthy heir," possessing by analogy with the Roman law of inheritance the full powers Peter himself had wielded, which he interpreted as monarchical, since Peter had been endowed with the *principatus* over the church.

The vicar of Peter

Medieval views. On the purely theoretical level the distance between the claims advanced by Leo I and the position embodied in Vatican I's primacy decree is not great. Medieval popes, such as Gregory VII, Innocent III, and Innocent IV, clarified by their practice as well as by their theoretical statements the precise meaning of that fullness of power (*plenitudo potestatis*) over the church to which, according to some scholars, Leo I himself had laid claim. In this they were aided not only by the efforts of publicists such as the Italian theologian and philosopher Aegidius Romanus (d. 1316), who magnified the pope's monarchical powers in unrestrained and secular terms, but also by the massive development during the late 11th, 12th, and 13th centuries of a highly romanized canon law. Gratian's *Decretum* (*c.* 1140), the unofficial collection of canons that became the fundamental textbook for the medieval student of canon law, laid great emphasis on the primacy of the Roman see, accepting as genuine certain canons that were the work of 8th- and 9th-century forgers—such as two principles that the 1917 Code of Canon Law restates: "that there cannot be an ecumenical council which is not convoked by the Roman Pontiff" and that "the First See is under the judgment of nobody."

The prevalence of such ideas and the absence of a formidable challenge to papal primatial claims during the High Middle Ages explains the lack of any conciliar definition of the Roman primacy at the great "papal" general councils of that period. Hence it took the (abortive) attempt at reunion with the Orthodox Church at the Council of Florence in 1439 to evoke the first solemn conciliar definition of the Roman primacy. This definition was included in the decree of union with the Greeks (*Laetentur Coeli*), and it went as follows:

The first conciliar definition of Roman primacy

> We define that the Holy Apostolic See and the Roman Pontiff hold the primacy over the whole world, that the Roman Pontiff himself is the successor of Peter, prince of the Apostles, that he is the true vicar of Christ, head of the whole church, father and teacher of all Christians, and [we define] that to him in [the person] of Peter was given by our Lord Jesus Christ the full power of nourishing, ruling and governing the universal church; as it is also contained in the acts of the ecumenical councils and in the holy canons.

Early modern and modern views. This decree was the basis for the solemn definition that Vatican I promulgated in 1870 as part of its dogmatic constitution (*Pastor Aeternus*). Having asserted as a matter of faith the primacy of Peter and the succession of the popes in that primacy and having quoted in full the Florentine definition, the constitution clarified what is to be understood by "the full power of nourishing, ruling, and governing" the church, which, according to that definition, inhered in the pope's primacy. Unlike the conciliar definition arrived at in Florence, *Pastor Aeternus* specified this to include the pope's judicial supremacy, insisting that there is "no higher authority," not even an ecumenical council, to which appeal can be made from a papal judgment.

This definition marked the culmination of a development reaching back at least to the 4th and 5th centuries. But the doctrinal development that culminated in Vatican I's definition of papal infallibility cannot lay claim to a comparable antiquity. There has always been much discussion about the meaning of the prerogative of infallibility and what it implies about the status of individual doctrinal pronouncements of the church's teaching authority. The notion that the church (conceived as the community of the faithful) is by virtue of Christ's own promise infallible—in the sense that it cannot totally deviate from the truth—is clearly scriptural in foundation and was not questioned even by the Protestant Reformers. Similarly, the notion that a preeminent authority attached to the doctrinal pronouncements of the Roman Church and its bishops was of great antiquity, long predating the exten-

sion of papal jurisdictional claims by the 4th- and 5th-century popes. But the combination of these two notions—*i.e.*, the identification of the supreme teaching authority of the universal church with that of the pope, and the claim that the infallibility promised to the church itself was possessed also by the pope acting as its head, thus guaranteeing the inerrancy even of his individual doctrinal pronouncements—is essentially a modern theological development and one characteristic primarily of the Roman or Ultramontane (propapal) theological school. This school rose to prominence in the 16th and 17th centuries; one of its most distinguished representatives was Cardinal Robert Bellarmine (d. 1621). Though it drew from earlier materials—notably from the Pseudo-Isidorian decretals and from the writings of such medieval theologians as St. Thomas Aquinas, Aegidius Romanus, and Augustinus Triumphus—the Ultramontane school derived much of its initial strength from the papalist reaction that followed in the wake of the conciliar movement, and it was shaped very much in opposition to the claims that the conciliarists and their Gallican successors made on behalf of the general council. This is evident in the solemn definition of the doctrine promulgated by Vatican I, with its insistence that the ex cathedra definitions of the pope (those made from "the chair," or papal throne), "are irreformable of themselves and not by virtue of the consent of the Church." The conciliar debates indicate that this sentence was intended to exclude the Gallican notion that a papal definition could not claim infallibility unless, subsequently or concomitantly, it received episcopal assent. Despite the maximalist (extremist) tendencies both of subsequent Catholic apologists and of their Protestant critics, the sentence apparently was not intended to restrict the church's infallible teaching authority to the pope alone or to suggest that the pope was free to define doctrine without making every effort to take into account the mind of the church.

Ultramontanism

Nevertheless, after 1870, when the memory of the heated conciliar debates had faded away, maximalist interpretations became prominent. In particular, there was a marked tendency to stress the absolute and unlimited nature of papal jurisdictional power and to end in favour of the papacy the hitherto unresolved question of the source of episcopal jurisdiction. In response to this development, Vatican II, in its dogmatic constitution, *De Ecclesia* (1964), while endorsing Vatican I's teaching on papal primacy and infallibility, also focused on the nature of episcopal authority. It insisted that bishops "are not to be regarded as vicars of the Roman Pontiff, for they exercise an authority which is proper to them," since, "by divine institution . . . [they] . . . have succeeded to the place of the apostles as shepherds of the Church" and are themselves, in fact, "the vicars and ambassadors of Christ." Also, "Just as, by the Lord's will, St. Peter and the other apostles constituted one apostolic college, so in a similar way the Roman Pontiff as the successor of Peter and the bishops as the successors of the apostles are joined together." This college, "together with its head, the Roman Pontiff, and never without this head" is "the subject of supreme and full power over the universal Church," a supreme authority that it can exercise in more than one fashion but "in a solemn way through an ecumenical council." The supreme authority in the church can be exercised not only personally by the pope himself but also in a collegial fashion by the *whole* episcopate, which of necessity includes the bishop of Rome as its head.

Episcopal collegiality

In so emphasizing the doctrine of episcopal collegiality, Vatican II was responding to the findings of modern New Testament and patristic scholarship concerning the nature of the primitive and ancient church, and it insisted that it was restoring an ecclesiological emphasis of great antiquity. Recent medieval scholarship indicates that this emphasis persisted into the Middle Ages and survived, in the writings of canonists and theologians, side by side with the more prominent concern with the papal primacy. The great conciliarists active at the Council of Constance made an unsuccessful attempt to effect a stable balance between these two emphases, and even in the modern period, despite the growing prominence of Ultramontane views and their eventual triumph at Vatican I, the collegial concern was never fully displaced. It was not lost sight of by the Gallican theologians of the 16th, 17th, and 18th centuries who, however much their subservience to the exigencies of royal policy may have damaged their credibility, apparently are now recovering in Catholic eyes, at least, a certain measure of esteem.

Eastern Orthodox views and critiques. The recovery of this ecclesiological emphasis has an importance outside Roman Catholic theology. It has never ceased to dominate in the churches of Eastern Orthodoxy, with their stress on episcopal equality, their respect for the autonomy of the national or regional churches, and their insistence that the supreme teaching authority in the universal church resides (if anywhere) in the collegial decisions of the bishops assembled together in an ecumenical council. Up to the 11th century Byzantine churchmen and theologians certainly accorded some sort of primacy to the church of Rome and its bishop. But with the growth of papal claims to a universal jurisdictional power, with the growing conviction that the Roman Church had fallen into heresy, and above all with the disastrous crusading onslaught on Byzantium in 1204, the attitude of Orthodox churchmen to Rome underwent an understandable shift. Though Byzantine theologians rarely questioned the fact of Peter's primacy among the Apostles, they concluded that their own fundamentally collegial ecclesiology necessitated the rejection of the primatial claims advanced on behalf of those who claimed to be uniquely his successors. The very attempt by the bishops of a single local church to claim a monopoly on the Petrine succession was regarded as something of a deviation, in that all bishops, insofar as they professed the faith of Peter, were to be understood as his successors.

Rejection of Roman primatial claims

In the modern period, then, the Eastern Orthodox churches have been unanimously adamant in their rejection of the papal claims to primacy and infallibility. Orthodox theologians are often careful to insist that what they are rejecting is not the notion of primacy itself but rather that actual primacy of jurisdiction as it was conceived in the Latin Middle Ages and as it has been exercised by Rome in the modern period—with its apparent corollary that all power in the church is to be regarded as proceeding outward from the primatial office and its concomitant tendency to stifle independent life in the local churches. The original primacy of honour, which these theologians argue, was one accorded to the Roman bishops by emperors and ecumenical councils, they clearly regard as a different matter altogether. Given this fact, and also the common ground shared in ecclesiological matters by the Roman Catholic and Orthodox churches, Vatican II's affirmation of episcopal collegiality may soften the edges of the Orthodox rejection of the papal primacy.

Protestant views and critiques. The impact of that doctrine on Protestant thinking is more difficult to predict. Historically, the Protestant rejection of papal claims has been much less qualified than that of the Orthodox. Thus, in the view of the 20th-century theologian Karl Barth, Vatican I's definition of papal infallibility completed the process by which the Roman Catholic Church abandoned the Christian belief in the unique character of divine revelation, identified itself instead with that revelation, and made the pope's teaching "the infallible revelation for the present age." Philipp Melanchthon, the Lutheran author of the Augsburg Confession of 1530, may have been willing to admit that a truly evangelical pope had a certain superiority over other bishops; however, even if one deemphasizes Luther's denunciations of the pope as the Antichrist, the rejection by the major Reformers and their successors of the Petrine theory and of papal primacy by divine institution is absolute. Peter, they argued, exercised no primacy. The powers communicated to him were the powers communicated to all the Apostles. By the "rock" Christ meant himself; upon him the church is founded, and in Matthew 16:17 f. Peter stands only as the type or figure of the Christian faithful who believe in Christ as the sole "rock."

Unlike such medieval predecessors as the Waldenses or the political philosopher Marsilius of Padua, who had likewise attacked the Petrine theory, the Reformers did

not base their attack upon the historical argument that Peter had never visited Rome. This argument was embraced by many a liberal Protestant theologian of the 19th century, but in the 20th it has lost most of its appeal. In the mid-20th century some Protestant theologians shifted toward the Roman Catholic understanding of the status and meaning of Matthew 16:17 f. According to Oscar Cullmann (the French Protestant biblical critic and theologian) any sound exegesis of the relevant scriptural texts points to the conclusion that Peter enjoyed a preeminence among the disciples even during Christ's lifetime; that the "rock" in Matthew refers not to Christ or to the faith of Peter but to his person; that Christ promised him, therefore, the leadership of the church; and that after the Resurrection Peter actually exercised that leadership. Though Cullmann argued that Peter did so only for a short time, being replaced in the leadership by James, other Protestant scholars have disagreed and have claimed for Peter a more enduring role. All, however, continue with Cullmann to distinguish sharply between conditions in the apostolic and post-apostolic church, to deny on exegetical grounds that the "primacy," or leadership, promised to Peter was intended to be passed on to any post-apostolic successors, and to insist on historical grounds that no such succession in the primacy actually occurred in the primitive church. Nevertheless, because of the degree of convergence already occurring between the Roman Catholic and Protestant exegesis of the Petrine texts, because of the reexamination of the Catholic tradition begun by Vatican II, and because of the growing Protestant sense of the need for some striking symbol of unity in the worldwide Christian community, some Protestant ecumenists in the last third of the 20th century have shown a degree of openness to the papal office that would have been unimaginable only 50 or 60 years before.

Modern Protestant reinterpretation of Roman primatial claims

Historical conceptions of the relationship of the papacy to the world. Theories concerning the relationship of the papacy to the world at large have both reflected the established political conceptions of the day and been in tension with them. The pope has been conceived successively as a leading dignitary in an imperial church headed in effect by the emperor, as a majestic potentate possessed of a supreme and direct authority even in temporal matters, and as a primarily spiritual figure who had in temporal matters no more than an indirect power of intervention. With the post-Reformation fragmentation of Christendom, the growth of secularism, and the emergence of the unified modern state claiming within its own borders jurisdictional omnicompetence, even such attenuated claims to an indirect power became increasingly anachronistic. In the 20th century, in his relations with the world at large, the pope, while affected by the conventions regulating the relationships of heads of state with one another, possesses primarily a moral authority deriving from the dignity and prestige of his office. The strength of that authority, however, depends upon his moral standing as a person, upon the persuasive force of his cause, and upon the degree of enthusiasm it can arouse within the church.

Contemporary teaching on papal authority. After the mid-20th century some voices were raised in Roman Catholic circles questioning both the doctrine of papal infallibility and the exercise of the papal primacy—at least as it is envisaged in the teaching of Vatican I and the Code of Canon Law. The church's official teaching on the papal office remains that of Vatican I, solemnly reaffirmed at Vatican II. Nevertheless, the latter council's juxtaposition of the doctrine of episcopal collegiality with the existing teaching on papal primacy and infallibility created something of a dilemma in Catholic ecclesiology. Though the text of *De Ecclesia* had insisted that the doctrine of episcopal collegiality in no way impugned the pope's primacy, a minority of the council fathers remained unconvinced and were commonly said to have been won over by the explanatory note that the Theological Commission by papal authority appended to the decree as an "authentic norm of interpretation." The note is framed in much more juristic terms than is the decree itself, and, in discussing the possession by the College of Bishops of "supreme and full power over the whole Church" it insists that "there is no distinction between the Roman Pontiff and the bishops taken collectively," that "necessarily and always, the College carries with it the idea of its head" so that the bishops acting independently of the pope cannot be considered to constitute a college. At the same time, the note insists that "since the Supreme Pontiff is the head of the College, he alone can perform certain acts which in no wise belong to the bishops, for example, convoking and directing the College, approving the norms of action etc.," norms that "must always be observed."

Recent questioning of the doctrines of papal infallibility and primacy

Already in 1964 there were some who regarded this note with considerable misgiving, feeling that it withdrew from the bishops, in practical and legal terms, that supreme authority in which they had been said, on theological grounds, to be sharers. Subsequent events did little to dispel such misgivings. Despite the unquestionable vitality shown at its 1967 and 1969 meetings, the Synod of Bishops was not really allowed to function as a decision-making rather than a merely advisory body, and it was no more consulted than were the bishops as a whole when, in 1968, the pope promulgated *Humanae Vitae* (the encyclical on birth control)—considered by some observers to be the most divisive papal initiative of recent times and one that amounted to a de facto negation of collegiality.

Humanae Vitae

Because of the dissent over *Humanae Vitae* and the tension engendered by the rigour of the pope's stand on the much-debated problem of clerical celibacy, attention probably will focus increasingly on the old and difficult question of the limits of papal power. Because of this, considerable importance attaches to the current revival of interest in the late medieval conciliar movement and to the assertion made by some Roman Catholic scholars (if hotly disputed by others) that a continuing dogmatic validity must be accorded to the decree *Sacrosancta,* promulgated in 1415 by the Council of Constance. This decree declared that the general council possessed an authority superior to that of the pope in matters pertaining to the faith, the ending of the schism, and the reform of the church. Those who assert this view do not always wish by so doing to cast any doubt on the dogmatic validity of Vatican I's teaching on papal primacy and infallibility, but the efforts thus far made to demonstrate the compatibility of the respective teachings of the two councils (*i.e.,* Constance and Vatican I) remain somewhat less than persuasive.

THE OFFICES OF THE CLERGY

The Roman Curia and the College of Cardinals. In the day-to-day exercise of his primatial jurisdiction the pope relies on the assistance of the Roman Curia, a name first used of the body of papal assistants in the 11th century. The Curia had its origins in the local body of presbyters (priests), deacons (lower order of clergy), and notaries (lower clerics with secretarial duties) upon which, like other bishops in their own dioceses, the early bishops of Rome relied for help. By the 11th century this body had, on the one hand, been narrowed down to include only the leading (or cardinal) presbyters and deacons of the Roman diocese, while, on the other hand, being broadened to embrace the cardinal-bishops (the heads of the seven neighbouring, or "suburbicarian," dioceses). From this emerged the Sacred College of Cardinals, a corporate body possessed, from 1179 onward, of the exclusive right to elect the pope. This right it still possesses, as it does the right to govern the church in urgent matters during a vacancy in the papal office. Recent popes have extended the size of the Sacred College beyond the traditional limit of 70 and have attempted, with growing success, to broaden its national complexion and to make its membership more representative of the church's international character.

Cardinals are selected by the personal choice of the pope, in consultation with the cardinals in Rome at the time, in a consistory, or solemn meeting, which is secret. The cardinals reside either as bishops in their own sees or in the Vatican as the highest rank of papal advisers and officers in the Roman Curia.

During the Middle Ages the cardinals played an important role as a corporate body, not only during papal vacancies, as today, but also during the pope's lifetime. In the 12th century the Roman councils that popes had

hitherto convoked when urgent matters were at hand were replaced by the assembly of the cardinals, or consistory, which thus became the most important collegial (corporate) body advising the pope and participating in his judicial activity. Eventually it began to make oligarchic claims to a share in the powers of the Petrine office and attempted, with sporadic success, to bind the pope to act on important matters only with its consent. During the 16th century, however, with the final establishment of the Roman congregations (administrative committees), each charged with the task of assisting the pope in a specific area of government, the significance of the consistory began to decline, and with it the importance of the cardinals as a corporate body. At the same time, there was an increase in the power and influence of the "curial" cardinals—those cardinals who did not administer local dioceses but served as the pope's representatives in important foreign affairs or resided permanently in Rome, holding responsibilities in the curial congregations, tribunals, and offices that proliferated in the course of the next three centuries.

Modern reforms of the Curia

By the early 20th century the growth of the Roman Curia had produced a bewildering tangle of administrative and judicial bodies, in which neither temporal and ecclesiastical functions nor executive and judicial powers were clearly demarcated. The reforms of Pius X (reigned 1903–14) and Benedict XV (reigned 1914–22) clarified and streamlined the work of the Curia, introducing a measure of order into its maze of overlapping jurisdictions. But in the wake of the complaints about abuses of curial power that were voiced at the second Vatican Council (along with requests for an internationalization of curial staff and a modernization of curial functions and procedures), Paul VI pledged himself to act.

Though Paul VI (reigned 1963–78) made some changes in detail, his reforms left intact the basic curial structure created by Pius X, with its tripartite division of the various curial bodies: the Roman congregations composed of cardinals nominated by the pope (*e.g.*, the Congregation for the Doctrine of the Faith, which was the former Holy Office and direct descendant of the Roman Inquisition); the tribunals, three in number, which compose the judicial branch of the Curia and one of which, the Rota, handles matrimonial cases; a group of offices, councils, and secretariates, the most important of which is the Secretariate of State, presided over by the cardinal secretary of state, who now emerges as the pope's "prime minister." To promote a higher degree of coordination among the various jurisdictions, provision was made for regular meetings of department heads, summoned and presided over by the cardinal secretary of state. Similarly, to prevent bureaucratic empire-building, most curial appointments were to be made for an initial term of five years. Finally, in response to Vatican II's request, some diocesan bishops were to be present at the plenary sessions of the congregations, efforts were to be made to internationalize the curial staff, and there was to be some attempt to consult the laity. These reforms went into effect in January 1968.

(F.C.O./Ed.)

The college of bishops. It has been noted that in Roman Catholicism the college of bishops is the successor to the college of the Apostles. This is said in spite of certain differences between the two offices. The Apostles in the New Testament were a college (except for Paul, not one of the Twelve); the bishops are individual officers, and their collegial function has not been operative in recent centuries. The Apostles had a power that was not defined locally; every Roman Catholic bishop is a bishop of a place, either a proper area, a jurisdiction, of which he is the ordinary (as he is called in church law), or a fictitious place, a see no longer existing, of which he is named titular bishop. Such a monarchical officer does not appear in the New Testament. Nevertheless, Ignatius of Antioch, whose letters (written about 107) provide an early description of the Christian community, was clearly a monarchical bishop, and he did not think himself the only one of his kind; thus, the institution must have arisen in apostolic or early post-apostolic times.

Monarchical officers

The bishops in Roman Catholic belief succeed to the apostolic power, which is understood as the power to teach Catholic doctrine, to sanctify the church through the administration of the sacraments, and to govern the church. The residential bishop is supreme in his territory in this threefold function, having no superior other than the Roman pontiff. An archbishop governs a metropolitan see, usually the largest or oldest see in a region of several dioceses called a province. The metropolitan archbishop convokes and presides at provincial synods, or meetings, and has certain rights of visitation; but he has no jurisdiction in the suffragan, or subordinate, sees. The power of the bishop in governing is only over his own diocese; even there, however, it is not absolute, because church law provides the bishop with certain advisory bodies.

Until the second Vatican Council the Roman Catholic Church had not dealt with the ambiguity of two concurrent jurisdictions, pontifical and episcopal. The pope cannot define or limit the powers of a bishop; the powers are "ordinary," inhering in the office itself. The second Vatican Council accepted the emphasis that recent theologians have laid on the collegial character of episcopacy, and the supremacy of the pope is understood as supremacy in the college; the pope needs the college of which he is head, although the first Vatican Council declared that he needs neither its consultation nor its approval. It is now understood that such solitary action should be the emergency rather than the rule; and the synod of bishops, established after the second Vatican Council, was a step toward involving the body of bishops in the policy of the entire church, hitherto formulated exclusively by the Roman see.

The qualifications for a bishop as defined in church law, which is known as canon law, are so general as to suit candidates for any office in the church, major or minor. Bishops have been chosen by the pope since the 11th century; the qualifications are not made public, and the decisions depend on a number of factors that are difficult to assess. In modern practice most bishops have been career administrators in the church, rarely pastors or scholars. Election is a much older tradition, and there have been many calls for a restoration of election of bishops. But because the selection of bishops is a basis of Rome's power over the whole Catholic Church, Rome has been generally unsympathetic to these calls.

Modern duties of a bishop

Bishops in modern times are more visible as managers of the business of the diocese than as pastors and teachers. The responsibilities of the office are great and demand leadership and the ability to delegate business to a competent staff. A common criticism from certain quarters within the Catholic Church in the mid-20th century was that the episcopacy has been conceived more in terms of power than in terms of leadership. One of the reasons for setting up episcopal conferences of nations and regions following the second Vatican Council was to promote leadership by giving bishops the strength that lies in community. No authentic "Catholic" activity is conducted in a diocese without at least the tacit approval of the bishop; his disposal of funds and persons makes it evident that the activity will flourish much more vigorously if it enjoys his active support and encouragement. His power to discourage or forbid activities, which he is free to use according to his own sole judgment, is both a strength and a weakness of the Roman Catholic structure.

The bishop is assisted in governing the diocese by a staff called, like the staff of the pope, a curia. The structure of the staff is to some extent determined by canon law—*e.g.*, vicar general, chancellor, and official, or head, of the diocesan tribunals. Otherwise, the bishop at his discretion may appoint a staff according to the needs of the diocese. In modern times a great amount of diocesan business has settled upon the bishop. This has become a constant strain on the structure of the Roman Catholic episcopacy; it cannot be a strength when so much of the time of the supreme officer must be consumed by purely routine business.

The power of the bishop over his clergy has been absolute, with almost no effective restraint except the human kindness of the bishop. This appeared to be changing, following the second Vatican Council, with the institution of senates of priests. The council strongly recommended that bishops introduce priests into the decisions of the diocese, but this was left to the discretion of the bishop.

The settling of doctrinal and disciplinary questions

Ecumenical councils. Regional councils of bishops to settle doctrinal and disciplinary questions appeared in the 2nd century. The first general council representing the bishops of the whole world (the Greek *oikumenē* referred to the inhabited world) occurred at Nicaea in Asia Minor in 325. The council was convoked not by an ecclesiastical authority but by Constantine, who wished to have a final decision on the Arian controversy. (According to Arius, the Son of God was a creature of similar but not the same substance as God the Father.) The representatives of Constantine's bishop, the bishop of Rome, presided over the council. The Roman Catholic Church has held 21 such assemblies. The chronological distribution of the last three (Trent, 1545–63; first Vatican, 1869–70; second Vatican, 1962–65) shows that in modern times the ecumenical council has been convened less frequently.

Canon law defines an ecumenical council and its procedure; actually, the law represents the procedure followed in the convocation of the first Vatican Council. There is no real criterion for an ecumenical council, and one can say only that those councils are ecumenical that the Roman Catholic Church regards as ecumenical. The Orthodox Churches recognize the first eight only.

The ecumenical council is recognized by the Roman Catholic Church as the supreme authority. With the pope this makes two supreme authorities; the Roman Church reconciles this logical dilemma by asserting that the ecumenical council, acting with the pope, is supreme. Only the pope can convoke an ecumenical council, and he or his legates must preside. There are no limits to the competence of an ecumenical council, but its decrees must be approved by the pope for validity.

The scandals of the Great Western Schism, which at its worst saw three men claiming the papacy, and the corruption of the papal court during the 15th century led to the movement of conciliarism, according to which the ecumenical council was the means of saving the church from scandal and corruption. Much of the policy of the Roman see since that time has been devoted to the suppression of any conciliarist sentiments. This has naturally led to questions about the value of ecumenical councils, which are cumbersome and expensive, when an omnicompetent office such as the papacy is prepared to handle the business of the Roman Catholic Church. Both the first and second Vatican councils illustrated the values of the ecumenical council. Apart from the public and psychological impact produced by a consensus so broad, the council not only makes available for the church a fund of worldwide wisdom and experience not available to the Roman Curia but also seems to generate a state of mind that raises the members of the council above their normal level of thought and action.

Assistant to the bishop and cultic officer

The priesthood. The title of priest (Greek *hiereus*) is given to no church officer in the New Testament. Nevertheless, the office appears in the 2nd century, no doubt with the development of the monarchical episcopate; the bishop needed assistance in his threefold task of teaching, sanctifying, and governing, and the priest exercised this power as an officer of the bishop. A priest is either a member of a diocese or of a religious community; but in the exercise of the threefold ministry every priest is subject to the bishop of the diocese where the ministry is conducted.

The priest is by definition a cultic officer, and the title designates the second element of the office, the work of sanctification. Certain ambiguities in the Roman Catholic clerical hierarchical system appear clearly in the priesthood. Ordination empowers the priest to administer the sacraments, but he cannot use this power except by receiving "faculties" (proper permission or license) from a bishop. Teaching and preaching are not powers conferred by ordination, but they are subject to the same "faculties." The priest is lowest in the system of government and actually does not govern unless he is a pastor. Governing is not exercised by curates (priests who assist the pastor) or by the large number of priests engaged in specialized works that can hardly be called ministries: administration, teaching, scholarship, journalism, and other activities.

The pastor of the parish is the model priest; in spite of the fact that in large parishes the pastor may be primarily an administrator, Catholics experience their church directly through the parochial clergy. Catholics hear sermons, worship, receive the sacraments, and look for religious counsel and direction in their parish. Many Catholics, particularly in the United States, have their children educated in a school run by the parish. The parish is also the centre of activities ranging from recreation to adult education and to social works, all under the direction of the clergy. In Roman Catholicism the parochial clergy are genuine pastors; the pastoral office has often been reduced for the bishop and is barely visible in the pope. The strength of the Roman Catholic Church historically has been rooted in its priests, especially in its parochial clergy.

Roman Catholicism for centuries has fostered a distinct clerical identity, symbolized by clerical garb, which sets priests as a class apart not only from non-Catholics but from Catholics. The most striking feature of this class, celibacy, has stirred up considerable dissatisfaction in the modern church with celibacy as well as a feeling that it interferes with the ministry. Critics point out that neither in the New Testament nor in the pre-Constantinian church was there a clerical class; the whole church was a people set apart with a mission to the unbelieving world. Because of this dissatisfaction with celibacy and issues related to it there have been a significant number of departures from the priesthood and an alarming fall in the number of candidates.

RELIGIOUS COMMUNITIES

Religious communities in the Roman Catholic Church consist of groups of men or women who live a common life and pronounce vows of poverty, chastity, and obedience (the evangelical counsels). The aim of such a life has traditionally been regarded as the achievement of Christian perfection (theologically defined as perfect love); thus it is an option only for a minority of the members of the church. Roman Catholic theology has never quite rationalized the elitism implicit in this idea nor escaped the implicit denigration of the lay state; but up to modern times both religious and seculars have overcome the need for rationalization by mutual respect and mutual services.

Hermits and monks. The origins of the religious life are seen in the anchorites, or hermits, of the 2nd and 3rd centuries, who escaped sin and temptation by flight from the world, mostly in the deserts of Syria, Egypt, and Palestine. Flight from the world became the rule of the cloister, forbidding both free entrance of "externs" into the enclosure and free egress of religious from the enclosure and imposing supervision in all dealings with seculars. The evangelical counsels meant a life of solitude and destitution and an effort to attain union with God by prolonged, almost constant contemplation. Where large numbers of hermits assembled in the same place, cenobitism (common life) emerged, and the hermits or monks (Greek *monachos*, "solitary") elected one of their members abbot (Aramaic *abba*, "father"). Eastern monasticism produced the rules of Pachomius and Basil in the 4th century, and travelers (most notably John Cassian) introduced monasticism into the Latin Church. Eastern monasticism, principally because of a lack of discipline, dissipated much of its energy and had no further influence on the West. Western monasticism was dominated by the rule of Benedict of Nursia in Italy, who founded his communities in the 5th century.

The Benedictine Rule emphasized less austerity and contemplation and more common life and common work in charity and harmony. It has many offshoots and variations, and it has proved itself sturdy; it is the longest continuous religious community in the Roman Catholic Church, and it has survived many near collapses and reforms. The monk did not join an "order" but a monastery. Benedictine monasteries were almost always located in remote areas. However, because the labour of the monks transformed them into food-producing areas, they became centres of settlement. Thus the monks who had fled the world found that the world sought them out for services, which they gladly rendered. Whatever charitable works existed, were done by them. The monks were also the only people who did anything to preserve the learning of

antiquity. They supported church reform and furnished many reforming popes and bishops. Benedict did not put contemplation into his rule; prayer was fulfilled by the chanting of the divine office (a set form of liturgical prayer), celebrated at specific times during the day.

Mendicant friars and clerks regular. The 13th century saw the rise of the mendicant friars (Franciscans, Dominicans, Carmelites, Augustinians). The friary was like a monastery, with common life and the divine office in choir; but the friars made excursions, sometimes at great length both in time and distance, for apostolic works, mostly preaching. All of the mendicant orders had apostolic work in mind in their foundation, and they desired a mobility that was had neither by the monks nor the diocesan clergy. They were thus at the ready disposal of the pope, and the principle of clerical exemption (exemption from the jurisdiction of the bishop) became much more important than it had been for the monks. Originally, the friars did not need even the approval of the bishop to preach in his diocese, although this freedom has been restricted in modern times. Preaching became almost the specialty of the mendicant friars in the Middle Ages, and they were important in the foundation of the universities of the Middle Ages.

The 16th century saw the emergence of the third major form of religious life, that of the clerks regular. These communities were formally and frankly directed to the active ministry. Even the friary, with the divine office in choir and other monastic restrictions, was dropped; they wore no distinctive religious habit. According to Ignatius of Loyola, founder of the Society of Jesus (Jesuits), the best-known example of clerks regular, their life imitated the manner of living of devout secular priests. The Jesuits, almost by accident, had no particular ministry and placed themselves at the disposition of the pope. The clerks regular had even greater mobility than the friars and had the resources to undertake specialized works. Since the 16th century the works of religious communities have been education, foreign missions, preaching, and theological scholarship. Orders founded since the 16th century have adopted the manner of life of the clerks regular.

Nuns and brothers. Religious communities of women until the 17th century were entirely contemplative and subject to rigid cloister, although from the 16th century onward they began to admit girls into the convent not as novices (those admitted to probationary membership in the community) but to educate them as gentlewomen. The modern communities of women all stem from the type of community instituted in France in the mid-17th century by Vincent de Paul under the name of the Daughters of Charity. At first these women were not religious and deliberately so; Vincent did not wish cloister. The group was founded to help the poor and sick and to train their children in religion and the rudiments of education. These have remained the major works of the communities of women.

Religious communities are orders if the members (or some of them) pronounce solemn vows; they are congregations if the members pronounce simple vows. Solemn vows are perpetual; simple vows may be perpetual or temporary. The difference is subtle; solemn vows, although dispensable, were meant to be a more permanent and durable consecration than simple vows. Men who make religious profession but who do not receive the sacrament of holy orders are "brothers."

Secular institutes have arisen since World War II. They are not religious (and therefore do not pronounce the three vows), have little or no common life in a common residence, have no superior but rather a manager of the few common affairs, and intend to bear Christian witness in the world in any type of secular employment.



Mendicant friars preaching the gospel, 13th-century manuscript illumination. In the Bodleian Library, Oxford (MS. Douce 180).

THE LAITY

The laity as a class do not appear in the New Testament; there could only be a laity when a clergy had come into being. When the laity appear, they are the passive element of the church. If the office of the clergy is conceived as teaching, sanctifying, and governing, then the function of the laity is to be taught, sanctified, and governed. Misleading identification of the church with the clergy (and, within the clergy, with the hierarchy) results.

The modern term Catholic Action (especially under Pius X and Pius XI) meant in general the assistance of the laity in the mission of the church. Yet, as it was more closely defined, the mission of the church was still entirely clerical, and lay action was accessory to the mission proper. The laity were merely the arm of the hierarchy. Furthermore, lay action fell under close direction and supervision of the hierarchy and clergy. It is not surprising that an action so vaguely defined, so patronized, and so uninspiring aroused relatively little response.

Anticlericalism and clericalism

Much of the 19th and 20th centuries saw the Roman Catholic Church engaged with anticlericalism in the "Catholic" countries of Europe; this seems to be a peculiarly Roman Catholic phenomenon. Actually, anticlericalism is a rejection of the medieval belief in the power of the clergy to direct all the decisions of the layperson that they thought themselves entitled to direct. Reaction in an exaggerated form nearly excluded the clergy from any activity except public worship in some countries.

The second Vatican Council definitely rejected clericalism. It called "secular" all nonecclesiastical activity and declared that the secular is the proper area of the layperson. This means that laypersons are the judges of how to realize their Christian destiny in the secular sphere. Proper does not mean exclusive, but the statement implies that the clergy can offer only principles and general directions, not make specific decisions. The Roman Catholic Church intended to make the laity the channel of its relevance in the world.

The council also took steps against the passive role of the laity in ecclesiastical life. It recommended the establishment of lay councils in each diocese and in each parish. This has moved slowly because Roman Catholics are not accustomed to the idea and are uncertain about how it should be implemented. As the secular is the proper but not the exclusive area of the laity, so the ecclesiastical is the proper but not the exclusive area of the hierarchy and the clergy.

CANON LAW

The earliest individual church law was called a canon (Greek *kanōn*, "rule, measure, standard"); the canons were finally called Canon Law. Church laws appear almost as soon as church authority, and some passages of the New Testament reflect early rules; whether they should be called law at this primitive stage is doubtful. Laws of dioceses or of regions appear even before Constantine; they were formed by diocesan synods or regional councils. Laws for the whole church appear with the earliest ecumenical councils. Canon Law remained scattered pieces

of papal, conciliar, and diocesan legislation until the 12th century. The first collection and synthesis of Canon Law was made by Gratian in 1142, the *Decretum Gratiani.* To this collection in the next 400 years were added the decretals (papal decrees on points of law) produced in the reigns of Gregory IX (1234), Boniface VIII (1298), and John XXII (1317) and two collections known as *Extravagantes* (1500). These formed the *Corpus Juris Canonici* ("Body of Canon Law"); no further collection of laws was made later than the *Corpus.* Effectively, although not formally, Canon Law included the opinions of canonists interpreting the *Corpus.*

This unsatisfactory and cumbersome collection led to calls for codification. No doubt the desire was influenced by the production of the Napoleonic Code, which became the basic law of most of the nations of western Europe. The codification was begun by a document of Pius X (1904) and was completed, directed by Cardinal Pietro Gasparri throughout, under Benedict XV (1917); it became law in 1918. This code remained the basic law of the Roman Catholic Church until 1983, when a new *Codex Juris Canonici* was instituted.

The history and structure of church law are treated more fully under CHRISTIANITY.

Beliefs and practices

FAITH

Concepts of faith. The idea of faith shared by all Christian churches is rooted in the New Testament. But the New Testament idea of faith is not simple, and it permits a breadth of meaning that has led to variations even within a single Christian communion. Most modern interpreters of the New Testament would agree to a description of New Testament faith as a total commitment of the self to God revealing himself in Christ. Yet it is doubtful whether the post-Reformation theology of any Christian church has presented faith simply in these terms.

Even before the Reformation, faith in Roman Catholicism had developed an emphasis that is not rooted in the New Testament but can be traced back to the Alexandrian school of theology and to Augustine. Faith appeared primarily as acceptance of revelation, and revelation appeared as a revelation of doctrine rather than as revelation of a person. This emphasis ultimately was formulated in the 13th century by Thomas Aquinas in a definition of faith—canonized by the Council of Trent and the first Vatican Council—as an intellectual assent given to revealed truth by the command of the will inspired by grace and motivated by the authority of God revealing.

Faith as an intellectual assent

The Reformers, with Martin Luther as the leader, rejected this idea of faith as nonbiblical and exclusively doctrinal; it seemed to place the teaching authority of the Roman Catholic Church between man and God not as a means of communication but as a replacement of God. Luther saw faith as confidence in the saving power of grace. This, Luther believed, was a return to the New Testament faith, but Roman Catholicism rejected this as a mere sentiment; these positions were crystallized up to the 20th century. At the risk of oversimplification, it is possible to say that both represented exaggerations of the New Testament. New Testament faith is more than either trust in the saving power and will of God or assent to revealed truth, although neither element can be entirely excluded. Efforts were wasted in trying to prove the adversaries wrong rather than in trying to understand the New Testament. The documents of the second Vatican Council reflect a shift in Roman Catholic theology from emphasis soley on faith as intellectual assent to recognition of faith as a loyal adherence to a personal God.

Roman Catholic theology, having chosen the option of faith as assent, was faced with the problems of showing that it was a rational assent rather than an irrational assent and of maintaining that faith was a deliberate and free meritorious act under the inspiration of grace. At first glance the two problems seem to cancel each other out; one can maintain one affirmation only by denying the other.

Preambles and motivation of faith. The study of the problems connected with faith involves the investigation of what are called the preambles of faith and also of the motivation of faith. The preambles of faith include those processes by which the believer reaches the conclusion that it is reasonable to believe—*e.g.,* the proof of the existence of God by the use of one's own reason. The freedom of faith is respected by affirming that this conclusion is as far as the preambles can take one. This process as proposed is a theoretical construction that actually occurs in no one, but the analysis can be of value in uncovering the psychological processes that occur without reflection. The preambles include the study of the scientific and historical difficulties raised against the Christian fact (*i.e.,* the incarnation, Resurrection, Ascension, and glorification of Jesus Christ) itself or against the Roman Catholic interpretation and proclamation of the Christian fact or against the Roman Catholic claim to be the exclusive custodian of revealed doctrine and the means of salvation. These studies were efforts to show what cannot be shown by scientific and critical methods, but in the exaggerated claims of their defenders they showed that faith was a necessary conclusion of a valid rational process. Such a faith could be neither free nor the result of grace.

The evidence for the act of faith

The study of the motivation of faith attempted to meet this difficulty. Some earlier analyses candidly presented faith as resting on evidence and clumsily postulated a movement of grace necessary to assent to this particular evidence. Normally, one "wills" to believe something because the evidence is not compelling; thus, people choose to believe that the candidate of their choice has the qualities desired for the office, although the evidence is less than overwhelming. The Roman Catholic thinks this is an assent to the probably rather than the certainly true and yet insists that the certainty of faith is the highest of all certainties. Ultimately, the Roman Catholic analysis must say that the evidence that belief is reasonable can never be so clear and convincing that it compels the radical deviation from worldly patterns that assent implies. At this point, the will inspired by grace chooses to accept revelation for other reasons than the evidence.

The motive of faith that has been presented by Catholic theologians is "the authority of God revealing." It is held that the preambles of faith show beyond reasonable doubt that God exists and that he has revealed himself. This evidence and an acceptance of the notion that, if God reveals himself, he does so authoritatively motivate a person to make the act of faith. The problem with such an analysis has been to define how the authority of the revealer is manifest to the believer. It seems that the notion of the authority of God revealing must be an object of faith rather than a motive, because the conjunction of this authority with the fact of revelation cannot be the object of historical experience. In the mid-20th century this dilemma caused an increasing number of Catholic theologians to move closer to a view that emphasized faith as a personal commitment to God rather than as an assent to revealed truth.

Heresy. Heresy is the denial by a professed, baptized Christian of a revealed truth or that which the Roman Catholic Church has proposed as a revealed truth. The unbaptized person is incapable of heresy, and the baptized person is not guilty of "formal" but only of "material" heresy if he does not know that he denies a revealed truth. The seriousness with which Roman Catholicism regarded heresy is shown by the ancient penalty of excommunication. Civil penalties, including the supreme penalty, did not appear until the Constantinian age. Lesser civil disabilities continued in force, although the law was often ignored, into the 20th century. Protestant governments often borrowed some of this severity from Roman Catholic governments.

Roman Catholic theologians often deal with heresy, paradoxically, as a necessary step in the development of dogma. In order to save themselves from an extremely crass and even cruel rationalization, they point out that the questions raised by heresy were legitimate but that heretics too quickly assumed a one-sided and exclusive view of doctrine that they wished to impose on the entire church. Modern studies have sometimes been less kind

to such champions of orthodoxy as Athanasius and Cyril of Alexandria, who were not themselves free of one-sided views and who showed themselves unwilling to listen to their adversaries with sympathy and understanding. In recent times most of the theses of Modernism (a movement to change the Catholic Church by means of radical renovation), which were condemned vigorously by Pius X in 1907, have found their way into Catholic theology. This may have something to do with the absence of the words heresy and heretics from the acts of the second Vatican Council. Like the use of the word church for Protestant churches, this indicates a substantial change of attitude toward a genuinely ecumenical position.

REVELATION

The concept of revelation. Although other religions have ideas of revelation, none of these bears a close resemblance to the idea of revelation found in the Old and New Testaments and in Christianity. Roman Catholic theologians distinguish between revelation in a broad sense, which means knowledge about God deduced from nature and man (and therefore actually philosophy), and revelation in the strict formal sense, by which they mean the utterance of God. This latter idea, of course, can only be conceived by analogy with the utterance of man, and its precise definition involves difficulties.

The earliest idea of revelation is the one found in the Old Testament in which the speech of God is addressed to Moses and the prophets. They in turn are described as quoting the words of God rather than interpreting them. Jesus, the fulfillment of the prophets, does not speak the word of God; he is the word of God. This phrase, which occurs only in the opening verse of both the Gospel and the First Letter of John, has become a technical term in theology; Jesus is the Incarnate Word. As such he is both the revealer and the revealed. He reveals the Father both by what he says and by what he is. Thus, the earliest Gospel (literally "good news") is the account of the life, death, and Resurrection of Jesus. The Gospel as the recital of his words appears in a later phase of development.

Jesus as revealer and revealed

It has been noted that the Roman Catholic Church has regarded revelation primarily as the revelation of propositions rather than the revelation of a person. Thus, even Jesus has been thought of more as a spokesman who tells of God than as a reality who himself in his being and actions manifests God. Though this latter aspect is found to some extent in the documents of the second Vatican Council, it has normally been considered only in the miracles of Jesus, which have been regarded in Roman Catholic apologetics as works of divine power that assure the credibility of the words of Jesus. These words, which were spoken in a particular historical context, have been preserved in a twofold way. They are written in the Gospels, which together with the Old Testament form a book of revelation that is distinct from the spoken words; but, because the Bible itself is written under divine inspiration, it has the same authority of revelation as the spoken words of Jesus. The Roman Catholic Church also preserves the words of Jesus, independently of the Bible, in its traditional teaching; but it does not utter the very words spoken by Jesus, and thus its words have a lower formal quality of revelation than the words of the Bible, although they are of equal authority. The idea of a book of revelation was taken by the early Christian Church from Judaism when it accepted the sacred books of the Jews as its own, just as it accepted the God of Judaism as the Father whom Jesus claimed for his own.

Mystery and the supernatural

The content of revelation. The proper content of revelation is designated in Roman Catholic teaching as mystery; this theme was important in the documents of the first Vatican Council. The development of the theme of mystery responded to those intellectual movements of the 18th and 19th centuries that are called by such titles as the Enlightenment, Rationalism, scientism, and historicism. To the Roman Catholic Church these movements were threats to the idea of a sacred revelation; they appeared to claim that human reason had no frontiers or that human reason had demonstrated that revelation was historically false or unfounded or that the content of revelation was irrational. The affirmation of mystery meant that the reality of God was unattainable to unaided human reason; theologians had long used the word incomprehensible, which says more than modern theologians wished to say. Mystery refers both to the divine reality and to the divine operations of the world. These operations can be observed only in their effects; the operation itself is not seen, nor is its motivation seen. The plan of God, which is realized in history, is mysterious. The first Vatican Council insisted that the existence of God and of a moral order is attainable to reason, and some of the fathers of the council wished to state that these truths were imposed upon reason by the evidence, a step that the council did not choose to take. Mystery does not mean the incomprehensible or the unintelligible; it means, in popular language, that man cannot know who God is or what God is doing or why God is doing it unless God tells him. Mystery also means that, even when the revelation is made, the reality of God and his works escapes human comprehension.

The term supernatural has been used in Roman Catholic theology since the 17th century to designate not only revelation but other aspects of the divine work in the world. The term has an inescapable ambiguity that has led many modern theologians to avoid its use. The "natural" that the supernatural presupposes is the world of human experience; the quality of this experience is not altered by technological and social changes as long as these are fulfillments of the potentialities of nature. Indeed, it is the spectacular growth in the knowledge of these potentialities in modern times that leads to doubt as to whether there can be a supernatural at all. The supernatural reality is identified with God in his reality and in his operations. This is a reality that man cannot create or control. The supernatural in cognition is this reality as it is perceptible to man; it is, for man, simply unknown as far as unaided reason can move. The first Vatican Council affirmed that without revelation human reason has not reached anything but a distorted idea of the divine and an imperfect idea of the moral order. This means also that human beings are unaware of their destiny, either individually or collectively, without revelation and that they are unable to achieve it without the entrance of the supernatural into the world of history and experience.

Contemporary theologians of revelation are aware of the problems raised by historical and literary criticism that render it impossible to cherish the primitive idea of revelation as the direct utterance of God to man. Roman Catholic theologians have not found a satisfactory way of describing revelation, but they do not see that the destruction of a naive idea of revelation destroys the whole idea. Theologians also recognize that the older idea of revelation of propositions as a collection of timeless and changeless verities, almost like a string of pearls, is no longer tenable. Every utterance that is called revelation was formed in a definite time and place and bears the marks of its history. There is no revealed proposition that cannot be restated in another cultural situation. Indeed, contemporary theologians are aware that these propositions must be restated if the Roman Catholic Church is to speak meaningfully in the modern world. Roman Catholicism does not accept the possibility of a new revelation; it believes that reason can never completely penetrate the "mystery" and that it must continue the exploration of the mystery that has already been revealed.

TRADITION AND SCRIPTURE

Tradition as channel and content

In Roman Catholic theology tradition is understood both as channel and as content. As channel it is identical with the living teaching authority of the Catholic Church. As content it is "the deposit of faith," revealed truth concerning faith and morals. In Roman Catholic belief, revelation ends with the death of the Apostles; the deposit was transmitted to the college of bishops, which succeeds the Apostles.

The Reformers contended that the Roman Catholic Church had imposed teachings that were not contained in the Scriptures, and this Protestant objection has been maintained in modern times. The objection was raised more intensely when the Immaculate Conception of Mary,

the mother of Jesus (Pius IX, 1854), and her Assumption (Pius XII, 1950) were defined as dogmas. For neither of these is there any biblical evidence; more significantly, there is no evidence in tradition for either before the 6th century.

The Roman Catholic Church recognizes that the Bible is the word of God and that tradition is the word of the church. In one sense, therefore, tradition yields to the Scriptures in dignity and authority. But against the Protestant slogan of *sola Scriptura* ("Scripture alone"), itself subject to misinterpretation, the Roman Catholic Church advanced the argument that the church existed before the New Testament. In fact, the church both produced and authenticated the New Testament as the word of God. For this belief, at least, tradition is the exclusive source; and this furnished a warrant for the Catholic affirmation of the body of truth that is transmitted to the church through the college of bishops and preserved by oral tradition (meaning that it was not written in the Scriptures). The Roman Church therefore affirmed its right to find out what it believed by consulting its own beliefs as well as the Scriptures. The Council of Trent affirmed that the deposit of faith was preserved in the Scriptures and in unwritten (not in the Bible) traditions and that the Catholic Church accepts these two with equal reverence. The council studiously avoided the statement that they meant these "two" as two sources of the deposit, but most Catholic theologians after the council understood the statement as meaning two sources. Protestants thought it meant the Roman Catholic Church had written a second Bible.

The question of the two sources of faith

Only in contemporary Catholic theology has the question been raised again, and a number of theologians believe that Scripture and tradition must be viewed as one source. They are, however, faced with the problem of nonbiblical articles of faith. To this problem several remarks are pertinent. The first is that no Protestant church preaches "pure" gospel; they have all developed dogmatic traditions, concerning which they have differed vigorously. It is true, on the other hand, that they do not treat these dogmatic traditions "with equal devotion and reverence" with the Bible. The second is that the early Christian Church through the first eight ecumenical councils (before the Eastern Schism in 1054) arrived at nonbiblical formulas to profess its faith. Protestants respond that this is at least a matter of degree and that the consubstantiality of the Son (*i.e.,* that he is of the same substance as the Father), defined by the Council of Nicaea, is more faithful to the Scriptures than the Assumption of Mary.

Roman Catholics and Protestants should be able to reach some consensus that tradition and Scripture mean the reading of the Bible in the church. Protestants never claimed that a man and his Bible made a self-sufficient Christian church. The New Testament itself demands that the word be proclaimed and heard in a church, and the community is formed on a common understanding of the word proclaimed. This suggests a way to a Christian consensus on the necessity and function of tradition. No church pretends to treat its own history of belief as nonexistent or unimportant. By reading the Scriptures in the light of its own beliefs it is able to address itself to new problems of faith and morals that did not exist in earlier times or to which the church did not attend.

Development of dogma

Catholic theologians of the 19th century dealt with the problem under the heading of development of dogma. To a certain extent the question can be reduced to epistemology (*i.e.,* theory of knowledge): is a new understanding of an ancient truth a "new" truth? The problem does not arise out of faith; Sir Isaac Newton's observations of falling bodies consisted of nothing that people had not seen for thousands of years. Yet the effects of Newton's insights and calculations altered an understanding of the universe and the actions of people within the universe. The problem is important in theology because of the necessity of basing belief on the historical event of the revelation of God in Christ. Unless the link is maintained, the church is teaching philosophy and science, not dogma. Hence, the Roman Catholic theological teaching has tended to say that dogma develops through new understanding, not through new discoveries.

THE TEACHING AUTHORITY OF THE CHURCH (THE MAGISTERIUM)

The concept of teaching authority. The Roman Catholic Church claims for itself a teaching authority that is unparalleled in the Christian community. The Reformation was primarily a rebellion against the teaching authority, and the Reformers did not claim for their own churches the authority they rejected in the Roman Church.

To teach with authority means that the teacher is able to impose his doctrine upon the listener under a religious and moral obligation. This moral obligation does not flow from the nature of teaching, which of itself imposes no obligation upon the learner; the learner is morally obliged only to assent to manifest truth. Instead it flows from the understanding that the Roman Church derives its teaching authority from the commission given by Jesus to the Apostles as contained in the New Testament ("He who hears you hears me"). But whereas the response of the hearers of the Apostles was faith, the response of the Roman Catholic is expected to go beyond faith. The Apostles were presumed to speak to those who did not yet believe, whereas the Roman Catholic Church imposes its teaching authority only upon its members. The definition of the teaching authority must show that these modifications do not exceed the limits of legitimate doctrinal development.

Organs of teaching authority. The teaching authority is not vested in the whole church but in certain well-defined organs. These organs are the hierarchy—the pope and the bishops. The Roman Catholic Church traditionally has divided the church into "the teaching church" and "the listening church." Clergy below the hierarchical level are included in "the listening church," even though they are the assistants of the bishops in the teaching office. The hierarchy alone teaches what the Roman Catholic Church calls "authentic" doctrine. There is an unresolved antithesis between this idea and the traditional belief that "the consent of the faithful" is a source of authentic doctrine; the conventional resolution that defines the consent as formed under the direction of the pastors of the faithful resolves the problem by depriving the consent of the faithful of any meaning.

The hierarchy

The Roman pontiff is vested with the entire teaching authority of the Roman Catholic Church; this was solemnly declared in the first Vatican Council. This means that he is the only spokesman for the entire Roman Church; the papacy carries in itself the power to act as supreme pastor. It is expected that he will assure himself that he expresses the existing consensus of the church, but in fact the documents of the first Vatican Council are open to the understanding that the pope may form the consensus by his utterance. The second Vatican Council clarified this ambiguity in the idea of the spokesman of the church by its emphasis on the collegial character of the primacy of the pope. The pope, however, does not always speak as the supreme pastor and head of the Roman Church, and he is expected to make this clear in his utterance.

The bishops are authentic teachers within their dioceses. Thus, the same implicit conflict exists in regard to teaching as was noted in connection with governing. The conflict is resolved by collegiality; that the authentic teacher teaches orthodox doctrine is recognized by comparing his doctrine with that of his episcopal colleagues. In this way doctrinal disputes were resolved in the pre-Constantinian church, and a regional council was called if necessary. Since the Reformation the Roman see has never admitted publicly that a bishop has fallen into doctrinal error; the united front of authentic doctrine is preserved, and the matter is dealt with by subtle means. What is taught by all the bishops is authentic doctrine; it is understood that they teach in communion with the Roman pontiff, and a conflict of doctrine on this level is simply not regarded as a possibility. This consensus of the bishops is known as "the ordinary teaching." "The extraordinary teaching" signifies the solemn declaration of an ecumenical council, which is the assembly of the bishops, or the most solemn type of papal declaration, known as a definition of doctrine ex cathedra ("from the throne"), a term that signifies that the declaration exhibits the marks of the teaching of the supreme pastor addressed to the universal church.

Object and response. The object of authentic teaching is defined as "faith and morals." Faith means revealed truth. Morals theoretically means revealed moral principles, but it has long been understood as moral judgment in any area of human conduct; thus, the Roman Catholic Church not only prohibits contraception for its members, but by declaring it contrary to "the natural law" the church declares contraception to be universally wrong. Thus, morals includes the declaration and interpretation of the natural law. The limits of faith and morals have never been defined by the Roman Catholic Church, nor can one take the exercise of the teaching authority as a reliable guide. Thus the teaching authority condemned the heliocentric theory of Galileo as contrary to the Bible because it has always understood that revealed truth involves propositions that are not themselves revealed but that must be affirmed or denied, at least in the present context of knowledge, because of revealed doctrine.

Dogma is the name given to a proposition that is proclaimed with all possible solemnity either by the Roman pontiff or by an ecumenical council. A dogma is a revealed truth that the Roman Catholic Church solemnly declares to be true and to be revealed; it is most properly the object of faith.

Infallibility

The first Vatican Council declared that the pope, when he teaches solemnly and in the area of faith and morals as the supreme universal pastor, teaches infallibly with that infallibility that the church has. The infallibility of the church has never been defined, and its extent is understood by theologians in the sense of pontifical infallibility as limited to faith and morals. These terms are ambiguous, as noted above. Infallibility is actually hedged in with many reservations; nevertheless, pontifical documents often have an aggressive tone that may mislead the incautious reader. The real problem is how a teaching authority that can and does make errors in doctrinal teaching can be called infallible, even with numerous and serious reservations. In the early 1970s some Catholic theologians (*e.g.,* Hans Küng) suggested that the church should be understood as indefectible (*i.e.,* not able to fail or be totally led astray) rather than infallible.

The proper response of the Roman Catholic to authoritative teaching that is "ordinary" and does not clearly deal with "faith or morals" is religious assent. This is extremely difficult to define; it admits dissent under poorly defined conditions. But the theory of religious assent does in fact permit the considerable dissent from the authoritative teaching of Paul VI in 1968 against contraception. Religious assent is particularly relevant to the pontifical document called the encyclical, a type of document that first appeared in the 18th century and became the normal mode of pontifical communication in the 19th century. The encyclical letter is a channel of ordinary teaching, not solemn and definitive and somewhat provisional by definition. Religious assent may be withheld, in popular language, by anyone who in good conscience thinks he knows better. The traditional discipline has made Roman Catholics slow to say this; in modern times they say it more quickly. At the same time, the documents of the second Vatican Council indicate that the authoritative teaching body will be slower to assert itself in the future.

MAJOR DOGMAS AND DOCTRINES

The Roman Catholic Church in its formula of baptism still asks candidates to recite the Apostles' Creed as a sign that they believe what they must believe. The early Church Fathers made the creed the basis of the baptismal homilies given to catechumens, those preparing for the rite. The homilies, like modern Roman Catholic doctrine, went considerably beyond the bare articles of the creed.

Roman Catholic faith incorporates into its structure the books of the Old Testament. From these books it derives its belief in original sin, conceived as a hereditary and universal moral defect that makes human beings incapable of achieving their destiny and even of achieving basic human decency. The importance of this doctrine lies in its explanation of the human condition as caused by the failure of man and not by the failure of God (nor, in modern Roman Catholic theology, by diabolical influence). Man can be delivered from the human condition only by a saving act of God. This act is accomplished by God in the death and Resurrection of Jesus. In Jesus, God is revealed as the Father who sends the Son on his saving mission, and through the Son the Spirit comes to dwell in the redeemed. Thus the Trinity of Persons is revealed, and the destiny of man is to share the divine life of the three Persons. The saving act of Jesus introduces into the world grace, a theological idea that has been much and hotly disputed. Grace signifies in Roman Catholic belief both the love of God and the effect produced in man by this love. The response of believers to the presence of grace is the three theological virtues of faith, hope, and charity; these enable them to live the Christian life. Human beings are introduced to grace and initiated into the church by baptism, which must be preceded by repentance and faith. The life of grace is sustained in the church by the sacraments.

Grace

The life of grace reaches its fulfillment in eschatology; in this area of belief about the end of the world and "the last things," there is some uncertainty in modern theology. Most theologians recognize the mythological character of most of the imagery of heaven, hell, and purgatory. The peculiarly Roman Catholic belief in purgatory was an effort to state that most men at death are neither good enough for heaven nor bad enough for hell. The theology of the last things is still unable to cope with the implications of this statement. Belief in a resurrection to eternal life has never been easy, and modern times have produced more difficulties than solutions. Christianity, in fact, shows oscillation between a transcendental direction and an immanent direction; in modern times the emphasis is on immanence—that is, on the meaning of religion in the world. The second Vatican Council reflected this in its statements on the "secular" and the response of the church to the secular.

This summary can state no more than the basic elements of the Christian fact. The complex Roman Catholic dogmatic structure has been mentioned several times, and probably no two statements of "major dogmas and doctrines" would be the same.

THE LITURGY

Cultic worship is so universal in religion that some historians of religion define religion as cult. Cultic worship is social, and this means more than a group worshiping the same deity in the same place at the same time. Cult is structured with a division of sacred personnel (priests) who lead and perform the cultic ceremonies for the people, who are in a more distant relation with the deity. The sacred personnel are designated by the choice and acceptance both of the deity and of the worshiping group. The words and actions of the cultic performance are divided into roles assigned to the leaders and to the worshipers. It is the tendency of cultic worship to replace spontaneity, which it once had, with set and even rigid forms of words and acts. These are preserved by tradition, and they generally have a sacredness that is based on the belief that the directions for cultic worship came ultimately from the deity.

The eucharistic assembly or mass. Roman Catholic liturgy has its roots in Judaism and the New Testament. The central act of liturgy from earliest times was the eucharistic assembly, the commemorative celebration of the Last Supper of Jesus. This was set in a structure of liturgical prayer. The first six centuries of the Christian Church saw the development of a rich variety of liturgical systems, many of which have survived in the Oriental churches. In the West the Latin liturgy appeared fully developed in Rome in the 6th and 7th centuries. From the 8th century the Roman liturgy was adopted throughout western Europe. In this same period, however, liturgy developed in Frankish territories; and the Roman rite that emerged as dominant in the 10th century was a Roman-Frankish creation. The Roman rite was reformed by the Council of Trent by the removal of some corruptions and the imposition of uniformity; after Trent the Roman see was the supreme authority over liturgical practice in the entire Roman Catholic Church.

Central act of liturgy

By the 11th century Roman liturgy had acquired the

classic form that it retained up to the second Vatican Council. The fullness of the liturgy could be witnessed only in some cathedrals, collegiate churches, and monastic churches. The full liturgy included the daily celebration of the solemn high mass and recitation of the divine office in choir. The solemn high mass was performed by at least three major officers (celebrant, deacon, and subdeacon), assisted by many acolytes and ministers. Except during the penitential seasons of Advent and Lent, the altar was decorated, and numerous candles (in the Middle Ages for light rather than ornamentation) and incense were employed. The singing and chanting were accompanied by the organ and in modern times even by orchestral music; Mozart once complained that the Archbishop of Salzburg compelled him to compose a mass without the resources of a full symphonic orchestra.

The divine office. The divine office was a legacy to the clergy from the monks. From the beginnings monks assembled several times daily for prayer in common. This developed into set common prayer at stated times each day (Matins, midnight; Lauds, first daylight; Prime, sunrise; Terce, mid-morning; Sext, noon; None, mid-afternoon; Vespers, sunset; Compline, before retiring). The divine office consisted basically of the chanting of the Psalms (in a weekly cycle), the recital of prayers, and the reading of the Scriptures (to which were later added selections from the writings of the Church Fathers, probably instead of a homily given by one of those present). Together with the mass the office has been the only "official" prayer of the Roman Catholic Church; all other prayer forms are "private," even if several hundred people recite them together. For this reason clerics in major orders for centuries since the Middle Ages have been obliged to recite the divine office, or "breviary," privately if they are not bound to attend the office in choir. It was long recognized that there is an inconsistency in the private silent reading of a prayer structure that is intended for choral chanting, and the second Vatican Council recommended a reform, after which time many priests abandoned the breviary.

Official prayer of the church

The cycle and the language of the liturgy. The liturgy has long been arranged in an annual cycle that is a re-enactment of the saving events of the life, death, Resurrection, and glorification of Jesus Christ. Even many Catholics do not realize that the cycle has an eschatological outlook; the events are reenacted as an assurance that the saving act will reach its eschatological fullness, and the liturgy is an expression and a support of the Christian hope. The cult of the saints is an intrusion into the liturgical cycle, and it has been much reduced in the contemporary liturgical reforms.

Latin did not become the language of the Roman rite until the 6th century; the language of imperial Rome was Greek. As a sacred language Latin really has no parallel. Jews have always made a genuine effort to learn some Hebrew, and other sacred languages are archaic forms of the vernacular; the English of the Authorized Version of the Bible became the language of prayer in many Protestant churches. The effect of Latin was to make the liturgy the preserve of the clergy, and the laity became purely passive. This was countered by the efforts to use sound and spectacle in the performance of the solemn liturgy. The Canon of the mass, the central eucharistic formula, for centuries was recited by the celebrant inaudibly; this was a kind of verbal "sanctuary" that the laity were not even supposed to hear. The abandonment of Latin as a result of the second Vatican Council excited deep antagonisms; one sees in the Latin liturgy an image, cherished by many, of the timeless and changeless Roman Catholic Church. Yet the restoration of the vernacular should restore to the liturgy two functions that it had in the early centuries: to instruct converts and to confirm members in their faith.

THE SACRAMENTS

The sacraments in general. In Roman Catholic theology a sacrament is an outward sign, instituted by Jesus Christ, that is productive of inner grace. The number of sacraments is seven (defined by the Council of Trent against the Reformers, who reduced the number). The number seven does not appear in Roman Catholic teaching before the 11th century, and it is an example of truth for which the Roman Catholic Church relies on its own tradition.

The sacrament in modern theology is frequently described as an encounter with mystery, the mystery being the saving act of God in Christ. Theological studies have been directed to the exploration of the idea of sign and significance. The traditional Roman Catholic statement of the effectiveness of the sacraments (defined by the Council of Trent) is described by the untranslatable *ex opere operato,* which is best explained briefly by saying that the faith and virtue of the minister neither add to the sacrament by their presence nor detract from it by their absence. The minister is merely the agent of the church, and the effectiveness of the sacrament is based on the saving act of God in Christ, which is signified by the rite and applied to the recipient of the sacrament.

The sacrament as an encounter with mystery

Protestant theologians formerly charged the Roman Catholic Church with a belief in magic; this controversial angle has generally been abandoned, but the theological explanation of the sign that effects by signifying is still difficult. Roman Catholic theologians remark that the mystery of God's saving act is not capable of complete rational explanation. There are analogies, however, in common experience, and there is no society that does not employ effective signs. These signs are not merely for display. The inauguration of the president of the United States makes the man president; the sign is effective because it signifies the reality of the election that this individual won. The sign of the coronation of a monarch is equally effective, but it is more difficult to define the reality signified. Such effective symbols are a part of human society.

The Roman Catholic Church adheres strictly to the external sign. Traditionally the church attributes the institution of the sign to Jesus Christ (although this has been the subject of discussion among modern theologians), and this removes the right of anyone to tamper with it. The Roman Catholic Church believes that, if God gave a sign, the alteration of the sign so that the significance is lost might render the sign ineffective. Hence, the use of the proper material and the retention of the traditional formula are treated as sacred. The Roman Catholic Church maintains its exclusive competence to supervise matter and form "in detail," a competence not precisely defined. Since Thomas Aquinas the material used is called matter, and the words are called form; the terms are borrowed from the Aristotelian theory of the constitution of matter. The material becomes sacred and salutary only by its conjunction with the proper words. The effect produced has for centuries been called grace, but it is difficult to assert a single effect and still explain why there are seven symbols.

The term sacramental is used to designate verbal formulas (such as blessings) or objects (such as holy water or medals) to which a religious significance has been attached. These are symbols of personal prayer and dedication, and their effectiveness is measured by the particular dispositions of the person who uses them. Although superstition has arisen in connection with sacramentals, the Roman Catholic with elementary instruction knows the difference between them.

Sacramentals

Baptism. Baptism is the sacrament of regeneration and initiation into the church. According to a theme of St. Paul, probably influenced by Jewish belief in the circumcision of adult proselytes, baptism is death to a former life and the emergence of a new person, signified by the conferring of a new name; it is the total annulment of the sins of one's past and the emergence of a totally innocent person. One becomes a member of the church and is incorporated into the body of Christ, thus becoming empowered to lead the life of Christ. Nothing but pure natural water may be used, and baptism must be conferred in the name of the Father, the Son, and the Holy Spirit. Baptism is normally conferred by a priest, but the Roman Catholic Church accepts the baptism conferred by anyone having the use of reason "with the intention of doing what the church does." As the sacrament of rebirth it cannot be repeated. The Roman Catholic Church baptizes conditionally in case of doubt of the fact of baptism or the use of the proper rite.

Two points of controversy still exist in modern times.

Points of controversy

One is baptism by pouring rather than immersion, even though immersion was probably the biblical and early Christian rite. The change was almost certainly the result of the spread of Christianity into Europe north of the Alps and the occurrence of the baptismal feasts, Easter and Pentecost, often in early spring. The Roman Catholic Church simply asserts that the symbolism of the bath is preserved by a ritual infusion of water.

The second is the baptism of infants. There is no certain evidence of this earlier than the 3rd century, and the ancient baptismal liturgies are all intended for adults. The liturgy and the instructions clearly understand the acceptance of baptism as an independent adult decision; without this decision the sacrament cannot be received. The Roman Catholic Church accepts this principle by introducing adults (sponsors, godparents), who make the decision for the infant at the commission of the parents. In Roman law as in modern law, adults are empowered to make decisions for minors. It is expected that the children will accept the decision made for them and will thus supply the adult decision that was presumed.

Until the recent liturgical renewal baptism did not have the religious and ceremonial importance that it had in the early church; the ceremonies were intended to make the adult aware that he had made the most important decision of his life, and the whole church witnessed the ceremony, performed only twice a year on a group of catechumens. Doubtless the baptism of infants contributed to this loss of ceremonialism and to a corresponding lower esteem of baptism.

Confirmation. Confirmation since the 11th century has been conferred by the bishop through the anointing with oil and the imposition of hands; the words are a declaration that the Holy Spirit is conferred. This is an echo of the accounts in the Acts of the Apostles (chapters 8 and 19) in which a distinction is made between baptism and the conferring of the Spirit. In Acts, however, the reception of the Spirit meant the reception and the manifestation of charismatic gifts (*e.g.*, prophecy, speaking with tongues, ecstasy); something else is now meant. Confirmation is normally conferred at or near the beginning of adolescence. The modern liturgical renewal has empowered pastors of parishes to confer confirmation.

Neglect of the theology of confirmation has left some ambiguities. The Oriental churches confer it on infants as a part of the initiation rites of baptism. The postponement of confirmation has led many Roman Catholic theologians to interpret it as a rite of passage from childhood, like the Jewish Bar Mitzvah ceremony; such rites of passage are common in tribal cultures. Early Christian baptism, however, was conferred on adults; thus the catechumenate was the period of "immaturity." It seems that there should be a return to the theology of the Spirit and a consideration of confirmation as the sacrament that empowers the Christian to take an active part in the church. The traditional Roman Catholic view of the laity as passive has contributed to the neglect of the theology of confirmation; it left no room for a charismatic laity.

The Eucharist. The Eucharist (the Lord's Supper, Holy Communion) is with baptism one of the two sacraments most clearly found in the New Testament; most Christian churches have it in some form. The Roman Catholic Church distinguishes the Eucharist as sacrifice (mass) and sacrament (communion).

The formula of institution of the Eucharist and the command to repeat it are found in the three Synoptic Gospels (Matthew, Mark, Luke) and in Paul. Originally the Eucharist was a repetition of the common meal of the local group of disciples with the addition of the bread and the cup symbolizing the presence of Jesus. Even in the 2nd century the meal became vestigial and was finally abandoned. The Eucharist was originally celebrated every Sunday; by the 4th century it was celebrated daily. The eucharistic formula was set in a framework of biblical readings, psalms, hymns, and prayers that depended in form somewhat on the synagogue service. This remained one basis of the various liturgies that arose, including the Roman rite.

The sacrificial character of the Eucharist was determined by its relation to the death of Jesus. The Eucharist is not seen as sacrificial everywhere in the New Testament, but the theme is so clearly elaborated in the Letter to the Hebrews that it is universally accepted as Christian belief. The Protestant churches denied the sacrificial character of the Eucharist and rejected the mass. Roman Catholic theology has never reached a universally accepted theory explaining the connection between the death of Jesus and the mass, but it has firmly insisted that the mass repeats the rite that Jesus told his disciples to repeat and that the rite is an effective symbolic commemoration of his death. The mass is the only act of worship that the Roman Catholic Church imposes upon its members. Historically, the Roman Church has attached great importance to the mass, conceding almost anything to secure its celebration.

Roman Catholicism believes in the Real Presence, and this has dominated Catholic-Protestant controversies about Holy Communion. Protestant belief can generally be called dynamic as contrasted with Catholic realism. The celebrated term transubstantiation is defined as the change of the substance of bread and wine into the substance of the body and blood of Jesus Christ. Protestants believe that Jesus is experienced as present. The Roman Catholic theory is difficult to explain in terms other than those of antiquated Aristotelian physics, and recent theories,

Bildarchiv Foto Marburg/Art Resource, New York City



"The Last Supper," relief sculpture on the rood loft of the Naumburg cathedral; after 1250.

not yet successful, have attempted to explore sacramental symbolism in other ways. The realism of belief in the presence is associated with the Roman Catholic practice of distributing only the bread to the laity, a serious modification in the sacramental sign. Not yet universally restored, Holy Communion under both species has become much more common since the second Vatican Council.

Sacrament of unity

Neither in Roman Catholic nor in Protestant eucharistic practice does the sacrament retain much of the symbolism of Christian unity, which it clearly has in the New Testament. Originally the symbolism was that of a community meal, an accepted social symbol of community throughout the whole of human culture. Roman Catholic efforts to restore this have included the use of the vernacular and the active participation of the laity. Furthermore, the ancient rite of concelebration—*i.e.,* several priests or bishops jointly celebrating a single eucharistic liturgy—was restored by the second Vatican Council as a means of symbolizing unity; and the practice of celebrating the Eucharist in an informal setting—*i.e.,* in private homes or classrooms—was instituted in some places as a way of drawing the laity more intimately into the rite. But a great obstacle to the symbolism of unity remains the liturgical isolation of the celebrant and the silence that suited the atmosphere of mystery and the presence of God.

Church law obliges the Roman Catholic to receive Holy Communion once a year (during the Lent-Easter season). Practice of frequency has varied over the centuries; the present law reflects the infrequency that was common in the Middle Ages. The symbolism of the sacrament as nutrition becomes rather feeble with such infrequency; it was rationalized both by the theology of the power of the sacrament and by considerations of the general unworthiness of Christians to receive it.

Penance. The name of the fourth sacrament, penance, reflects the earliest discipline of the penitential rite. Those who sinned seriously were excluded from Holy Communion until they showed repentance by undergoing a period of public penance that included such practices as fasting, public humiliation, the wearing of sackcloth, and other austerities. At the end of the period they were publicly reconciled to the church. There were some sins, called capital (murder, adultery, apostasy), for which certain local churches at certain times did not perform the rite; this did not mean that God did not forgive but that good standing in the church was permanently lost. Elsewhere it was believed that the rite of penance could be performed only once; relapsed sinners lost good standing permanently. Rigorist sects that denied the power to forgive certain sins were regarded as heretical. The penitential rite did not endure beyond the early Middle Ages, and there can be no doubt that it was too rigorous for most Christians. It may also be noticed that the penitential discipline did not reflect the forgiveness of Jesus in the Gospels with all fidelity.

Confessing of sins

It is impossible to assign an exact date for "auricular confession"—the confessing of faults by an individual penitent to a priest—but it must have arisen in the early Middle Ages with the disappearance of the penitential system. This is the penitential rite that has endured into modern times. It was rejected by most of the Reformers on the ground that God alone can forgive sins. The Roman Catholic Church claims that the absolution of the priest is an act of forgiveness; to receive it the penitent must confess all serious (mortal) sins and manifest genuine "contrition," sorrow for sins, and a reasonably firm purpose of amendment. No quality or quantity of sin is too great for sacramental absolution. Roman Catholic theologians have not arrived at an explanation of the process of absolution. They do not admit that absolution is merely a recognition by the priest of dispositions on the part of the penitent that merit forgiveness nor that it is merely a process whereby the penitent is reconciled with the church. There seems to be an unspoken belief that it is a rare person who is really sorry for his sins and that the sacrament is a manifestation of the graciousness of God to human weakness.

Indulgences, which caused such a stir at the beginning of the Reformation, are neither instant forgiveness to the unrepentant nor licenses of sin to the habitual sinner. They are declarations that the church accepts certain prayers and good works, listed in an official publication, as the equivalent of the rigorous penances of the ancient discipline.

The anointing of the sick. This sacrament was long known in English as "extreme unction," literally rendered from its Latin title, *unctio extrema.* This non-English designation concealed the meaning of the Latin, "last anointing." It is conferred by anointing the sense organs (eyes, ears, nostrils, lips, hands, and formerly the feet and the loins) with blessed oil and the pronunciation of a formula. It may be conferred only on those who are seriously ill; seriousness is measured by the danger of death, but a danger, however certain, from external causes (such as the execution of the death sentence) does not render one apt for the sacrament. It may be administered only once during the same illness; recovery renders one apt again. Its effects are described as strengthening both of soul and body; it is an ancient rite that continues Jesus' ministry of healing. The sacrament is directed against "the remains of sin," an ill-defined phrase; but it was long ago recognized that illness saps one's spiritual resources as well as one's physical strength, and one is not able to meet the crisis of mortal danger with all of one's powers. In popular belief anointing is most valuable as a complement to confession or, in case of unconsciousness, as a substitute for it.

The anointing is not the sacrament of the dying; it is the sacrament of the sick. The New Testament passage (James 5:14–15) to which the Roman Catholic Church appeals for this rite does not envisage a person beyond recovery. Postponement until the patient is critically ill in modern medical terms means that the sacrament is often administered to an unconscious or heavily sedated patient. Under such circumstances the rite can no longer be effective as a sacrament of the sick, and to the uninformed a magical rite of forgiveness is suggested.

Marriage. The inclusion of marriage among the sacraments gives the Roman Catholic Church jurisdiction over an institution that is of concern to the state and to non-Catholic persons and groups within society. The Roman Church claims complete jurisdiction over the marriages of its members, even though it is unable to urge this jurisdiction in modern secular states. The sacrament in Roman Catholic teaching is administered by the spouses through the exchange of consent; the priest, whose presence is required, is an authorized official witness; in addition, the church requires two other witnesses. Marriage is safeguarded by a number of impediments that render the marriage null and void whether they are known or not, and the freedom of the spouses must be assured. This means that the Roman Catholic Church demands an unusually rigorous examination before the marriage, and this in turn means that it is practically impossible to marry on impulse in the Catholic Church. All of this is for the purpose of assuring that the marriage so contracted will not be declared null in the future because of some defect.

The claim of jurisdiction over marriages

The rigid Roman Catholic rejection of divorce has been a major point of hostility in the modern world. Absolute indissolubility is declared only of the marriage of two baptized persons (Protestants as well as Catholics). The same indissolubility is not declared of marriages of the unbaptized, but the Roman Church recognizes no religious or civil authority except itself that is empowered to dissolve such marriages; this claim is extremely limited and is not used unless a Roman Catholic is involved. Because of its rigorous conditions for contracting marriage, the Roman Catholic Church finds grounds for nullity that do not exist in civil law, and it is willing to make a more searching examination. Declarations of nullity, however, should not be confused with divorce nor be thought a substitute for divorce. Some Roman theologians have suggested that Roman Catholic rigour is based on a misunderstanding of the Gospel texts that reject divorce; but a position maintained for centuries is not easily modified.

The onerous conditions that Roman Catholicism formerly imposed upon non-Catholic partners in mixed marriages have been notably relaxed since the second Vatican Council, particularly as regards written promises that the children would receive religious education in the

Roman Catholic faith. The former coldness of the Roman Church toward such marriages is also relaxed; they may be celebrated in church during the mass, and a Protestant minister or a Jewish rabbi may share the witness function with the priest.

Power to administer the sacraments

Holy orders. This sacrament confers upon candidates the power over the sacred, which means the power to administer the sacraments. The Latin Church had long recognized four minor orders (porter, lector, exorcist, acolyte) and four major orders (subdeacon, deacon, priest, bishop). The minor orders represented church services rendered by persons not ordained. In 1972 Pope Paul VI issued the apostolic letter *Ministeria quaedam,* which abolished the major order of subdeacon and all minor orders and which created the lay liturgical ministries of lector and acolyte. Only the major orders are held to be sacramental, but they are regarded as one sacrament within which a tripartite hierarchy of sacramental effects is administered separately. Ordination is conferred only by the bishop; the rite includes the imposition of hands, anointing, and the delivery of the symbols of the order. The power of the sacred peculiar to the bishop is shown only in the sacraments of confirmation and orders. Ordination can neither be repeated nor annulled. Priests who are suspended from priestly powers or laicized (permanently authorized to live as a layman) retain their sacred power but are forbidden to exercise it except in emergency. The priest is always ordained to a "title," meaning that he is accepted in some ecclesiastical jurisdiction. Lectors and acolytes are instituted by a bishop or by the major superior of a clerical religious institute. Following a calling of the candidates, instruction, and prayer, lectors are presented a Bible and acolytes a vessel with bread or wine.

Other theological developments following the second Vatican Council concerned the ordination of women, against which no solid theological objection has been shown; the restoration of the permanent diaconate (with the powers to baptize, preach, and administer the Eucharist), to which both married and single men are admitted; and the idea of ordination for a fixed period of service. Except for the diaconate, these are radical suggestions in Roman Catholicism.

PARALITURGICAL DEVOTIONS

In the Roman Catholic Church, liturgy in the proper sense is the liturgy of the mass, the divine office, and the sacraments. The Latin language, the clerical character of the liturgy, and the search for novelty for hundreds of years have combined to produce forms of worship that are paraliturgical—by which is meant that they lie outside the liturgy and in some cases in opposition to it. These acts are also known as devotions or devotional practices, by which is meant that they are accepted voluntarily and not from obligation.

Eucharistic devotions. A number of eucharistic devotional practices arose in the Middle Ages, when Catholics rarely received the Eucharist more than once a year. These were cultic forms that were directed to the Real Presence of Jesus in the Eucharist rather than to sacrifice and Holy Communion. Such were Benediction of the Blessed Sacrament and "exposition." Benediction was a blessing conferred by a priest holding a consecrated Host in a vessel of display called the monstrance; the priest's hands were covered to signify that it was the blessing of Jesus and not his own. This blessing was accompanied by hymns and the use of the organ and incense. Exposition was the public and solemn display of the eucharistic bread, again with the accompaniment of hymns, the organ, incense, and processions. The reservation of the Eucharist in churches was a way in which Catholics could address themselves in personal prayer to Jesus really present. These have often functioned as substitutes for mass and Holy Communion, and since the modern renewal of liturgy they occur much less frequently.

Benediction and "exposition"

Cult of the saints. Other devotions revolve about the cult of the saints, a practice repudiated by the Reformers as a denial of the total mediation of Christ. This objection oversimplified Catholic practice, but the devotions did sometimes approach superstition. Catholic theologians distinguish (by Greek technical terms) the worship paid to God (latria, "adoration") from the veneration addressed to Mary (hyperdulia, "super-service") and the saints (dulia, "service"). Protestants do not disagree with the principle of admitting the saints as examples of genuine Christianity, but they reject the intercession of the saints as utterly superfluous and ineffective. The Roman Catholic understanding of the intercession of the saints is an extension of the belief in the communion of saints. Although such veneration does tend to multiply mediators, it has often fostered a simple and not unpleasing familiarity with the world of the supernatural. The excesses of the cult of Mary have stirred up controversy, and the tendency to superstition and the deification of Mary have sometimes been painfully present. Mary represents the feminine principle in Roman Catholicism; often in other religions this principle has been personified as a goddess. Mary is given the feminine traits of sympathy and tenderness that are not improper to the deity but are somewhat improper to the father figure and the king figure. The multitude of apparitions of Mary (*e.g.,* at Lourdes, Fr., and Fatima, Port.) come from the need of a local and national symbol of presence, which enables the Roman Catholics of a nation or region to identify with Mary. Because Mary as a historical person is almost totally unknown, Catholics have been able to find in her all the traits of the ideal person that they needed to find.

Roman Catholicism has always insisted on its right to official supervision of devotional cults, and only approved forms of devotions may be used in the churches or under clerical auspices. Approval does not imply the historical reality of the vision or apparition involved; no Roman Catholic is obliged to believe that Mary appeared to anyone at Lourdes or Fatima, that the rosary (prayer beads) was delivered by a private revelation, or that Jesus manifested himself as the Sacred Heart. Nor is any Catholic obliged to practice any of these devotions. Generally, they serve the purpose of emphasizing some element of Christian faith that is obscured in the preaching and the liturgy at a particular time and place. Devotion to the Sacred Heart, for example, turned the attention of Catholics to the humanity of Jesus and to Christian love in the somewhat arid spirituality of the 17th and 18th centuries. It may be urged that more authentic biblical proclamation would have brought out these things; Roman Catholicism has often manifested itself through devotions when authentic biblical preaching was not available. In approving devotions the Roman Catholic Church simply declares that they are not in conflict with Roman Catholic faith and morals. It does not deny that they may be entirely products of the imagination.

Direct experience of God

Mysticism. The search for God through mysticism has never been received cordially by the official Roman Catholic Church. In general terms, the mystical experience can be described as a direct experience of the reality of the divine. A sufficient number of mystics have been proved fraudulent to justify caution but not to justify a blanket antecedent disapproval. Every saint who has been recognized as a mystic had some trouble with church authority. Indeed, one may see in the mystical experience of God something that the official church can neither furnish nor control. In addition, mystics have often had a prophetic character that expressed itself in criticism of abuses in the official church. Whatever the explanation, mystical phenomena have become extremely rare in the modern Roman Catholic Church.

THE ROLE OF THE CHURCH IN SOCIETY

Missions. From its beginnings Christianity alone among the great religions has regarded itself as a true world religion that appeals to all men without distinction of race, nation, or culture. Roman Catholicism believes that it has preserved this missionary thrust more faithfully than any of the non-Roman churches. From the 4th to the 10th century the Roman Church devoted itself to the evangelization of the barbarians. The barbarians wished to become "Roman," and they accepted the church as a component of Roman civilization. The spread of Islām was met with crusades and not with missionaries, and

the Roman Catholic Church has never mounted more than a feeble missionary effort toward Muslims. Thus, the missionary movement languished from the 10th to the 16th century; but the ages of the expansion of Europe, in which the Catholic countries were the early leaders, spread Roman Catholicism to the Americas, Asia, Oceania, and Africa.

Centralized control by the Roman see

This missionary effort differed from both the New Testament missions and the missions to the European barbarians in its very close, centralized control by the Roman see. Missionary churches have begun to achieve that independence proper to the diocesan structure only in the 20th century. It has been difficult for the Roman Catholic missions to divorce themselves from colonialism, and many missionaries did not want the divorce. Again until recent times most of the clergy and all the hierarchy in mission countries were European or American, as were the heads of educational and benevolent operations. Even the peoples of the mission countries, including their clergy and religious personnel, generally wished to give their church a European identity rather than an Asian or African identity. The Roman see, which had suppressed efforts to admit Chinese rites in the 18th century, was unsympathetic to what appeared to be "non-Roman" practices. The second Vatican Council officially ended the colonial phase of missions; in practice, however, the end will take longer. Where possible—meaning where the personnel are available—the operation of the mission churches has been given to native hierarchy and clergy.

Education. Between the barbarian invasions and the Protestant Reformation, education in Europe, except for the Arabic and Jewish centres of learning, was conducted by Roman Catholicism. Learning during the early Middle Ages was preserved by the monasteries; and, although the monks did little more than copy the manuscripts of Greek and Latin pagan writers and of the Church Fathers, they educated the few people who had any learning. The foundation of the European universities after 1200 was also the work of Roman Catholicism; these institutions were stimulated by the learning of Arabic scholars, through whom Europeans became acquainted with the philosophy of Aristotle and produced the learning of Scholastic philosophy and theology. The cultivation of literature and the arts in the 15th century flourished under the patronage of the papacy and Catholic princes and prelates.

The birth of modern science was coincidental with the Reformation and the age of the expansion of Europe. The Roman Catholic response to the new science, accompanied by new philosophical systems, was hostile; and the world of European learning after 1600 was dissociated from the Roman Catholic Church, which patronized only defensive learning. At the same time, Roman Catholic initiatives in educating the poor were gaining momentum. The invention of printing had diffused education far beyond earlier possibilities, and the churches were all interested in reaching the minds of the young. This interest was matched after the French Revolution by the modern states, which in the 19th century moved toward the exclusion of church influence from education. But the Roman Catholic Church, through its religious communities, was a pioneer in the elementary education of the children of the poor.

Critical problems for Catholic schools

In the 20th century the Roman Catholic educational endeavour in many European and American countries, particularly in the United States, had become a vast enterprise. In the second half of the 20th century, however, mounting costs and diminished religious personnel created critical problems for Catholic schools, and even their survival was at stake in many regions. The problems were not lightened by the realization that Roman Catholic education, even where it was strongest, reached only a minority of Catholic students; and the Roman Church had to face its established reputation as an adversary of the intellectual freedom that the modern academic world cherishes.

Eleemosynary activities. Institutional benevolence to the poor, the sick, orphans, and other helpless people has been characteristic of the Christian Church from its beginning. It involved organized assistance, supported by the contributions of the entire community and rendered by dedicated persons. The church in this way fulfilled the duty of "the seven corporal works of mercy" mentioned in the Gospel According to Matthew (chapter 25) and carried on the healing mission of Jesus. Protestant churches continued the works of institutional benevolence after their separation from the Roman Church, and the history of Christian benevolence is a noble portion of church history. Institutional assistance to the helpless is a legacy from the church to modern governments.

This work, which would seem to be above criticism, was beset with troubles in the latter half of the 20th century. Costs for these works, like the costs for education, soared beyond the possibilities of individual contributions. Governmental assumption of responsibility for benevolence both rendered the necessity of church works doubtful and narrowed the base of contributions. Church organizations as they existed were not well equipped to deal either with modern urban poverty or with the problem of international poverty.

Church and state relations. The most important modification in Roman Catholic theory and practice of church–state relations was the declaration of the second Vatican Council in which the Roman Catholic Church recognized the modern, secular, pluralistic nation as a valid political society. Union of church and state had been the common pattern since the era of Constantine, and all pontifical declarations of the 19th century rejected separation of church and state as pernicious. This position was steadfastly maintained in spite of the fact that the union of church and state had been accepted by the Protestant countries of Europe; it reflects a long history of domination of the church by the state and of the church's involvement in political power struggles. The second Vatican Council declared that the Roman Catholic Church is not a political agent and will not ask political support for ecclesiastical ends. A significant change in the Roman attitude toward the state is the council's express declaration of freedom of religion.

Recognition of the pluralistic state

Economic views and practice. In the centuries when the Roman Catholic Church was Christendom, there was a place for every member in the church that corresponded to the individual's place in the social structure. In modern times the church hierarchy became identified with the landed aristocracy; this dangerous identification led the revolutionaries of 18th-century France to attempt to destroy the Roman Catholic Church with other components of the old order. The Roman Catholic Church entered the 19th century with a firm official bias against revolutionary movements, and the brief liberalism of Pius IX was ended with his experiences in the Italian revolution of 1848. The Roman Catholic Church was inflexibly opposed to all forms of Socialism, and its opposition to Marxist Communism was implacable. Thus, the Roman Catholic hierarchy was identified with the new capitalist classes of the industrial society. In many European countries this meant that the church lost membership among the working classes. Leo XIII in *Rerum Novarum* (1891; "Of New Things") was the first pope to speak against the abuses of capitalism. Social teaching was further elaborated by Pius XI in *Quadragesimo Anno* (1931; "In the 40th Year"), John XXIII in *Mater et Magistra* (1961; "Mother and Teacher") and *Pacem in Terris* (1963; "Peace on Earth"), and Paul VI in *Populorum Progressio* (1967; "The Progress of Peoples"). Catholic opposition to Socialism has gradually been diminished, although Catholic teaching tends in the direction of the diffusion of capital and not in its nationalization. In some features, however, such as the approval of labour unions, Catholic teaching has reached points that Pius IX would have regarded as Socialism.

In its own practices the Roman Catholic Church has accepted full control over the ownership of property and of productive investments. It does not admit accountability to the laity for its funds, which are managed by the hierarchy; hence the wealth of the Catholic Church has long been a mystery, often attractive to greedy anticlerical governments. Their raids as well as some public disclosures indicate that popular belief exaggerates the wealth of the Catholic Church. Following the second Vatican Council, there was a strong movement in Catholicism for public financial reports.

Stability of the family

The family. Roman Catholic teaching attributes to the family a social and moral centrality that many people think it no longer has. The teaching has grown up around a number of factors, not all of which come from the New Testament. The guiding principle of Roman Catholic teaching is the stability of the family. This principle does not admit divorce. The family is preserved by a strong authority structure in which the father is the head; this reflects not only the Old Testament but also Roman law. The family, moreover, is child-centred; traditional Catholic teaching makes the primary end of marriage the procreation and rearing of children. Only recently have Catholic theologians begun to speak of mutual love as an end "equally primary." The principle of the stability of the family also mirrored a certain tribal view of the family that saw the individual's chief security not in the law and the courts but in kinsmen. Such a view of stability is, however, not highly compatible with the mobility of the modern family and the independence possible to the person who wishes it strongly enough. In the late 20th century the Roman Catholic Church was faced with the problem of preserving for its members the values of mutual love and responsibility associated with the family without imposing on it an antiquated structure. But the major problem was certainly the practice of birth control. The moral arguments for the Catholic position against birth control had suffered general erosion, and many Catholics regarded the declaration of Paul VI in 1968, reiterating the traditional prohibition, as a blind exercise of authority. The church also struggled with the issue of homosexuality among the laity and clergy.

Roman Catholicism following the second Vatican Council

The Roman Catholic Church has been experiencing a renewal that reached its official peak in the second Vatican Council. Renewal has brought benefits, but it has also brought internal disturbances greater than any the church has known since the Protestant Reformation. There has been a clear polarization between liberal and conservative wings of the type that tends to leave no room for moderates. Although such disunity poses a real threat of schism, there have been no group departures except in a few instances. The number of individual departures, however, has been large enough to cause concern. The exact number is unknown, because discontented Catholics in modern times leave quietly.

The Roman Catholic Church has officially abandoned its "one true church" position. It has entered into ecumenical conversations with the Protestant churches that could lead to Christian union; the Catholic Church has expressed a readiness to make doctrinal and disciplinary concessions, but how far these may go is not yet clear. The church has even made gestures of friendliness to Islām and Judaism and does not speak of the great Asian religions as simple paganism. Indeed, efforts at improving relations with other religions were pivotal to the papacy of John Paul II.

The openness of the Catholic Church following the second Vatican council took surprising form in places such as Latin America, where the church supported liberation theology in the 1970s. Catholic leaders also adopted a less confrontational approach to communist governments in the hope of improving the life of the church in those countries. Pope John Paul II, however, supported opposition movements in communist eastern Europe and suppressed liberation theology, but at the same time the church remained involved in international affairs as he traveled the world in support of peace and the church.

Problems within the church

Problems, however, are more in evidence than progress. The most spectacular was the pedophilia scandal that broke in the early 21st century in which numerous priests were accused of molesting children and their superiors were charged with covering up or ignoring the problem. The church also faced the challenge of resolving the long, latent conflict between hierarchy and lower clergy. Priests are resistant to the traditional total obedience in style of life and ministry. This conflict has come to a focus on the issue of clerical celibacy; although there are no sure statistics, it is a reasonable assumption that at least half of the Catholic clergy wish celibacy to become an option. The discontent with life and ministry has led to a large number of losses in the priesthood and in religious communities, some of which face the possibility of extinction. Much of this discontent revolves around ministry as much as around a way of life; many religious workers feel that the conventional ministries are not reaching enough people and are not touching their most urgent needs. What might appear to be a minor point in some places has become major; priests and religious (women religious in particular, who have had more of a problem) no longer wish to wear the identifying garb; they believe that it immediately becomes an obstacle to personal relations.

Roman Catholic liturgy has been profoundly changed. The results have not been altogether satisfactory, but some observers say that the effects of the new liturgy cannot be assessed until a new generation has grown up that knows no other liturgy. On this point minor local schisms have occurred, led by reactionary Catholics wishing to return to the traditional liturgy in Latin. Others find the new liturgy stodgy; but the degree to which liturgy ought to be exciting has never been established.

The place of the laity, like that of the clergy, in church decisions remains uncertain. Bishops, clergy, and laity generally are timid in undertaking a modification in church government for which nothing in their previous church experience has prepared them. They seem hesitant to draw on their experience in government and business, where shared responsibility is the rule rather than the exception. Many Catholics find it difficult to examine the role of their hierarchical officers without also questioning their credibility. Yet the direction of the movements where the problems lie is toward greater responsibility of each member of the Catholic Church—hierarchy, clergy, and laity, each in its own way. (J.L.McK.)

BIBLIOGRAPHY

General. Reference works include ROBERT C. BRODERICK (ed.), *The Catholic Encyclopedia*, rev. and updated ed. (1987); *New Catholic Encyclopedia*, 17 vol. (1967–79, reissued 1981), which treats all phases of Roman Catholicism and includes a volume on change in the church; *Sacramentum Mundi: An Encyclopedia of Theology*, ed. by KARL RAHNER *et al.*, 6 vol. (1968–70), which deals with Catholic doctrine and theological thought; and F.L. CROSS and E.A. LIVINGSTONE (eds.), *The Oxford Dictionary of the Christian Church*, 2nd ed. (1974, reprinted 1983), with informative articles on Roman Catholic subjects and helpful bibliographies. An excellent brief compendium of doctrine is *A New Catechism: Catholic Faith for Adults* (1967; originally published in Dutch, 1966). The contemporary Roman Catholic Church is surveyed by JOHN L. MCKENZIE, *The Roman Catholic Church* (1969, reissued 1971). A balanced and comprehensive introduction is RICHARD P. MCBRIEN, *Catholicism*, 2 vol. (1980), with additional bibliographies. See also BARRIE RUTH STRAUS, *The Catholic Church* (1987). JAROSLAV PELIKAN, *The Christian Tradition: A History of the Development of Doctrine* (1971–), of which 4 vol. had appeared by 1987, opens with the apostolic Fathers and is to close with the second Vatican Council. ROSEMARY RUETHER and ELEANOR MCLAUGHLIN (eds.), *Women of Spirit: Female Leadership in the Jewish and Christian Traditions* (1979), is a step toward redressing the imbalance in most scholarship. For developments in Roman Catholic theology after the second Vatican Council, see HANS KÜNG, *On Being a Christian* (1976, reissued 1984; originally published in German, 1974). On recent developments in Roman Catholic feminist theology, see MARY JO WEAVER, *New Catholic Women: A Contemporary Challenge to Traditional Religious Authority* (1985).

The papacy. Of the general histories, JOHANNES HALLER, *Das Papsttum: Idee und Wirklichkeit*, rev. and enlarged ed., 5 vol. (1950–53, reissued 1965), is a classic. A chronological listing of the popes and antipopes, with concise biographical information, is found in J.N.D. KELLY, *The Oxford Dictionary of Popes* (1986). Useful largely for reference is HORACE K. MANN, *The Lives of the Popes in the Early Middle Ages*, 18 vol. (1902–32), which covers the period to 1304; it is continued for the period to 1800 by LUDWIG PASTOR, *The History of the Popes: From the Close of the Middle Ages*, various editions, 40 vol. (1891–1953; originally published in German, various editions, 16 vol. in 21, 1866–1938). EAMON DUFFY, *Saints and Sinners: A History of the Popes*, 2nd ed. (2002) is the best one-volume history of the papacy. The standard collection of documents is CARL MIRBT, *Quellen zur Geschichte des Papsttums und des römischen*

Katholizismus, 6th ed. rev. by KURT ALAND (1967). An excellent brief introduction to papal history up to the Reformation, including a good bibliography, is GEOFFREY BARRACLOUGH, *The Medieval Papacy* (1968, reissued 1979). For the development of medieval papal claims, see W. ULLMANN, *The Growth of Papal Government in the Middle Ages: A Study in the Ideological Relation of Clerical to Lay Power,* 3rd ed. (1970). For the subsequent crisis of papal authority, see FRANCIS OAKLEY, *Council over Pope? Towards a Provisional Ecclesiology* (1969). EDWARD CUTHBERT BUTLER, *The Vatican Council, 1869–70: Based on Bishop Ullathorne's Letters,* new ed. edited by CHRISTOPHER BUTLER (1962), is a history of the first Vatican Council; see also JAMES J. HENNESEY, *The First Council of the Vatican: The American Experience* (1963). WALTER M. ABBOTT (ed.), *The Documents of Vatican II* (1966, reissued 1982), is an introduction to the achievement of the second council, including commentaries and responses by Catholic, Protestant, and Orthodox scholars.

Works on the papacy from the theological perspective include PAUL C. EMPIE and T. AUSTIN MURPHY (eds.), *Papal Primacy and the Universal Church* (1974), an ecumenical dialogue; and RUDOLF SCHNACKENBURG, *The Church in the New Testament* (1965, reissued 1974; originally published in German, 1961), which presents the results of 20th-century Roman Catholic biblical scholarship. RAYMOND BROWN, KARL P. DONFRIED, and JOHN REUMANN (eds.), *Peter in the New Testament: A Collaborative Assessment by Protestant and Roman Catholic Scholars* (1973), considers the biblical problems in the Petrine question. KARL RAHNER and JOSEPH RATZINGER, *The Episcopate and the Primacy* (1962; originally published in German, 1961), is an analysis of the pope–bishop relationship; and HANS KÜNG, *Infallible? An Inquiry* (1971, reissued 1983; originally published in German, 1970), *The Church* (1967, reissued 1976; originally published in German, 1967), and *Structures of the Church* (1964, reissued 1982; originally published in German, 1962), are basic to an understanding of contemporary "liberal" Roman Catholic thinking on the papacy. For Eastern Orthodox views on the papal primacy, see FRANCIS DVORNIK, *Byzantium and the Roman Primacy* (1966, reprinted 1979; originally published in French, 1964); and J. MEYENDORFF *et al., The Primacy of Peter,* 2nd ed. (1973; originally published in French, 1960).

The Latin Church in the West (1000–1517). The only large-scale work covering the entire period (except for the century 1274–1378) is AUGUSTIN FLICHE and VICTOR MARTIN (eds.), *Histoire de l'église depuis les origines jusqu'à nos jours,* vol. 8–10, 12–14 (1940–64). Another classic is ALBERT HAUCK, *Kirchengeschichte Deutschlands,* 9th ed., 5 vol. in 6 (1958), which covers most of continental Europe to 1437. Good introductions to the medieval church are JOHN A. F. THOMPSON, *The Western Church in the Middle Ages* (1998); and, especially, JOSEPH H. LYNCH, *The Medieval Church: A Brief History* (1995). For a perceptive introduction, see R.W. SOUTHERN, *Western Society and the Church in the Middle Ages* (1970, reprinted 1985). See also R.W. CARLYLE and A.J. CARLYLE, *A History of Mediaeval Political Theory in the West,* 6 vol. (1928–36, reprinted 1970), especially vol. 3–5. ETIENNE GILSON, *History of Christian Philosophy in the Middle Ages* (1955, reprinted 1980), is a masterly summary with full bibliography; DAVID KNOWLES, *The Monastic Order in England: A History of Its Development from the Times of St. Dunstan to the Fourth Lateran Council, 940–1216,* 2nd ed. (1963, reprinted 1966), also covers part of Europe. Other recommended studies include HENRY CHARLES LEA, *The Inquisition of the Middle Ages: Its Organization and Operation* (1954, reissued 1969), 8 chapters from the author's original 1887 3-vol. work; GUILLAUME MOLLAT, *The Popes at Avignon, 1305–1378* (1963, reprinted 1965; originally published in French, 9th ed. 1949); and SCHAFER WILLIAMS (ed.), *The Gregorian Epoch: Reformation, Revolution, Reaction?* (1964), a useful collection of studies by early and more recent authorities.

The late Middle Ages. FRANCIS OAKLEY, *The Western Church in the Later Middle Ages* (1979, reissued 1985); and STEVEN OZMENT, *The Age of Reform (1250–1550): An Intellectual and Religious History of Late Medieval and Reformation Europe* (1980), both cover the period with sound judgment. See also W.A. PANTIN, *The English Church in the Fourteenth Century* (1955, reissued 1980); ROGER AUBENAS and ROBERT RICARD, *L'Église et la Renaissance (1449–1517)* (1951); E. DE MOREAU, PIERRE JOURDA, and PIERRE JANELLE, *La Crise religieuse du XVIe siècle* (1950, reprinted 1956); and L. CRISTIANI, *L'Église à l'époque du concile de Trente* (1948). GEORGES DE LAGARDE, *La Naissance de l'esprit laïque, au déclin du Moyen Âge,* 3rd ed., 5 vol. (1956–70), studies the lay movement in the Middle Ages. JOSEPH LORTZ, *History of the Church,* 2nd ed. (1939, reprinted 1948; originally published in German, 5th–6th ed., 1937), analyzes the history of the church from the point of view of the history of ideas. HEIKO AUGUSTINUS OBERMAN, *The Harvest of Medieval Theology: Gabriel Biel and Late Medieval Nominalism,* 3rd ed. (1983), looks at the theology of the late Middle Ages in its entirety, with special emphasis on Nominalism. See also BRIAN TIERNEY, *Foundations of the Conciliar Theory: The Contributions of the Medieval Canonists from Gratian to the Great Schism* (1955, reprinted 1968).

Reformation and Counter-Reformation. CONRAD BERGENDOFF, *The Church of the Lutheran Reformation: A Historical Survey of Lutheranism* (1967), provides a study of Reformation history from its beginnings to the 20th century. COMMISSION INTERNATIONALE D'HISTOIRE ECCLÉSIASTIQUE COMPARÉE, *Bibliographie de la Réforme, 1450–1648,* vol. 1–7 (1958–70), is a reference work on the history of the Reformation. An important resource is *The New Cambridge Modern History:* vol. 1, G.R. POTTER (ed.), *The Renaissance, 1493–1520* (1957); and vol. 2, G.R. ELTON (ed.), *The Reformation, 1520–1559* (1958). A.G. DICKENS, *Reformation and Society in Sixteenth-Century Europe* (1966, reprinted 1979), is an account of the sociological relationships in the 16th century. HAROLD JOHN GRIMM, *The Reformation Era, 1500–1650,* 2nd ed. (1973), presents a study of the Reformation and the Counter-Reformation. Also useful is HUBERT JEDIN (ed.), *Handbuch der Kirchengeschichte:* vol. 3, pt. 1, *Die mittelalterliche Kirche* (1966); vol. 3, pt. 2, *Vom kirchlichen Hochmittelalter bis zum Vorabend der Reformation* (1968); and vol. 4, *Reformation, katholische Reform und Gegenreformation* (1967). See also G.R. ELTON, *Reformation Europe, 1517–1559* (1963, reissued 1967); PHILIP HUGHES, *The Reformation in England,* 5th rev. ed., 3 vol. in 1 (1963); and JOSEPH LORTZ, *How the Reformation Came* (1964; originally published in German, 3rd ed., 1955), on the causes of the Reformation in England, and *The Reformation in Germany,* 2 vol. (1968; originally published in German, 2 vol., 1939–40), a standard work on the history of the Reformation. JAROSLAV PELIKAN, *Obedient Rebels: Catholic Substance and Protestant Principle in Luther's Reformation* (1964), is an investigation of Luther's thought. Other scholarly works include MAURICE POWICKE, *The Reformation in England* (1941, reissued 1973); GOLO MANN and AUGUST NITSCHKE (eds.), *Propyläen Weltgeschichte: Eine Universalgeschichte,* vol. 7, *Von der Reformation zur Revolution* (1964, reissued 1976); and HERBERT MAYNARD SMITH, *Henry VIII and the Reformation* (1948, reprinted 1964). GEORGES HUNTSTON WILLIAMS, *The Radical Reformation* (1962), is a synoptic presentation of the "left wing" of the Reformation. ERNST WALTER ZEEDEN, *Die Entstehung der Konfessionen* (1965), discusses the formation of various confessions of faith at the time of the development of different denominations, and *Das Zeitalter der Gegenreformation* (1967) covers the battle for the reorganization of the Roman Church.

Roman Catholicism outside Europe in modern times. The standard work on church history is KARL BIHLMEYER, *Church History,* rev. by HERMANN TÜCHLE, 3 vol. (1958–66, reprinted 1966–68; originally published in German, 13th ed., 3 vol., 1952–56). The documents of Roman Catholicism are assembled in HENRICUS (HEINRICH) DENZINGER, *Enchiridion Symbolorum: Definitionum et Declarationum de Rebus Fidei et Morum,* 36th ed. (1976). ROLAND H. BAINTON, *The Horizon History of Christianity* (1964, reissued with the title *Christianity,* 1985), is a well-written, beautifully illustrated, comprehensive introduction to Western Christianity through the centuries and includes references to modern Catholicism worldwide. Much more extensive and valuable, especially because of the excellent bibliography, is KENNETH SCOTT LATOURETTE, *Christianity in a Revolutionary Age: A History of Christianity in the Nineteenth and Twentieth Centuries,* 5 vol. (1958–62, reissued 1973); vol. 1, 3, and 5 concentrate on Roman Catholic themes. AUGUST FRANZEN, *A History of the Church,* rev. and ed. by JOHN P. DOLAN (1969; originally published in German, 2nd ed., 1968), is a convenient brief introduction that includes some modern materials. E.E.Y. HALES, *The Catholic Church in the Modern World: A Survey from the French Revolution to the Present,* new rev. ed. (1960), concentrates on Europe and America. STEPHEN NEILL, *Colonialism and Christian Missions* (1966), and *A History of Christian Missions,* 2nd ed. rev. by OWEN CHADWICK (1986), provide brief and generally fair comments on Catholic ventures. A much more conservative Protestant bias is present in the standard work by ROBERT HALL GLOVER, *The Progress of World-Wide Missions,* rev. and enlarged by J. HERBERT KANE (1960). ROBERT L. DELAVIGNETTE, *Christianity and Colonialism* (1964; originally published in French, 1960), is written by a Roman Catholic and concentrates on Catholic experience but in too narrow a scope. Two works that make aspects of the American Catholic experience readily available to readers are JOHN TRACY ELLIS, *American Catholicism,* 2nd ed. rev. (1969); and the somewhat less adequate THEODORE MAYNARD, *The Story of American Catholicism* (1941, reprinted 1960). GUSTAVO GUTIÉRREZ, *A Theology of Liberation: History, Politics, and Salvation* (1973; originally published in Spanish, 1972), is a provocative introduction to Roman Catholicism in the Third World.

(J.L.McK./F.C.O./M.E.M./M.D.K./J.J.Pe.)

Rome

A capital of kingdoms and of republics and of an empire that defined the Western world in antiquity and left seemingly indelible imprints thereafter, a city called eternal, as the spiritual and physical centre of the Roman Catholic church, and a city whose name evokes pinnacles of artistic and intellectual achievement, Rome has retained all of these attributes: the capital of Italy, a font of religious authority, and a memorial to the cultural glory of the past. Probably more than any other city in the West, possibly more than any other in the world, it is a city whose history has shaped nearly every aspect of its being but, at the same time, whose historical legacy gives it abiding influence in the modern world.

For well over a millennium, Rome (Latin and Italian: Roma) controlled the destiny of all civilization known to Europe, but it then collapsed. Physically mutilated, economically paralyzed, politically senile, and militarily impotent by the late Middle Ages, Rome nevertheless remained a world power—as an idea. The force of Rome the lawgiver, teacher, and builder continued to radiate throughout Europe. Although the situation of the popes from the 6th to the 15th century was often precarious—at times tragic, ridiculous, or shameful—Rome knew glory as the fountainhead of Western Christianity.

When Italy overcame its political division to become a unified nation in the 19th century, ancient Rome—the only state ever to have unified the peninsula in the past—served as a model, and Rome, with its central location between northern and southern Italy, was an obvious choice for the capital of the new country. As Italy's capital, Rome also is the country's most important cultural centre, but it lags behind Milan as an economic force. At the beginning of the new millennium, Rome remained a spiritual pinnacle for millions of Catholics worldwide, Italy's political centre, and an international cultural hub.

This article is divided into the following sections:

Physical and human geography 914
  The landscape 914
    Location and layout
    Climate
    The main streets and their monuments
  The people 917
  The economy 919
    Industry
    Services
    Transportation
  Administrative and social conditions 919
    Government
    Public services
Monuments of the city 919
  The seven hills 919
    The Palatine
    The Capitoline
    The Aventine
    The Caelian
    The Esquiline
    The Viminal and Quirinal
    Other hills
  The Forum 922
  The riverlands 923
    Castel Sant'Angelo and the bridges
    The lower east bank
    The Campus Martius
    The palaces
  The churches 925
    The great basilicas
    Other major churches
  The fountains 927
History 927
  Rome of antiquity 927
    Founding and the kingdom
    The early Roman Republic
    The city of world power
    The late republic
    Municipal reforms of Augustus
    Contributions of later emperors
    Slow decline of the late empire
  The city of the popes 930
    Decay of imperial authority
    Factional struggles: papacy and nobility
    Emergence of the Roman commune
    Period of the Avignon papacy
    The city of the Renaissance
  Evolution of the modern city 931
    Rebuilding and repopulation
    Decline and fall of the papal empire
    Capital of a united Italy
Bibliography 931

Physical and human geography

THE LANDSCAPE

Location and layout. Rome is located in central Italy on the Tiber (Tevere) River, 15 miles (24 kilometres) inland from the Tyrrhenian Sea. The Roman countryside, the Campagna, was one of the last areas of central Italy to be settled in antiquity. The city was built on a defensible hill dominating the last downstream high-banked river ford, where a midstream island facilitated crossing.

The city of the seven hills, of fountains, ruins, monuments, and tourists, lies mostly within the old city walls. Although earlier walls existed, the city's surviving fortifications date to the time of Aurelian, who began building in brick-faced concrete in AD 270. Almost 12 miles long and girdling about four square miles (10 square kilometres), this is the wall that Italian troops had to breach to claim their capital in 1870, and it is still largely intact.

Extent of modern Rome

The ancient walled city of Rome embraces only 4 percent of the modern municipality's 582 square miles (1,507 square kilometres) and is the smallest of the city's 12 administrative zones. The walled centre is divided into 22 *rioni* ("districts"), the names of most dating from classical times, while surrounding it are the city's 35 *quartieri urbani* ("urban sectors"). Within the city limits on the western and northwestern fringes are six large *suburbi* ("suburbs"), while beyond the municipal boundaries the commune of Rome about doubles the area of the city itself.

About six miles out from the centre of Rome, a belt highway describes a huge circle around the capital, tying together the antique roads that led from everywhere to Rome: the Via Flaminia, Via Aurelia, Via Appia. Masses of modern apartment buildings rise in the districts outside the centre, in which the small amount of contemporary construction is inconspicuous. Street frontages and show windows are often rebuilt to keep pace with the times, and the Romans succeed in harmonizing the new, the simply old, and the antique with a talent that they have demonstrated since the first extensions of the republican Forum were made under the emperors.

Small as it is, the old city contains some 300 hotels and 300 *pensioni*, more than 200 palaces, 20 churches, eight of the city's major parks, the residence of the Italian president, the houses of Parliament, offices of city and national government, and the great historical monuments, in addition to thousands of offices, workshops, restaurants, and bars. It is there that the millions of tourists seem to descend annually.

Climate. Rome's hot, dry summer days, with temperatures often above 75° F (24° C), are frequently cooled



The Esposizione Universale di Roma on the outskirts of Rome. The tall building (centre) serves as a corporate headquarters. The dome (right) is that of the covered stadium built for the 1960 Olympics.

Paolo Koch—Rapho/Photo Researchers

in the afternoons by the *ponentino*, a west wind that rises from the Tyrrhenian Sea 15 miles away. The city receives about 33 inches (840 millimetres) of precipitation annually; spring and autumn are the rainiest seasons. Frosts and occasional light snowfalls punctuate the otherwise mild winters, when temperatures average about 45° F (7° C). The *tramontana*, a stormy wind from the north, frequents the city in the winter.

The Via del Corso

The main streets and their monuments. The main street in central Rome is the Via del Corso, an important thoroughfare since classical times, when it was the Via Flaminia, the road to the Adriatic. Its present name comes from the horseraces (*corse*) that were part of the Roman carnival celebrations. From the foot of the Capitoline Hill, the Corso runs to the Piazza del Popolo and through a gate in the city wall, the Porta del Popolo, there to resume its ancient name. It begins spectacularly with the Vittoriano, the monument to Victor Emmanuel II, first king of united Italy. The nation's unknown soldier was interred there after World War I. A Neo-Baroque marble mountain, it is the whitest, biggest, tallest, newest (1911), and possibly the most pompous of Rome's major monuments. Useful as well as ornamental, it contains a museum of the 19th-century cultural revival.

Along the Corso among the smart shops are five churches, eight palaces (and one palazzetto), and the column of Marcus Aurelius. The first church is S. Marco, the first of Rome's parish churches to be built (*c.* AD 336) on the plan of a classical basilica. The present church, third on the site, dates from the 9th century and was restored in the 15th by the Venetian pope Paul II, who built the Palazzo and the Palazzetto Venezia around the church in 1445, when he was cardinal, enlarging the residence when he became pope. Thereafter, the basilica's priest was always a Venetian cardinal, sharing the palace with the Venetian embassy. Mussolini had his headquarters there and harangued the crowds from the balcony from which Paul II had cheered the carnival races and given his papal benediction. The palace is now a Renaissance art museum and contains the Biblioteca dell'Istituto Nazionale d'Archeologia e Storia dell'Arte (Library of the National Institute of Archaeology and Art History).

While her son Napoleon languished on St. Helena, Madame Letizia languished in the Palazzo Bonaparte, now Palazzo Misciatelli. Across the way is the Palazzo Salviati, built by the Duc de Nevers in the 17th century, owned in the 19th by Louis Bonaparte. The Palazzo Doria is a late 15th-century building behind a 1734 facade. Four mornings a week the public is admitted—through a side door—to the state rooms and the art gallery, in which there are many Titians, Bruegels, and Caravaggios, a Bronzino, a Memling, and a Velázquez portrait and Bernini bust of the family pope, Innocent X. Behind S. Marcello, the Baroque reworking of a church founded in the 4th century, is the mid-17th-century Palazzo Ballestra, in which Bonnie Prince Charlie of Scotland was born in 1720 and to which he returned in 1788 to die.

The column of Marcus Aurelius, with reliefs showing his victory over Danubian tribes, was preserved from the assorted Christian looters of Rome because it was the property of a religious order. In the square around the column, the Piazza Colonna, are the Palazzo Chigi (1562), for many years the Ministry of Foreign Affairs and now the official residence of the prime minister, and the Palazzo Wedekind. Although built in the 19th century, the Wedekind, which now houses a daily newspaper, is not without its plundered antique columns.

The Piazza del Popolo

The Corso emerges onto the splendid oval Piazza del Popolo, which is monumental without being intimidating, a sort of toy theatre stage set magically magnified. Over a period of 300 years, it was constructed as the ceremonial entryway to Rome, and, although its elements are diverse

in style and in age (13th century BC–19th century AD), a remarkable harmony prevails. In 1561 the Porta del Popolo, the medieval gate in the city wall, was rebuilt. Ninety-four years later its inner face was redone by Bernini for the grand entrance of Queen Christina, who had abandoned the Protestant throne of Sweden for the Catholic hospitality of Rome. In 1589 Pope Sixtus V punctuated the plaza centre with an obelisk (13th century BC) brought by Augustus from Heliopolis to the Circus Maximus.

The church next to the gate, Sta. Maria del Popolo, which stood for centuries before the piazza existed and gives its name to the area, was founded in 1227 to replace a 1099 chapel built over what was presumed to be Nero's tomb. It was replaced in 1472–77 by the present-day church, further disguised on the piazza frontage by a Neoclassical facade. The interior is fraught with the works of great Renaissance and Baroque artists. The main chapel has tombs by Andrea Sansovino and frescoes by Pinturicchio. In the Cerasi Chapel are Caravaggio's "Conversion of St. Paul" and his "Crucifixion of St. Peter." The Chigi Chapel, unique for the early 16th century in being a miniature church, was designed by Raphael. Bernini sculpted two of the four prophets in the corners.

At the opposite end of the piazza stand "twin" churches (1662) framing the entrance to three streets. The streets were there first, so the churches were ingeniously squeezed into awkward, different-sized plots between them. Sta. Maria in Montesanto, on the east, has an oval plan and dome, while Sta. Maria dei Miracoli, on the narrower plot toward the Tiber on the west, has a round dome. Carlo Rainaldi, the architect, turned both facades slightly inward to frame the welcoming parades that would proceed up the Corso between the two churches. One of the streets, the Via del Babuino, was one of many built by Sixtus V (1585–90) to try to repopulate parts of Rome deserted after the Gothic wars.

Since lack of water had driven residents off the high ground, he restored the aqueduct of Alexander Severus, the Aqua Alexandrina, and gave it his own first name, Aqua Felice. He laid out new roads, the basis for the modern street plan of Rome. He also built the Vatican Library, saw to the completion of St. Peter's dome, rebuilt the papal palaces of the Vatican, the Quirinal, and S. Giovanni in Laterano (St. John Lateran), refurbishing the squares in front of the last two, and built a new square at Sta. Maria Maggiore. He reerected four obelisks found among the ruins and restored a great number of fountains, dearly beloved of the Romans.

The Piazza di Spagna

An obelisk in the Piazza di Spagna is not his work but was discovered in the piazza in Campo Marzio in 1778 and erected in 1857 to commemorate the 1854 promulgation of the dogma of the Immaculate Conception. The fountain is fed by the Aqua Vergine, Agrippa's aqueduct of 19 BC, which escaped Gothic destruction because it was mainly underground and which was repaired in 1447. When the fountain was planned in the early 1600s by Bernini (whether by father or son has not been established), there was insufficient water pressure for spouting jets, so the Barcaccia (Scow) was conceived, an ancient marble boat foundering endearingly in its marble bath.

The most striking architectural element in the piazza—indeed, one of the most striking in all Rome—is the renowned Scalinata della Trinità dei Monti (known as the Spanish Steps, or Stairs). The staircase is a rare case of the failure of French cultural propaganda, for while they are called the Spanish Steps—the Spanish Embassy moved onto the square in the 17th century—they are unequivocally French. First suggested by the French about the time the Spanish Embassy was being installed, the idea was approved by papal authorities 100 years later and paid for with a legacy from a French diplomat. The stairs ascend to the French-built church and convent of Trinità dei Monti, begun in 1495 with a gift from the visiting French king Charles VIII and restored by Louis XVIII.

Garofalo—Grimoldi



The Piazza Navona with the church of S. Agnese designed by Borromini and, facing it, Bernini's "Fountain of the Four Rivers"; at right is Bernini's "Fountain of the Moor."

Charles Dickens described the steps as thronged with unengaged "artist's models" in regional costume. They are still crowded with loiterers in distinctive dress, students from all over the world. Artists were among the first to move into the area, and some few who have not been shouldered out by galleries and ultra-modish shops retain their studios among the walled gardens of the Via Margutta. Since the end of the 16th century, the Piazza di Spagna, with its innkeepers who followed the artists, has been a stopping place for tourists. Young lords on the Grand Tour of Europe left their heavy touring coaches for refitting in a side street still called Via delle Carozze (Carriage Street). The room on the piazza in which John Keats died in 1821 has been made into a museum. The surrounding streets at both the top and the bottom of the steps are among the smartest shopping streets in Rome.

THE PEOPLE

The knowledge that Rome is eternal, that nothing lasts but nothing changes, gives rise to the local watchword, *pazienza* ("patience"). In this overcrowded, understaffed city, *pazienza* is demonstrated everywhere, every day. Except for brief, lowering, summer-lightning flashes of an underlying volatility, the Roman is apt to be cheery and courteous, a little less operatic in his reactions than many other Italians.

Family life

In Rome, as in the rest of Italy, all children are godsends and are demonstrably, publicly loved, patted, petted, cuddled, and kissed. Unmonied families make sacrifices to provide the biggest possible dolls and the flashiest possible tricycles. This continues far into life, with the man playing the role of adored but respectful princeling to his queen mother and imperious but indulgent king to his wife and children. In society outside the family the important thing is *bella figura,* or keeping face. Thus the *dottore* (the only degree the university of Rome gives is the doctorate) salutes the street sweeper as *capo* ("chief"), a gesture of respect called for by the uniform.

For 1,000 years, to be a citizen of Rome was to hold the keys to the world, to live in safety, pride, and relative comfort. Today there is still considerable pride in being a *Romano di Roma,* a Roman Roman. Among such are the "black nobility," families with papal titles who form a society within high society, shunning publicity and not given to great intimacy with the "white nobility," whose titles were conferred by mere temporal rulers.

Both Romans and visitors alike continue to congregate at the café tables ranged on the plane-tree-shaded sidewalks of the Via Vittorio Veneto, a street of grand hotels, airline offices, and government buildings. Laid out in 1887 from the Villa Borghese gardens to the Piazza Barberini, it runs downhill in a dogleg. During the 15 or so years of peak prosperity in Italian filmmaking, about 1950–65, international film celebrities abounded, and clouds of beautiful career hopefuls drifted among the tables, making the Via Veneto one of the most intriguing—in both senses of the word—streets in the world. The street remains a fashionable thoroughfare, gaily and expensively animated until long after midnight.

At the same hour, less glittering Romans can be found in the Piazza Navona, on the flat plain in the bend of the Tiber that was the Campus Martius of classical times. The piazza retains the shape and some of the remains of Domitian's circus (AD 81–89), which remained intact until at least 1450. This is far more typical of central Rome than the Via Veneto, a mere centenarian and therefore a new street. Mussolini's regime cut some new routes through the city, mainly to render historic sites more accessible, but modern streets are rare in the historic centre.

The most nobly Roman of the Romans

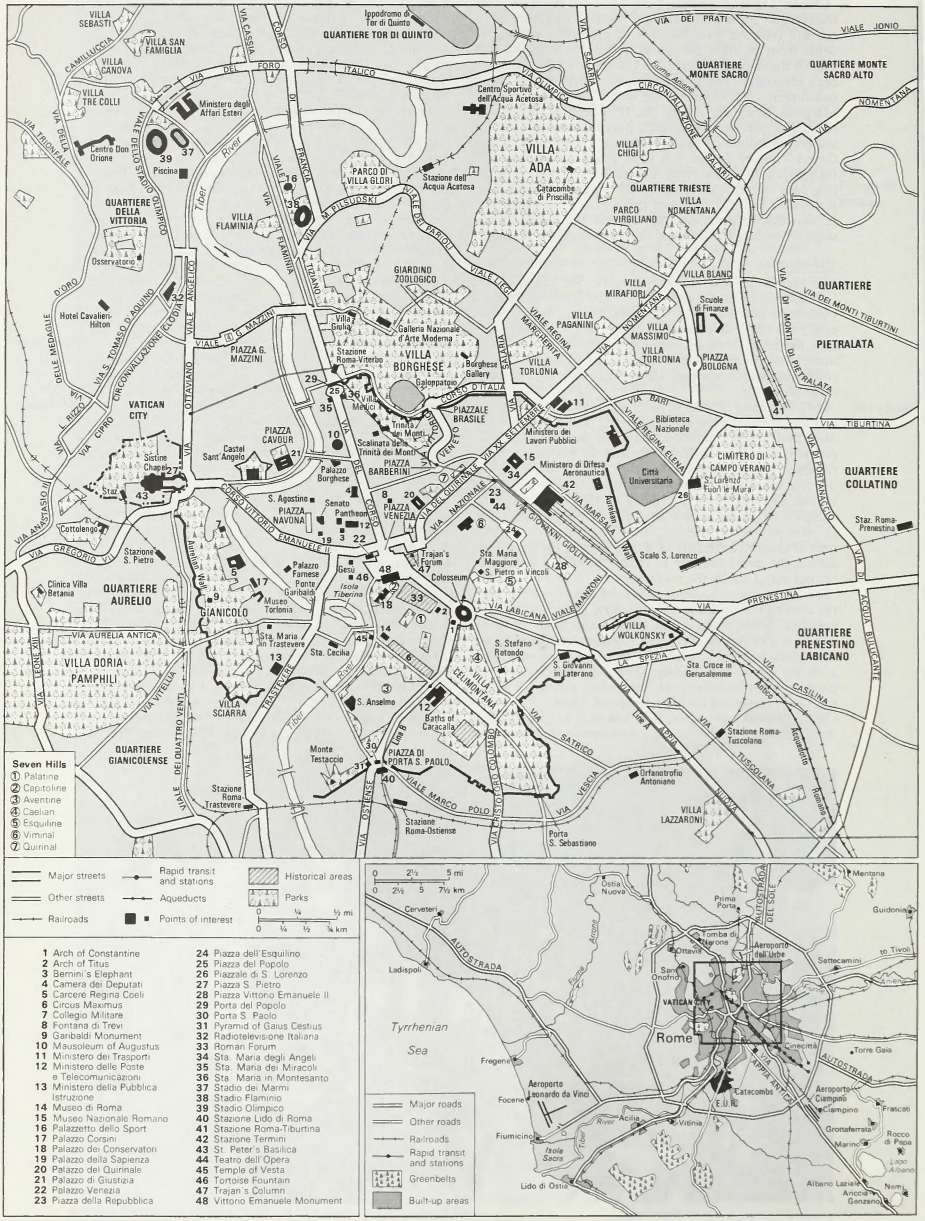
The inhabitants who consider themselves the most nobly Roman of them all are the people of Trastevere (Across the Tiber). They have been in their neighbourhood for a very long time, although they are of neither pure nor primordial stock. Trastevere was the quarter for sailors and foreigners, whereas the founding fathers eastward across the river were soldiers and farmers. In the Middle Ages a number of palaces were the homes of powerful families, and palaces continued to be built during the Renaissance (the Palazzo Farnesina) and even in the 18th century (the Palazzo Corsini). Some authorities—not all from Trastevere—claim Sta. Maria in Trastevere as the oldest church in Rome, pointing out that under the empire the district was the home of Orientals with alien religions, among them a goodly number of Jews proselytized by SS. Peter and Paul. It is said that Alexander Severus (reigned 222–235) permitted Christians to foregather at this site under the leadership of Pope St. Calixtus I, and it is recorded that Pope St. Julius I either raised or rebuilt a church there in 341–352. Today's church is largely 12th-century Romanesque, with a beguiling mosaic facade.

Over the millennia the area has lost little of its vigour. The people have maintained the earthiest of Roman accents, and their taverns have remained generally faithful to simple fare, robust wine, and the unison bawling of irreverent songs. One of Rome's few secular statues—a top-hatted marble effigy of Giuseppe Gioacchino Belli, a

Paolo Koch—Rapho/Photo Researchers



Sidewalk cafe along the Via Vittorio Veneto.


Seven Hills
① Palatine
② Capitoline
③ Aventine
④ Caelian
⑤ Esquiline
⑥ Viminal
⑦ Quirinal
Major streets
Other streets
Railroads
Rapid transit and stations
Aqueducts
Points of interest
Historical areas
Parks
1 Arch of Constantine
2 Arch of Titus
3 Bernini's Elephant
4 Camera dei Deputati
5 Carcere Regina Coeli
6 Circus Maximus
7 Collegio Militare
8 Fontana di Trevi
9 Garibaldi Monument
10 Mausoleum of Augustus
11 Ministero dei Trasporti
12 Ministero delle Poste e Telecomunicazioni
13 Ministero della Pubblica Istruzione
14 Museo di Roma
15 Museo Nazionale Romano
16 Palazzetto dello Sport
17 Palazzo Corsini
18 Palazzo dei Conservatori
19 Palazzo della Sapienza
20 Palazzo del Quirinale
21 Palazzo di Giustizia
22 Palazzo Venezia
23 Piazza della Repubblica
24 Piazza dell'Esquilino
25 Piazza del Popolo
26 Piazzale di S. Lorenzo
27 Piazza S. Pietro
28 Piazza Vittorio Emanuele II
29 Porta del Popolo
30 Porta S. Paolo
31 Pyramid of Gaius Cestius
32 Radiotelevisione Italiana
33 Roman Forum
34 Sta. Maria degli Angeli
35 Sta. Maria dei Miracoli
36 Sta. Maria in Montesanto
37 Stadio dei Marmi
38 Stadio Flaminio
39 Stadio Olimpico
40 Stazione Lido di Roma
41 Stazione Roma-Tiburtina
42 Stazione Termini
43 St. Peter's Basilica
44 Teatro dell'Opera
45 Temple of Vesta
46 Tortoise Fountain
47 Trajan's Column
48 Vittorio Emanuele Monument
Major roads
Other roads
Railroads
Rapid transit and stations
Greenbelts
Built-up areas
Tyrrhenian Sea
Rome
VATICAN
Fregene
Aeroporto Leonardo da Vinci
Fiumicino
Isola Sacra
Lido di Ostia
Ostia Antica
Ostia Nuova
Cerveteri
Ladispoli
Prima Porta
Tomba di Nerone
Ottavia
Aeroporto dell'Urbe
Settecamini
Guidonia
to Tivoli
Torre Gaia
Cinecittà
Catacombs
E.U.R.
Aeroporto Ciampino
Frascati
Grottaferrata
Rocco di Papa
Marino
Lago di Albano
Albano Laziale
Nemi
Anzio
Genzano
Acilia
Vitinia
Focene
Torvaianica
Pomezia
Ardea
Velletri
Tivoli
Monterotondo
Castelnuovo di Porto
Sacrofano
Bracciano
Lago di Bracciano
Cerveteri
Tarquinia
Civitavecchia
Fiumicino
Aurelia
Cassia
Flaminia
Tiburtina
Prenestina
Casilina
Tuscolana
Appia Nuova
Ardeatina
Laurentina
Pontina
Aurelia Antica
Via Ostiense
Via Portuense
Via Trionfale
Via Cassia
Via Flaminia
Via Salaria
Via Nomentana
Via Tiburtina
Via Prenestina
Via Casilina
Via Tuscolana
Via Appia Nuova
Via Ardeatina
Via Laurentina
Via Pontina
Via

19th-century satirical dialect poet—stands near the Ponte (Bridge) Garibaldi.

Most of the streets are still narrow and without sidewalks, appearing only on the most detailed maps and baffling taxi drivers who do not live there. Every 100 paces or so the haphazard cobbled lanes open upon some surprising small plaza with a church, a palace, a cloister, or a group of cafés.

THE ECONOMY

Industry. Rome cannot be called an industrial city, although it has some medium and light industries, including food processing, engineering, electronics, chemicals, printing, and clothing manufacture. The industrial sector steadily declined in size during the late 20th century.

Services. The service sector accounts for most of Rome's employment. Government services are the most important single component of this sector, and state-owned enterprises such as the state television channels, the national railways, and the national airline (Alitalia) also are major job providers. The Roman Catholic church is another important economic force in the city, with its own administrative apparatus, newspapers, television channels, radio stations, and financial entities. Rome is an important publishing centre for Italian books, magazines, and newspapers, including such leading dailies as *La Repubblica, Il Messagero,* and *Il Tempo*. It is also home to Italy's most important film concerns, centred at Cinecittà (Cinema City, built in 1937) a few miles outside of central Rome. The city also is a major commercial hub for central and southern Italy, but it is subordinate to Milan as a national commercial and financial centre. Nevertheless, its retail industry draws shoppers from across Italy. Perhaps the most important private service activity, however, is tourism, which sustains many of Rome's hotels, restaurants, museums, and night clubs. About half of the annual visitors to the city go for religious reasons, either to visit St. Peter's Basilica and the many other Catholic holy places or on church business. Rome also is a frequent host to conferences and trade fairs.

Transportation. Traffic is a typical Roman dilemma because much of the municipal revenue derives from the more than one million automobiles and motor scooters that plague the city. The average noise during waking hours is at or above the level that gradually induces deafness, whereas the speed of motor traffic, in spite of the audacity of the drivers, is about four miles per hour. Damage from pollution has threatened Rome's historical monuments. In the late 20th century, the city made efforts to reduce these problems. Private vehicles were banned from parts of the city's ancient core, to reduce both congestion and noise and air pollution. The city's subway system was gradually extended, and new tram lines were built. Despite these efforts, traffic problems, a lack of parking, and pollution have persisted. Rome is the focal point of Italy's superhighway network. It is also the country's main rail hub, and frequent suburban and long-distance trains link Rome's central Termini Station with the surrounding region and the rest of Italy. The city's main international airport is Fiumicino Airport (also known as Leonardo da Vinci Airport), on the coast 15 miles southwest of the city centre. Local and express trains connect Fiumicino directly with Termini Station. Ciampino Airport, about seven miles southeast, offers limited international services.

ADMINISTRATIVE AND SOCIAL CONDITIONS

Government. Rome is governed by an elected council of 80 members. A mayor and an executive council of 14 members (with four reserves) are selected from among the council members. The council is responsible for such amenities as police protection, health services, transportation, and certain aspects of public assistance.

Public services. Rome is one of the most beautiful and exciting capitals in the Western world. According to local authorities, it is also the filthiest, noisiest, and most heavily indebted city in Italy.

Housing shortage

The city that invented both concrete and the apartment house (*insula*) suffers a perennial housing shortage. The housing shortage persists because of the incessant arrival of job-seeking migrants from all over Italy but mostly from the impoverished south. All the plans, powers, agencies, and even state building funds are available, but three things impede construction: first, land cannot be built upon until the municipality supplies public services and schools (the city is so short of school space that schools sometimes operate classes in three successive shifts for 12 straight hours a day); second, Roman politics are more Byzantine—more labyrinthine and convoluted—than a 5th-century mosaic; third, the notorious glacier-slow Roman bureaucracy can, by paper shuffling alone, delay an approved project up to five years. Life in Rome remains an endless paper chase through the obscure corridors of petty authority.

The city's main institution of higher education is the University of Rome (founded 1303), whose buildings, the Città Universitaria, are located east of Termini Station.

Monuments of the city

Many of the treasures of Rome no longer can be seen where they were placed originally; many can be seen only in other cities of the world, while many others still in Rome represent the spoils of conquest brought to the city from around the ancient world or the cannibalizing of one age or of one faith upon the creations of an earlier one. Rome was sacked first by the Gauls in 390 BC and subsequently by the Visigoths in AD 410, the Vandals in 445, the Normans in 1084, and Spanish troops in 1527. Muslims laid it under siege in 846. The Great Fire of Rome—Nero's fire—occurred in AD 64, and fires and earthquakes ravaged individual buildings or whole areas fairly often over the millennia. But, of all these scourges, it was the stripping of the structures of antiquity for building materials, especially from the 9th century through the 16th, that destroyed more of Classical Rome than any other force. The heritage of the past that survives in Rome is nevertheless unsurpassed in any city of the West, and it is so ubiquitous that its highlights must be comprehended in terms both of geography and of type.

The historic scourges of Rome

THE SEVEN HILLS

The Palatine. The origins of Rome, as of all ancient cities, are wrapped in fable. The Roman fable is of Romulus and Remus, twin sons of Mars, abandoned on the flooding Tiber and deposited by the receding waters at the foot of the Palatine. Suckled by a she-wolf, they were reared by a shepherd and grew up to found Rome, Romulus being obliged to execute Remus for disobeying one of the city's first laws. The Etruscan bronze statue of the maternally ferocious wolf (late 6th or early 5th century BC; Capitoline Museum) is one of the greatest works among the thousands of masterpieces in Rome. The nursing infants were sculpted and placed under the Etruscan statue in 1509.

The wolf cave, the Lupercal, was maintained as a shrine at least until the fall of the empire but is now lost. On the same side of the Palatine, "Romulus' House," a timber-framed circular hut covered in clay-plastered wickerwork, was kept in constant repair. Modern excavations have revealed the emplacement of just such Iron Age huts from the period (8th–7th century BC) given in the fable for the founding of Rome.

On this hill the columns of lost palaces rise in uncompromised beauty from fields of wildflowers and the dust of history. Ilex and pine and bay frame views of Rome. This is the landscape—classical, with figures—that has stirred romantics since it was first limned by 17th-century etchers and sketchers. Before the emperors departed, virtually the entire hill was one vast palace.

Residences of the 1st century

The Palatine was a superior residential district by the 3rd century BC. Augustus was born there in 63 BC and continued to live there after he became emperor. His private dwelling, built about 50 BC and never seriously modified, still stands. Known as the House of Livia, for his widow, it has small, graceful rooms decorated with paintings. Other private houses, now excavated and visible, were incorporated into the foundations of the spreading imperial structures, which eventually projected down into



The Stazione Termini in the Piazza dei Cinquecento and the remains of the 4th-century-BC Servian Wall (left) that abuts the building.
Garofalo—Grimoldi

the Forum on one side and onto the Circus Maximus on the other. The three crests of the hill were flattened in the course of building. The palace was begun by Tiberius, to whose work Nero, Caligula, Trajan, Hadrian, and Septimius Severus made their own additions.

The biggest and richest structure of all was created for Domitian (reigned AD 81–96), whose architect achieved feats of construction engineering not seen before in Rome. Parts of the lavish structure—the richly marbled, centrally heated dining hall of which is among the chambers visible today—were occupied by popes after there were no more emperors, and then the hill was abandoned.

After some six centuries the great Roman families returned to the Palatine, planting 16th-century pleasure gardens and pavilions over past glories. A whole set of rooms from the private wing of Domitian's palace was preserved by incorporation into the Villa Mattei. Atop Tiberius' palace the Farnese family built two aviaries and a garden house and laid out one of Europe's first botanical gardens—some parts of which have escaped archaeological excavation.

The Capitoline. The seat of Roman government, the Capitoline is little changed from Michelangelo's design and represents one of the earliest examples of modern town planning. The centrepiece of this piazza of three palaces is a bronze equestrian statue of Marcus Aurelius, which stood unmolested for ages by the barracks of the imperial guard (later the Palazzo del Laterano) because it was believed to be a statue of Constantine, the first Christian emperor.

The Palazzo Senatorio incorporates remains of the facade of the Tabularium, a state-records office constructed in 78 BC and one of the first buildings to use concrete vaulting and employ the arch with the Classical architectural orders. After a popular uprising in 1143, a palace was built on the site for the revived 56-member Senate, supposedly elected by the people but by 1358 a body of one appointed by the pope; when it was rebuilt to Michelangelo's design, it was called the Palazzo Senatorio (Senate Palace).

The palace of the municipal councillors, the *conservatori*, is on the south side of the square opposite the Palazzo del Museo Capitolino (Capitoline Palace), which, as a papal collection of Classical works offered back to the citizens of Rome by Sixtus IV in 1471, became the first public museum of sculpture in the Western world. Now occupying both the Capitoline Palace and the Palazzo dei Conservatori, as well as a later private palace, the museum contains only objects found in Rome, including the famed Romulus and Remus wolf, the "Capitoline Venus," the "Dying Gaul," and the "Boy with Thorn," as well as the host of portrait busts that can, in imagination, repeople the Forum just below.

The hill was the fortress and asylum of Romulus' Rome. The northern peak was the site of the Temple of Juno Moneta (the word money derives from the temple's function as the early mint) and the citadel emplacements now occupied by the Victor Emmanuel monument and the church of Sta. Maria d'Aracoeli. The southern crest, sacred to Jupiter, became, in 509 BC, the site of the Temple of Jupiter Optimus Maximus, the largest temple in central Italy. The tufa platform on which it was built, now exposed behind and beneath the Palazzo dei Conservatori, measured 203 by 174 feet (62 by 53 metres), probably with three rows of six columns across each facade and six columns and a pilaster on either flank. The first temple, of stuccoed volcanic stone quarried at the foot of the hill, had a timber roof faced with brightly painted terra-cottas. Three times it burned and was rebuilt, always of richer materials. The temple that Domitian built was marble with gilded roof tiles and gold-plated doors. It was filled with loot by victorious generals who came robed in purple to lay their laurel crowns before Jupiter after riding in triumph through the Forum. The Clivus Capitolinus, the antique pavings of which can be walked today, was lined with 40 elephants bearing torches to light the way for Caesar coming in triumph from Gaul. In this centre of divine guidance, the Roman Senate held its first meeting every year. When Petrarch was crowned with laurel among the ruins of the capitol in 1341, it was a harbinger of the Renaissance.

The church of Sta. Maria d'Aracoeli, built before the 6th century, remade in its present form in the 13th, is lined with columns rifled from Classical buildings. It is the home of "Il Bambino," a much loved miracle-performing wooden Christ child who is called to save desperately ill children. At Christmas, adorned in jewels given by the grateful, he can be seen in the church's celebrated manger scene, where he is serenaded by shepherd pipers.

The Aventine. Though considerably built over with modern houses and travelled by modern bus lines, the Aventine still bespeaks a Rome of the past, if not the Classical past. The repeated fires that swept the city de-

stroyed all the republican buildings, and the Temple of Diana remains only as a street name. Under the 4th-century church of Sta. Prisca is one of the best preserved and maintained Mithraic basilicas in the city. The basilica of Sta. Sabina, little altered since the 5th century, is lined with 24 magnificent matching Corinthian columns rescued out of Christian charity from an abandoned pagan temple or palace. The Parco Savello, a small public park, was the walled area of the Savello family fortress, one of 12 that ringed the city in medieval times.

A romantic gem is the Piazza dei Cavalieri di Malta, designed in the late 1700s by Giambattista Piranesi, an engraver with the heart of a poet and the eye of an engineer. To the right of this obelisked and trophied square, set about with cypresses, is the Knight's Priory, residence of the grand master of the Knights of Malta. In 1113 the newly founded order, the Knights Hospitaller of St. John of Jerusalem, was in the Holy Land, whence it was driven to Rhodes, which it held until 1522, thence to Malta until 1789, when the order repaired to its stronghold in a Roman side street. The sovereign military order continues its long history of international medical work.

The Knights of Malta

The Caelian. Almost half parkland, the Caelian includes the public park of Villa Celimontana, once the garden of the Mattei family, who had another on the Palatine, a clutch of palaces in the Campus Martius, and another in the Trastevere quarter. The six churches on the hill date from the 4th to the 9th century.

In the medieval confines of the only fortified abbey left in Rome stands SS. Quattro Coronati, today sheltering nuns and their charges, deaf-mute children. The basilica of SS. Giovanni e Paolo, from the 5th century, stands in a piazza that has few buildings later than the Middle Ages. Alongside the church are the remains of the platform of the Temple of Claudius, partly dismantled by Nero, completely by Vespasian. The round church of S. Stefano Rotondo (460–483) may have been modelled on the Church of the Holy Sepulchre in Jerusalem.

The Hospital of St. John was founded in the Middle Ages as a dependence of S. Giovanni in Laterano (St. John Lateran), just off the hill, and maintains its Romanesque gateway. The Hospital of St. Thomas, established at the same period, has disappeared save for its mosaic gateway, signed by the original Cosmate of the Cosmati school of carvers and decorators and by his father Jacobus. Nearby stands the Arch of Dolabella (AD 10), and not far away are the ruins of Nero's extension of the Claudian aqueduct. Also on the hill is the extensive Military Hospital of Celio.

The Esquiline. Between the Esquiline and the Caelian, the end of the Forum valley is filled by the Colosseum and the Arch of Constantine, with the Palatine edging down from the north. After the fire of AD 64 had destroyed so much of the city, Nero undertook to rebuild the end of it—200 acres (81 hectares)—as a palace for himself: seawater and sulfur water were piped into its baths; flowers were sprinkled down through its fretted ivory ceilings; and the facade was covered in gold, from which the name Domus Aurea, the Golden House. The expropriation so enraged the citizens that his successors hastened to efface all trace of Nero's incredible palace: the ornamental artificial lake was drained and on its bed the Colosseum was erected for free entertainment; Trajan built magnificent baths—also with free admission—atop the domestic wing of the Golden House; and Domitian converted the portico on the edge of the Forum into Rome's smartest shopping street. The obliterators were aided by the fire of AD 104. In 131 Hadrian erected his Temple of Venus and Rome where the vestibule had stood at this end of the Forum; the church and former convent buildings of Sta. Maria Nova were built on the western corner of the temple platform in the 10th century. Less than 70 years after the Golden House had been started, nothing was left of it but a 150-foot gilded statue of Nero. Popular tradition has it that the face was changed with each succeeding emperor, but it was destroyed by one of the early popes.

Nero's palace

The removal was so complete that later Romans could not remember where the Golden House had stood. When the domestic wing was discovered under Trajan's Baths in the 15th century, the rooms painted in the Pompeiian style were thought to be decorated grottoes. Some years later, when Raphael and his friends were let down on ropes to look, the style they imitated in decorating the Vatican loggias was called *grottesche.*

The Colosseum that replaced Nero's lake is more correctly called the Flavian Amphitheatre. It was begun by Vespasian and inaugurated by Titus in AD 80. The oval stadium measures one-third of a mile around, with external dimensions of 615 by 415 feet. The 160-foot facade has three superimposed series of 80 arches and an attic story. The attached columns follow the order applied on the Theatre of Marcellus (13 BC): sturdy, unadorned Doric on the ground floor, more elegant Ionic next, and luxuriant Corinthian on top. The attic story bore corbels supporting masts from which royal sailors manipulated awnings to protect the 50,000 seats from the sun during the gladiatorial contests, combats with wild animals, sham battles, and, when the arena was flooded, naval displays. The main structural framework and facade are travertine, the secondary walls of volcanic tufa, the inner bowl and the arcade vaults of concrete. Until Pius VIII (reigned 1829–30) began conserving what was left, it had been a convenient quarry for 1,000 years.

The Colosseum

The nearby Arch of Constantine was erected hastily in 315 to celebrate a victory two years earlier. Almost all the sculpture on this splendid arch was snatched from earlier monuments: a battle frieze from the Forum of Trajan, a series of Hadrianic roundels, and eight panels from a Marcus Aurelius monument.

Not all the rooms of the Golden House on the Oppio have been excavated. Above them spread the remains of Trajan's Baths, theatrical decorations for the public garden, Parco di Traiano. They served as models for the baths of Caracalla (*c.* 212–217) and Diocletian (298–305/306), which, in turn, served as a pattern for the Basilica of Maxentius. The bath building that housed the hot, warm, cold, and exercise rooms and the swimming pool was a huge, rectangular concrete structure lined with marble. It was surrounded by a garden enclosed in an outer rectangle of libraries, lecture halls, art galleries, and other facilities of a big community centre.

Caracalla's baths on the river flats behind the Caelian Hill covered more than six acres, part of which is occupied today by the modern glass-fronted buildings of the United Nations Food and Agriculture Organization and the Ministry of Posts and Telecommunications. Among the towering remains set in a large park, the caldarium (steamroom) is now used for summer opera performances. Much of the famed Farnese collection of marbles was stripped from these baths.

The Baths of Diocletian are over the brow of the Viminal, and some idea of their size (130,000 square yards, or 110,000 square metres, for the main bath block) can be gained from the fact that the church of S. Bernardo was built into one of the chambers some 500 feet west of the central hall of the 92-foot-high *frigidarium* ("cold room"), into which Michelangelo built the cloister church of Sta. Maria degli Angeli in 1561.

The first seven halls of the Museo Nazionale Romano, also called the Museo delle Terme, are rooms of the *frigidarium* block. This matchless collection of antiquities includes wall paintings from villas, mosaics, sarcophagi, and sculptures, including the famous Ludovisi throne (Greek, 5th century BC), the Niobid from the Gardens of Sallust where the Via Veneto wends today, and the bronze "Pugilist" (2nd century BC), discovered in 1884 in a building site on the Quirinal.

The Basilica of Maxentius (also named after Constantine, who completed it after dispatching Maxentius) was started about 311. This massive hall of justice and commerce was an oblong 265 feet long and 120 feet high, covered by three groin vaults with three deeply coffered tunnel-vaulted bays on either side. Probably ruined by the earthquake of 847, it was also mined for its materials. One of the great Corinthian columns stands obelisk-like before Sta. Maria Maggiore on the Esquiline. The head of Constantine's 40-foot-high statue reposes in the courtyard of the Palazzo dei Conservatori.

The Viminal and Quirinal. Like much of the Esquiline,

the Viminal and Quirinal lie in the heart of modern Rome. Heavily built upon and sclerotic with traffic, the former seems almost flattened under the Ministry of the Interior, the weighty department that directs the state's police forces. The Quirinal, pierced by a modern traffic tunnel, has been a distinguished address since Pomponius Atticus, recipient of Cicero's letters, was a resident. Starting with the Crescentii, who planted the family fortress there in the Middle Ages, powerful Roman families built their homes in this location. The Palazzo Colonna, at the foot of the hill near the Corso, is an art gallery open to the public; and its gardens, climbing the slope to the Piazza Quirinale, contain remnants of Caracalla's Temple of Sarapis. The piazza has been graced since antiquity with two large statues of men with rearing horses, "The Horsetamers" or "Castor and Pollux." Closed on three sides by palaces, the piazza opens on the fourth to a splendid view over the Tiber. The Quirinal Palace, built by Pope Gregory VIII in 1574 as a summer palace away from the heat and malaria of the Vatican, was enlarged and embellished over the next 200 years by a succession of noted architects. The palace, with many extensions and wings, is huge, and its garden is five times as big as the building. From 1550 to 1870, the Quirinal rather than the Vatican was the official papal residence. In 1870 it became the royal palace of the new Kingdom of Italy and in 1948 was made the presidential palace. Both monarchs and presidents, however, have preferred to inhabit the homier palazetto at the far end.

The hill of family strongholds

The handsome buildings opposite are the stables (1730–40), built on the site of the Crescentii 10th-century stronghold. The Palazzo della Consulta (1734) was erected for part of the papal administration. The Palazzo Pallavicini-Rospigliosi, built by a Borghese cardinal in 1603, is still a private house. The Palazzo Barberini farther up the hill, constructed 1629–33 on the site of the old Palazzo Sforza, was occupied by the family until 1949. Part of the collection of the Galleria Nazionale d'Arte Antica is housed here, the rest across the river in the Palazzo Corsini in Trastevere. The 1,700 pictures, most of them works by celebrated masters, were contributed by distinguished families, including the Barberinis. Architecturally, the palace is important, because it marks a departure from the heavy-set four-square town houses of the early and High Renaissance. In the Rome region, only country villas had previously been built on so open a plan, with two wings coming forward from an open, arcaded facade. Further, it pioneered the Baroque style in domestic architecture.

Carlo Maderno, who put the facade on St. Peter's, made the plans, which were carried out after his death by Bernini, assisted by Borromini. Each of these two rivals has a church just around the corner. After 20 years of apprenticeship, Borromini was given his first chance to do his own building. It was an impossibly tiny site at the crossroads of the Quattro Fontane (Four Fountains, one of which is built into a niche in the church wall), but S. Carlo alle Quattro Fontane was a triumph. To his revolutionary solutions of site problems, for which he employed a brilliant variation on the oval, Borromini added a facade in 1667, the year he died, which responded to the waves of motion generated by the spatially complex interior. Walls that flow and sway created a sensation, and the idea was seized upon by Baroque artists, especially from other nations.

Bernini's S. Andrea al Quirinale is also small, but it took 12 years to build (1658–70), late in his career. An oval building with the naves sculpted into the outer wall, it enlarges on concepts advanced by Michelangelo. Bernini's use of coloured marbles and shrewd lighting effects gives the small structure extra dimension. Nearby is the Teatro dell'Opera (Opera House), built in 1880 by Achille Sfondrini. It was acquired by the state in 1926 and is Rome's most important lyric theatre.

Other hills. Behind the river plain of Trastevere is the Gianicolo (Janiculum), and behind the Piazza del Popolo across the river is the Pincio. Both are now parkland, with villas, gardens, and churches discreetly disposed. The Janiculum crest was made into a park in 1870 to honour Garibaldi for his heroic but unsuccessful defense of the Roman Republic in 1849. During the Roman Empire the Pincio was covered with villas and gardens, but it was made into a public park only in the 19th century. By day, nannies wheel their charges through the greenery, and toward sunset Romans arrive to carry on the tradition of the before-dining Pincio promenade. Down the road toward Trinità dei Monti is the 1554 Villa Medici, bought by Napoleon in 1801 to house the Accademia di Francia (French Academy), which is still in occupation. This academy, founded in 1666, is the oldest of many national academies established from the 17th to the 19th century to give architects, artists, writers, and musicians the opportunity to study the vast textbook that is the city itself and to use its museums and libraries.

The French Academy of Rome

The Villa Giulia and the Villa Borghese are also on the hill, both housing art collections of world importance. The Villa Giulia was a typical mid-16th-century Roman suburban villa, conceived not as a dwelling but as a place for repose and entertainment during the afternoon and early evening. It houses the Museo Nazionale di Villa Giulia, which has a collection of Etruscan art and artifacts of singular beauty and historical value. The Borghese collection is small but choice, with a roomful of Caravaggios and, in addition, Titian's "Sacred and Profane Love." Canova's Neoclassical nude statue of Pauline Bonaparte, for a time a Borghese princess, as Venus retains its capacity to scandalize. The Italian government bought the grounds, house, and contents in 1902. The Zoological Garden (established in 1911) on the grounds of the Villa Borghese, is the largest of its kind in Italy and is landscaped to reproduce the natural habitats of the animals. To the west is the Galleria Nazionale d'Arte Moderna, founded in 1883, with an important collection of 19th- and 20th-century Italian art.

THE FORUM

The Forum was the religious, civic, and commercial centre of pastoral, royal, and republican Rome. After Julius Caesar, though it became more imposing, it was only one (albeit the most distinguished) of several complexes serving the same functions. Essentially, it was a small, closed valley ringed by the Seven Hills. There were two meeting places, formal open spaces in the northwest corner, the political Comitium and the social Forum—the name later applied to the entire valley—with shops down both sides. At the other end of the valley was the precinct of the high priest next to the Vestals, the keepers of the sacred flame. Between these two were the temples of the gods. Various emperors opened up the ends of the valley, and there was more building; but the poles of activity did not alter.

Fires, earthquakes, and invasions repeatedly levelled the buildings, and new ones were erected on their remains until the valley was covered by 50 feet of debris, earth, and ashes. Medieval Romans called it Campo Vaccino (Cow Field) and the abutting Capitoline Hill, Monte Caprino (Goat Hill). Excavation began late in the 19th century, and most of the accumulation has been dug away, down to the level at which Julius Caesar knew it.

The little stream cutting diagonally across the valley floor was, according to tradition, canalized as the Cloaca Maxima in the 6th century. Stratigraphic excavations again support the folktales and date the sewer construction at about 575 BC. Although later buildings perpetuated the name and roughly the position of the first halls and temples, they do not necessarily stand where earlier buildings stood, and many details of the earlier Forum are still the subject of scholarly speculation.

Janus and Saturn, both of whom have temples there, were among the gods of early Rome, and the Temple of Vesta, even in its last marble version (AD 191), retained the circular shape of a primitive clay-and-wattle hut. The forge of Vulcan, the Volcanal, had very early beginnings. The Regia, traditionally described as the residence of Numa Pompilius, the priest-king, became the administrative building for the *pontifex maximus*, who took on the monarchy's priestly duties. The Temple of Castor and Pollux was built at the establishment of the republic.

The oldest formally consecrated monument was the open space of the social Forum. A roughly trapezoidal stretch of ground about 125 by 70 feet, it was bare save for three

Building and rebuilding of the Forum

plants essential to Mediterranean agriculture: the grape, the fig, and the olive. Centuries later, when the basilicas were built behind the bordering shops, they served as a protective palisade for the Forum and a covered extension of its open space. At the wide end of the Forum and to one side was the Comitium, in which the Popular Assembly met. Between the two clearings lay the orators' platform, the Rostra, decorated in 338 BC with the iron rams (*rostra*) taken as trophies from the warships of Antium.

At the other end of the Comitium stood the Curia, where the Senate met. When it was destroyed by fire, along with the Basilica Porcia (184 BC, the first of the basilicas), Julius Caesar built a new and greatly enlarged one that encroached on the open space of the Comitium. For the assembly, he built a meeting hall in the Campus Martius, outside the valley altogether. He built a new and much bigger Rostra, though, across the wide end of the Forum. He supplanted the Basilica Sempronia (170 BC) on the western side of the Forum with his own Basilica Julia (54 BC), installing new shops in place of the old Tabernae Veteres. On the other side of the Forum already stood the shop-fronted Basilica Aemilia (179 BC), named for the censor who constructed the Tiber bridge now called the Ponte Rotto.

Caesar also carried his building program onto the flat ground just north of the valley between the Quirinal and Esquiline hills, making his own forum of shops and temple, alongside which Augustus, Trajan, Nerva, and Vespasian later constructed their forums. Pompey's theatre in the bed of the Tiber (55 BC) was followed by the Theatre of Marcellus (13 BC). The great baths, Agrippa's grand concourse in the Campus Martius, the circuses, and the Colosseum all drew the populace away to other centres of activity. The political attraction of the Forum, already vitiated in Caesar's day, continued to decline.

Nevertheless, the halls and temples of the Forum were assiduously rebuilt, ever grander, and more were added. Caesar, after his death, was made a god, and his temple was erected between the Forum proper and the Regia. Eventually, the sacred open space was defiled with honorary columns and an equestrian statue of Domitian. The last thing to be erected in the Forum was a column, raised by Phocas, a Byzantine usurper (608), to honour himself. Septimius Severus placed his arch over the Via Sacra. Other temples were rammed into empty places, and the whole became a forest of towering columns, gleaming walls, and ornate statuary. The dazzling marble mountain of the Palatine flowed down into the Forum as well, and the opposite rim glittered with the splendours of the imperial forums.

The Forum is now a confusing boneyard of history. Of the thousands of columns, not many more than 50 stand erect. Amid the ruins are Christian churches, thickets of trees and bushes, and hundreds upon hundreds of free-living cats.

THE RIVERLANDS

Along a 1½-mile stretch of the Tiber, around a big kangaroo-nosed bend, lie all the historic quarters of the river plain. On the left (east) bank are the Campus Martius, Circus Flaminius, Forum Boarium, and Forum Holitorium; on the right, the Palazzo di Giustizia (Palace of Justice, built 1889–1910), the Castel Sant'Angelo, or Hadrian's Tomb, the entrance to Vatican City, and Trastevere. At the bottom of the bend is Tiber Island.

Castel Sant'Angelo and the bridges. Four of the 11 bridges along this part of the Tiber are of special interest. The Ponte Sant'Angelo, to which Bernini was asked to add angels, is in the main the Pons Aelius built in AD 134. A year later Hadrian began his tomb, just off the end of the bridge. A towering cylinder 20 metres high on a square base, it was in size and form a typical imperial mausoleum. In 271 it was built into the Aurelian Wall and became a key fortress in the defense of Rome. In 587 Gregory the Great, leading a procession to pray for the end to a plague, allegedly had a vision of the archangel Michael atop the tomb. The epidemic ceased and the tomb-citadel became known as the Castel Sant'Angelo (Castle of the Holy Angel). In time it became a papal castle, with richly furnished and frescoed rooms, loggias for the view, a siege store of 5,800 gallons (22,000 litres) of oil and 770,000 pounds (350,000 kilograms) of grain, a centrally heated bathroom, a prison that incarcerated Benvenuto Cellini, among others, and a still-intact fortified passage from the Vatican to carry the pope to refuge there. It is now a state museum with an arboured terrace.

The palace-citadel-tomb on the Tiber

At Tiber Island are two bridges. The Ponte Cestio, often rebuilt since the 1st century BC, leads to Trastevere, while

Authenticated News International



Ruins of the Roman Forum with the Colosseum (left background).

the Ponte Fabricio (62 BC), the oldest in Rome, runs from the shore below the Capitoline. The island, 1,100 feet long and less than 330 feet wide at its widest, has been a place of healing since the Temple of Aesculapius was erected after the plague of 291 BC; the largest building there is the Fatebenefratelli Hospital (also called the Hospital of S. Giovanni di Dio). Facing the hospital is another of Rome's towered medieval family fortresses, this one built by the Pierleone. The traffic howls along both banks, noisier and more voracious than the wolves of the Pierleone's anarchic Rome, but on the island peace prevails. Just downstream are the remains of the Ponte Rotto (Broken Bridge) of 179 BC and two bridges farther along. The modern Ponte Sublicio is named for the wooden bridge defended by Horatius and his comrades on this part of the river.

The lower east bank. On the shore by the Ponte Rotto is the site of the earliest cattle market (Forum Boarium) and vegetable market (Forum Holitorium), girt with temples, of which two remain: the elegant, circular Pentellic marble structure of the 1st century BC and a nicely proportioned, rectangular Ionic building, perhaps a few decades older. Their dedications are disputed, save that they are not, as they are popularly called, temples of Vesta and of Fortuna Virilis. In the 6th century the church of Sta. Maria in Cosmedin was built into the antique grain-commission offices. Some of the Forum Boarium columns can still be seen on the interior of the church, and one of its drain lids, fixed to the outer wall, was carved to represent a face with a gaping mouth. This classical manhole cover became the dread Bocca della Verità (Mouth of Truth), which allegedly would crunch down upon the hand of anyone telling a lie.

Nearby is the Theatre of Marcellus, begun by Caesar and completed in 13 BC by Augustus, who named it for a short-lived nephew. It owes its preservation to its conversion into a fortress for one of the quarrelsome clans of the Middle Ages. Converted into a palace for the Orsinis in the 16th century, it remains private property. The classical orders of the facade, adopted for the Colosseum, became the model for Renaissance architects.

From there northward to the tomb of Augustus and as far inland as the Via Flaminia (modern Corso), the river plain was a vast plantation of temples, baths, and sports grounds until the Middle Ages, when the remaining Romans took up residence there. Today, three major imperial monuments survive: the Pantheon, the reconstructed Ara Pacis Augustae (Altar of Augustan Peace), and Hadrian's Column. Interspersed among the 40 palaces and 100 churches are remnants of what the emperors built.

The portion closest to Tiber Island was once a major republican racing and sports ground, the Circus Flaminius (220 BC), which in the 16th century became the Jewish ghetto. Jews were not persecuted in Rome until Pope Paul IV (1555–59) herded them into a ghetto under curfew. Although Paul was so loathed that the Romans decapitated his statue when he died, other popes carried on his anti-Jewish program. Except for brief respites under Napoleon and the momentary Roman Republic of 1848, Jews until 1870 were debarred from all the professions, government service, and landownership. For many years the neighbourhood retained a Jewish flavour, with some 3,000 Jews living there in the 1960s, but only a few remain, as the ghetto, like Trastevere, became ripe for conversion to luxurious flats. Nearby, the Largo Argentina, excavated 1926–29, contains four small temples of the 1st and 2nd centuries BC.

The Jewish ghetto

The crescent of buildings between the Piazza del Biscione and the Piazza dei Satiri take their curved shape from having been built into and around Pompey's Theatre, the first stone theatre building in Rome. Inspired by the Greek theatre of Mytilene, in which Pompey had been so spectacularly entertained, it had a portico of 100 columns that was equipped to be a community centre almost as much as the baths. The Senate met there on the Ides of March in 44 BC, when Julius Caesar was stabbed 23 times and fell at the foot of Pompey's statue. For almost 400 years a piece of sculpture, unearthed nearby in 1550 and deposited in the Palazzo Spada, was erroneously believed to be the Pompey statue. A part of the theatre was fortified by the Orsinis in the 12th century and later converted into the Palazzo Righetti, or Pio.

The Campus Martius. The rest of the river bend northward was known as the Campus Martius. Marshy in places, with a few temples and public buildings, it was made into one of the grandeurs of Rome by Agrippa (died 12 BC), a landscape of lawns, baths, temples, and parks. The swamp became a lake, the Stagnum Agrippae, where—according to Tacitus—Nero led one of his more elaborate orgies from a sumptuous raft.

Of all this splendour almost nothing remained after the fire of AD 80. Hadrian undertook to restore some of it. Among his works was the new Pantheon, one of the West's great buildings, extraordinary as architecture, remarkable as a feat of engineering. This "Temple of All the Gods," imperial property, survived because it became a church, the gift of the Byzantine emperor Phocas to Pope Boniface IV in 608. This protected the building from everyone but the pope: the bronze roof beams of the grandiose pedimental porch of 18 sixty-ton columns of Egyptian granite were stripped by Urban VIII, the Barberini pope, who took them as raw material for the baldachin in St. Peter's, provoking the celebrated anonymous comment, *Quod non fecerunt barberi, fecerunt Barberini*, "What was not done by the barbarians was done by the Barberinis."

The Pantheon

It has been suggested that the temple was designed by Hadrian himself, whose villa at Tivoli is another landmark in the development of architecture. The Pantheon was possibly the first monumental building of antiquity conceived as an interior. Evenly lighted from a single source—the open eye (*oculus*) in the centre of the dome—the enormous interior, circular and richly marbled, is almost unchanged from classical times. Until the 20th century the dome was the largest ever built, 141 feet in diameter, exactly the height of the building. Two things made its construction feasible: the magnificent quality of the mortar used in the concrete and the meticulous selection and grading of the aggregate, which became lighter in weight with increasing height. Roman concrete was essentially a hydraulic cement, deriving its unique strength from the properties of the dark volcanic ash (*pozzolana*) of the Roman subsoil that was substituted for sand. There is some brick ribbing in the lowest part of the dome and thrust-containing brick outer facing, but, in general, brick was not used by the Romans as a building material in itself. Brick and tile were used to help hold the concrete until it dried, making for a less brutal exterior. The stamped trademarks on the bricks from the big yards behind Vatican Hill and up the Tiber Valley help in determining chronology. The Pantheon, for example, bears the original dedicatory inscription of Agrippa, modestly replaced by Hadrian. The latter's name does not appear, but the stampings on the bricks show that construction does indeed date from Hadrian's reign. The original bronze doors are still in place. Italy's first two kings are buried in the Pantheon, as are many artists, of whom Raphael is the most notable. Nearby are fragments of Agrippa's baths, and the Rome stock exchange gains considerable dignity from the incorporation of some of the Temple of Hadrian.

The shattered drum of Augustus' tomb marks the spot where he was buried AD 14. The mausoleum became a 12th-century Colonna fortress, a 16th-century garden, a ring for Spanish bullfights in the 17th century, and then a concert hall until 1936, when it was scraped down to its impressive but mournful foundations by Mussolini, who may have planned to be buried there himself. Next to the tomb is the delicately beautiful white marble Ara Pacis Augustae (Altar of Augustan Peace, designed 13 BC, dedicated 9 BC). The altar, raised on steps, is enclosed in a sculptured screen. Bits of the friezes were discovered off the Corso in the 15th century, and the altar itself was dug up there in 1938 after 35 years of labour. The pieces unearthed earlier were bought back from museums, and the whole was reassembled to stand four streets away from its original location.

In the Campus Martius Italy's Chamber of Deputies sits in the Bernini-designed Palazzo di Montecitorio, its Senate in Palazzo Madama (17th century), and its Council of State in Palazzo Spada (*c.* 1540), the picture gallery

of which is open to the public. The Museo di Roma, which illustrates the life of the city through the ages, is in Palazzo Braschi (18th century). The Brazilian embassy is in the Palazzo Pamphili, which has a gallery designed by Borromini and painted by Pietro da Cortona. The early 16th-century Palazzo di Firenze was the Florentine embassy until the union of Italy; it is now occupied by the Società Dante Alighieri. The Palazzo della Sapienza, located near the Senate, is now the National Archives, but from 1431 to 1935 it was the seat of the University of Rome (founded 1303).

Architecture of the palaces

The palaces. The three architecturally celebrated palaces in this palace-studded quarter are the Cancelleria, the Farnese, and the Massimo alle Colonne. Because all the pertinent documents were destroyed in the Spanish sack of Rome in 1527, the architect of the Cancelleria remains unknown. Dated 1486–98, it was built by Cardinal Raffaelo Riario out of a night's winnings at the gaming table. Seized by the Medici Pope Leo X (1513–21), it has housed some portion of the Vatican chancellery ever since, except for Napoleonic and revolutionary interruptions. A square building with a rusticated ground floor, its upper stories are plain and rhythmically pilastered, while the columned inner court is noble and deeply harmonious. The city's first High Renaissance building, it could be said to symbolize Rome's displacement of Florence as art capital of the world—its artists drawn from north and south but not from Rome.

The Farnese, the most monumental of Rome's Renaissance palaces, was designed by Antonio da Sangallo the Younger, who was succeeded after his death by Michelangelo, Vignola, and Giacomo della Porta. Sangallo followed the Renaissance precepts regarding the architectural orders on the lower floors, but Michelangelo's top story uses the traditional elements in a willful way, capping it all with an overpowering cornice—a personal expression that foreshadowed Mannerism, a leaching of Renaissance ideals, and the subsequent theatrical self-expression of Baroque. Michelangelo's project to join this palace to the Farnesina by a bridge over the river was actually begun. It can be seen from the surviving arch over the Via Giulia, one of the city's most charming streets.

Mannerist architecture is typified by Baldassare Perruzi's Palazzo Massimo alle Colonne (1535), the name of which comes from a colonnaded palace on the site destroyed in the 1527 sack. It disregards all Renaissance canons, with its brooding entry and heavy cornice below a slightly bowed and airy facade punched with small windows. The Massimo family gave shelter to Konrad Sweynheym and Arnold Pannartz, who produced Rome's first printed book in their house in 1467.

THE CHURCHES

Some 25 of the original parish churches, or *tituli*, the first legal churches in Rome, still function. Most had been private houses in which the Christians illegally congregated, and some of these houses, as at SS. Giovanni e Paolo, are still preserved underneath the present church buildings. Since the 4th century the *tituli* priests have been cardinals who, over the centuries, have rebuilt, enlarged, and embellished their churches.

In the 4th century, basilicas were built to mark the burial places of martyrs. Most martyrs had been interred beyond the city walls in the catacombs, underground galleries with recesses used as tombs. When later sieges of Rome laid waste the Campagna, saintly relics were removed to the safety of city churches. During the Middle Ages, when the prevalence of malaria and of tomb robbers—there was a brisk commerce in religious relics—made ventures beyond the walls risky, some of the oratories and basilicas fell almost to ruin and the location of some catacombs was forgotten.

The great basilicas. Among the basilicas, seven are designated as great (*maggiore*): St. Peter's, S. Paolo Fuori le Mura (St. Paul's Outside the Walls), and S. Giovanni in Laterano, all built by Constantine; and those of S. Lorenzo Fuori le Mura (St. Lawrence Outside the Walls), Sta. Croce in Gerusalemme (Holy Cross in Jerusalem), S. Pietro in Vincoli (St. Peter in Chains), and Sta. Maria Maggiore. Under the 1929 concordat with Vatican City, the Italian government grants them extraterritorial privileges.

The basilicas established the model for Western ecclesiastical architecture for centuries to come. Basilica, a Greek word meaning royal, was used by the pre-Christian Romans to designate a public hall, but no surviving example of a Roman basilica anywhere in the empire is the architectural predecessor of the Christian basilica.

Basic design of the basilicas

The basilical church has a nave higher than the aisles, from which it is separated by a colonnade on each side. It has either a cloistered court (atrium) or anteroom (narthex) or both at the west end and a semicircular projection (apse) at the east. The basilicas in Rome that are closest to the early Christian structures are the churches of competing cults, as strikingly exemplified by the Neo-Pythagorean-sect basilica of the Porta Maggiore, unearthed by the railroad viaduct in 1926.

Some early Christian churches were centrally rather than longitudinally organized, a plan dictated by the circular form of the imperial mausoleums into which they were built. A good example is Sta. Costanza (*c.* AD 320), which also has a superb series of 4th-century vault mosaics in pagan designs. Although churches of this type were few, they had a strong influence on the development of the centrally planned house of worship.

St. Peter's. Protected by the fortified Castel Sant'Angelo, St. Peter's Basilica and the Vatican Palace gained precedence over the cathedral church and Lateran Palace during the papacy's troubled centuries. St. Peter's was built over the traditional burial place of the Apostle from whom all popes claim succession. The spot was marked by a three-niched monument (*aedicula*) of AD 166–170. Excavations in 1940–49 revealed well-preserved catacombs, with both pagan and Christian graves dating from the period of St. Peter's burial.

Constantine enclosed the *aedicula* within a shrine and during the last 15 years of his life (died 337) built his basilica around it. The shrine was sheltered by a curved open canopy supported by four serpentine pillars that he brought from the Middle East. The design, enormously magnified, was followed in making the baldachin (1623–33) over today's papal altar.

In spite of fires, depredations by invaders, and additions by various popes, the original basilica stood for 1,000 years much as it had been built, but in 1506 Julius II ordered it razed and a new St. Peter's built. His architect was Donato Bramante, a Florentine who in 1502 had completed the first great masterpiece of the High Renaissance, the Tempietto in the courtyard of S. Pietro in Montorio, a mile away on the Janiculum Hill. Built to mark the spot where, according to tradition, St. Peter had been crucified, the Tempietto is round, domed, and unadorned. Its outer face is a colonnade of bare Tuscan Doric, the earliest modern use of this order. Because of its proportions, the tiny temple has the majesty of a great monument.

Work of Bramante and Michelangelo on St. Peter's

Bramante's ground plan for St. Peter's was central: a Greek cross, all of the arms of which are equal, around a central dome. Both he and the Pope died before much could be built. Successive architects, including Raphael, drew fresh plans. The last of them, Antonio da Sangallo, died in 1546, and the 71-year-old Michelangelo was solicited to complete Sangallo's projects, which included St. Peter's, the Palazzo Farnese, and the Capitol. He accepted but refused payment for his work on the basilica.

Michelangelo adapted Bramante's original plan, the effect being more emotional and mighty, less classically serene. Of the exterior, only the back of the church, visible from the Vatican Gardens, and the dome are Michelangelo's. After his death Giacomo della Porta and Domenico Fontana, who executed the dome, altered the shape, making it taller and steeper than the original design.

The east end remained unfinished, and it was there that Carlo Maderna was ordered to construct a nave, the clergy having won its century-long battle to have a longitudinal church for liturgical reasons. Thus, St. Peter's orientation reverses the normal. Maderna added a Baroque facade in 1626. He was followed by Gian Lorenzo Bernini, who worked on the building from 1633 to 1677, both inside and outside. His pontifical crowd-funnelling colonnade in

the shape of a keyhole around the piazza, a fountain for the piazza, the breathtaking baldachin, his several major pieces of sculpture, his interior arrangements for the church, and his dazzling Scala Regia (Royal Stair) to the Vatican exhibit his legendary technical brilliance and his masterful showman's flair. Before the lamentable assault in 1972, which damaged the sculptural masterpiece, one could enter the church and, in the first chapel at the right, see the "Pieta" (1499) of Michelangelo in the original splendour.

All the planning, plotting, labour, and faith of all the popes, priests, artists, and artisans produced a vast, gorgeous ceremonial chamber. Amid the gleam and glitter of gold and bronze and precious stones eddy throngs of awed, dwarfed humanity.

S. Giovanni in Laterano. When Borromini redid the interior of S. Giovanni in Laterano (St. John Lateran) in 1646–50, little of the original Constantinian fabric remained after destruction by the Vandals (5th century), damage by earthquake (9th), two devastating fires (14th), and four consequent rebuildings. The Emperor had built a five-aisled basilica over the remains of the barracks of the imperial guard, the Equites Singulares. The bronze doors come from the Curia (the Senate chamber in the Forum); the silver reliquaries containing the heads of SS. Peter and Paul are copies of the twice-stolen originals.

The octagonal 5th-century baptistery replaced that of the 4th, which had been built into the baths of the House of Fausta, Constantine's second wife. (Later, in another palace, she was strangled in the hot room of the bath, a conventional Roman device for suggesting accidental suffocation of an awkward relative.) Its chapels are decorated with mosaics of the period. The cloisters contain some of the finest examples of early 13th-century carved and inlaid decoration called Cosmatesque after the Cosmati, one of several families of traditional craftsmen. (The cloisters of S. Cosimato, S. Paolo Fuori le Mura, and SS. Quattro Coronati are notable examples of this work, which often was accomplished with porphyries and marbles robbed from classical buildings.)

On the exterior a 1732 facade is topped with 15 giant statues that were visible across the city. The piazza around which the Lateran buildings are grouped is decorated with another obelisk, the oldest and tallest in Rome (15th century BC), one of those erected by Sixtus V late in the 16th century. At the same time, he demolished the old patriarchate, from which the Sancta Sanctorum (the papal chapel) and the Scala Santa (Holy Stairs) were preserved.

The Holy Stairs

The Scala had been the principal ceremonial stairway of the palace, but about the 8th or 9th century it began to be identified popularly as having been brought from Jerusalem by St. Helena, Constantine's mother, reportedly from Pilate's palace and thus the stair climbed by the Saviour. The steps are protected by a wooden cover, and believers mount on their knees. The Scala Santa is not mentioned, however, in ecclesiastic, imperial, or personal writings from the 4th, 5th, or 6th century.

Sta. Croce in Gerusalemme. There is similar lack of record regarding St. Helena's acquisition of the True Cross, which is at Sta. Croce in Gerusalemme. This basilica was built into the palace in which St. Helena lived (317–322). At about this time a hall of the palace was converted into a church and two adjoining small rooms were converted into chapels. The rest of the palace continued to be lived in for centuries. The alleged relics of the cross, found in 1492 walled into a niche, are now in a modern chapel. The facade and narthex of the church are 1743 Rococo, the interior an earlier Baroque with a 12th-century Cosmatesque pavement, some antique columns, a few Renaissance details, and, somewhere within it all, part of a palace built around 180–211.

S. Lorenzo Fuori le Mura. Now in the midst of the Campo Verano cemetery, Rome's Catholic burying ground since 1830, S. Lorenzo Fuori le Mura (St. Lawrence Outside the Walls) dates from the 4th century. The nave is a 13th-century basilica built by Pope Honorius III, and the chancel is another basilica built by Pope Pelagius II in the late 6th century as a replacement for the 4th-century original. On the inner part of the triumphal arch between the two is a 6th-century mosaic, and along the walls are giant Corinthian columns of rare marble taken from a non-Christian building.

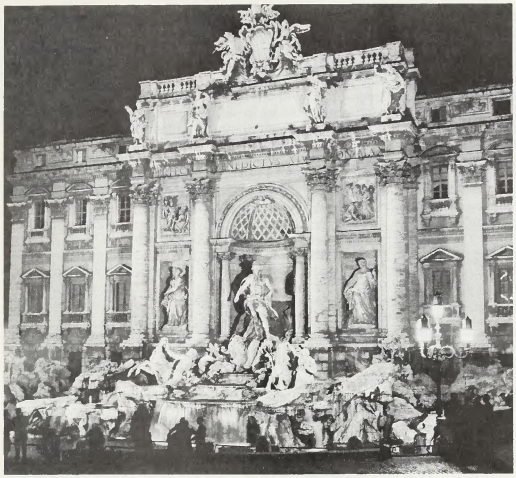
S. Paolo Fuori le Mura. A basilica built by Constantine over the Apostle's grave, S. Paolo Fuori le Mura (St. Paul's Outside the Walls) was replaced starting in 386 by a structure mammoth for its time, 328 by 170 feet. It was faithfully restored after a fire in 1823 and thus remains an outstanding example of early basilical architecture. It has a single eastern apse, a lofty transept, and five majestic nave aisles. Before the Muslim rampage around the walls in 846, the approach to the basilica was a mile-long colonnade down the Ostian Way from the Porta S. Paolo. Today, after leaving the tomb of Gaius Cestius (died 12 BC), a 120-foot pyramid that has inspired many monument builders since, one-third of the route is fenced by gasworks on one side and warehouses on the other.

Sta. Maria Maggiore. Located on the Esquiline, Sta. Maria Maggiore was founded in 432, just after the Council of Ephesus, which raised the Virgin above all created things; it was thus the first great church of Mary in Rome. Behind its Neoclassic facade (1741–43), the original basilica has resisted change. Most of the mosaics date from the time it was built, lining the walls and bursting with blue and gold around the altar. When a new apse was added in the 13th century, it was also decorated with mosaics. Although the ceiling is Renaissance, the slabs of fine marble and the classical columns are pieces of original plunder from other buildings. The great treasure of the church is the Crib of Christ, five pieces of wood connected by bits of metal. Another pope, St. Liberius (352–366), built another church on the Esquiline in response to a vision of the Virgin, who told him to erect a church where snow fell on the night of August 5. In remembrance, it "snows" white flower petals from the roof of the Pope Paul V chapel in Sta. Maria Maggiore every August 5.

Other major churches. *Gesù.* The mother church of the Jesuit order, Gesù, was built 1568–84. Over the following four centuries, it supplied one of the most pervasively influential designs for church building. Michelangelo offered the new order plans for their first church but died before his plans could be acted upon. Building began under Giacomo Vignola (1507–73), very possibly following Michelangelo's ideas. The Jesuits, shock troops of the Counter-Reformation, proselytizers rather than liturgists, needed a new kind of church for their new approach. Vignola combined the central plan (for preaching) with the longitudinal plan (for ritual) by transforming the aisles into a series of chapels opening into the nave. The facade carried the classical orders upward, though only across the width of the tall nave, and the space above the lower aisles to either side was filled with a scroll. The ideas were not new in the history of architecture, but they were new to Rome and new to the age; and they spread with rapidity.

S. Pietro in Vincoli. Originally the Basilica Eudoxiana, S. Pietro in Vincoli (St. Peter in Chains) was built in 432–440 with money from the empress Eudoxia for the veneration of the chains of St. Peter's Jerusalem imprisonment. Later, his Roman chains were added. The chains became famous after they were mentioned at the Council of Ephesus (431). Michelangelo's thunderous Moses is on the tomb of Julius II. Behind the main altar is a 4th-century sarcophagus with seven compartments, brought to Rome from Antioch during the 6th century in the belief that it contained relics of the seven Maccabees.

Sta. Maria della Vittoria. Built 1605–26, Sta. Maria della Vittoria harbours an unfailing crowd pleaser, Bernini's "Ecstasy of St. Teresa" (1645–52). It is a chapel conceived entirely in theatrical terms, even to having the Cornaro family (in marble) seated in opera boxes at the sides of the chapel. Their eyes are directed at the central group in a niche framed in columns, exactly like a proscenium arch, the back wall concealed by gilded metal beams of glory, the scene lighted from above and behind by a hidden yellow-paned window. Amid this setting the angel hovers above the swooning saint, who is—and the illusion is nigh to perfect—borne into the air at the moment of her ecstatic mystical union with Christ. Extraordinarily convincing and utterly voluptuous, it has been both praised as



The Trevi Fountain, designed by Niccolò Salvi in 1732.
Fenno Jacobs—Photo Researchers

a masterwork of consummate spirituality and condemned as an impious, pornographic peepshow.

S. Agostino. Of the scores of churches in the Campus Martius of historical, architectural, and artistic interest, S. Agostino (1479–83) is the most Roman, the church to which would-be mothers come and in which they have offered ex-votos when their prayers have been answered. The "Madonna and Child" (1521) by Jacopo Sansovino, obviously derived from a pagan Juno, is covered with gold and jewels given by the gratified. The church was constructed entirely of travertine looted from the Colosseum. Caravaggio painted the "Madonna with Pilgrims"; Raphael did the fresco of Isaiah. This was these artists' favourite church, and some of the more celebrated among them managed to be interred in it.

THE FOUNTAINS

Rome is as much a city of fountains as it is of churches or palaces, antiquities or urban problems. The more than 300 monumental fountains are an essential part of Rome's seductive powers. Part of the everyday, yet part of the daily surprise, they are points of personal, often sentimental attachment to the city. The Roman composer Ottorino Resphigi found in them inspiration for his orchestral tone poem *Fontane di Roma* (1917). In their ceaseless pouring forth, they also provide a sense of luxury: on her arrival Queen Christina, having watched the fountains in St. Peter's Square, gave her permission for them to be turned off only to learn that they flowed all the time.

Legends of the waters

Every fountain has its history and many have legends, the best known of which guarantees a return to Rome to those who toss coins into the Trevi Fountain. Restored after 1,000 years of silence by Pope Nicholas V in 1485, the fountain was renewed in the 17th century and then transformed from a handy source of household water into a scenic wonder. The huge fountain bulges into most of a tiny square and takes up the entire end of an abutting palace. Niccolò Salvi won a 1732 competition by designing a late Baroque marble mass of rocks and rills, rush and gush, beards and buttocks, all very allegorical and damp. It took 30 years to complete. Its water, from Acqua Vergine, was considered Rome's softest and best tasting; for centuries, barrels of it were taken every week to the Vatican and carried off by the jugful by expatriate English tea brewers. Declared nonpotable in 1961, the waters are now recycled by electric pumps.

Out of the Bernini–Borromini rivalry that so enriched the Roman cityscape arose a legend, still believed and recounted today. This explains that, on Bernini's allegorical Piazza Navona fountain, the statue of the Nile River, whose source was then unknown, hides its head to avoid seeing the Borromini facade on the church opposite, and that of the Río de la Plata raises its arm in alarm to prevent the building from falling. The fountain was, in fact, unveiled in 1651, a year before the church of S. Agnese was begun, two years before Borromini was called in, and 15 years before the facade was completed. The church is owned and maintained by the Doria-Pamphili family.

The oldest of the city's fountains is really a spring, the Lacus Juturnae in the Forum, restored in 1952 to the appearance it had in Augustan times. The newest fountain in the old city is one of the most admired. Inaugurated as simple jets of water in the Piazza Esedra (now the Piazza della Repubblica) by Pius IX just 10 days before the troops of united Italy broke into the city, it was probably the last public work dedicated by a pope in his role of temporal magistrate of the city. In 1901 the nymphs frolicking with sea beasts were added.

The least-liked fountain figure in Rome, unpopular since it was installed in 1587, is on the triumphal arch fountain in the Piazza S. Bernardo, commissioned by Sixtus V. The figure is a pallid Moses, apparently in imitation of Michelangelo's, and its sculptor, Prospero Bresciano, is said to have been so hurt by the public's jeers that he died of a broken heart. (B.E./Ed.)

History

ROME OF ANTIQUITY

Founding and the kingdom. Although the site of Rome was occupied as early as the Bronze Age (*c.* 1500 BC) and perhaps earlier, continuous settlement did not take place until the beginning of the 1st millennium BC. By the 8th century BC, separate villages of various iron-using Indo-European peoples appeared, first on the Palatine and the Aventine hills and soon thereafter on the Esquiline and Quirinal ridges. The artifacts and especially the funerary customs of these communities indicate that, from the beginning, diverse culture groups—including Latins, Sabines, and perhaps others—played important roles in the formation of the future city.

Early peoples and cultures

With the settlement of the valleys between the Palatine,

Esquiline, and Caelian hills in the 7th century, the independent villages began to merge. Before the end of this century, the Forum valley, originally used as a cemetery, was partially drained and occupied by wattle-and-daub huts. The mixed agricultural and pastoral economies of the earliest settlements were slowly exposed to commercial contacts with both Etruscan and Greek traders. The formation of a politically unified city probably occurred in the early 6th century BC under the influence of the Etruscan city-states to the north. Under the rule of its kings, traditionally seven in number (the last three probably Etruscans), Rome became a powerful force in central Italy.

During the regal period, social and economic differences began to shape the two classes, patrician and plebeian, whose struggles for political power dominated the early republic. The tribal organization of the populace was replaced by one based on military units, whose composition in the late regal period depended on property qualifications.

The early Roman Republic. The overthrow of the last Roman king and the establishment of the republic, either in 509 BC or a generation or two later, coincided with the decline of Etruscan power in central Italy. The new government under the leadership of two patrician consuls was at first a mixed blessing. Although Etruscan techniques and symbols survived in republican Rome, commercial ties with the Etruscans and with the Greek colonies in southern Italy gradually withered. During the ensuing economic crisis, grain shortages occurred, a problem that was to plague the city intermittently for a millennium and more; the government was forced to make purchases from as far away as Sicily.

Political upheaval followed economic depression. The first major confrontation between the patricians and plebeians in the mid-5th century led to the writing down of the customary laws in the Law of the Twelve Tables (451–450) and to the formation of a plebeian political organization whose leaders, the tribunes, acted to protect the plebeians from arbitrary patrician actions. In the last half of the 5th century, Rome began again to expand its control over neighbouring territories and peoples, a process that culminated in the conquest of the Etruscan city of Veii in 396.

In 390 Rome suffered a disastrous check when a Gallic army laid siege to the city. After seven months, during which only the Capitoline remained in Roman hands, the Gauls were bought off but left Rome in ruins. The Romans set about reconstructing their city almost immediately, surrounding it with a continuous wall of huge tufa blocks. Later writers attributed Rome's haphazard appearance to the rapid rebuilding during this period; Livy described Rome as looking more like a squatters' community than a planned community. For eight centuries, however, no foreign invader was to breach Rome's walls.

Origins of the haphazard municipal plan

The economic dislocation caused by the Gallic attack helped renew the conflict between the patricians and the plebeians; but, before the end of the 4th century, through a series of judicious compromises, the plebeians had won access to all of the offices of the state, and the actions of the plebeian assembly (plebiscites) had been made legally binding on all Romans. Economic legislation dealing with debt and land distribution was directed toward relieving the distress of the lower classes.

The city of world power. The remarkable though largely unplanned territorial expansion of Rome between 375 and 275 brought lasting economic gains. With control of all of peninsular Italy, Rome established colonies on some of the conquered territories and elsewhere assigned lands to individual Roman citizens. The nearly 60,000 holdings distributed before the middle of the 3rd century helped to solve the pressure of Rome's land-hungry population; nevertheless, by *c.* 250 the city's population had grown to almost 100,000. The booty from conquests also helped to defray the costs of such public works as the building of temples and roads and the improvement of the city's water supply. By the early 3rd century, two aqueducts carried fresh water into the city.

In 264 Rome was drawn into a war with Carthage, the great Phoenician emporium in North Africa. After more than a century of conflict, Rome emerged as the strongest power in the Mediterranean; but the acquisition of an empire, which, for the most part, had not been the conscious desire of the Roman people, brought new social and economic problems to the city itself. During the Second Punic War (218–201) large areas of the peninsula were devastated by invading troops from Carthage, led by the famous general Hannibal; much land was abandoned and many peasants sought refuge in Rome. The growing requirements of a standing army depopulated the countryside and concentrated veterans in the city. The Roman nobility, prohibited by law and by custom from investing in commerce or industry, profited from the economic distress of the peasantry by buying up large tracts of land in central and southern Italy. Slaves, whom Rome's wars in the Mediterranean made available in large numbers, were introduced into Italy as farm labourers and herdsmen, causing further dislocation among the free peasantry. In general, the Roman economy lagged well behind the political development of both city and empire.

The late republic. During the 2nd century, the rapid growth of the urban population and the extension of Roman citizenship led to the effective disenfranchisement of the urban vote. The Senate, now the chief policy-making body of the Roman state, was preoccupied with the problems of the empire and too often ignored the needs of the city. With no separate municipal government, public works and the management of food and water supplies were left to private initiative or to amateur public officials. Nevertheless, some progress did occur. Some of the main streets were paved; drains were covered; and several large basilicas and a new row of shops were built in the Forum. The first stone bridge across the Tiber was completed in 142, and the first high-level aqueduct was erected in 144, allowing settlement on the higher ground of the city's eastern ridges. From the early 2nd century, the river port at the base of the Aventine acquired new warehouses and docking facilities.

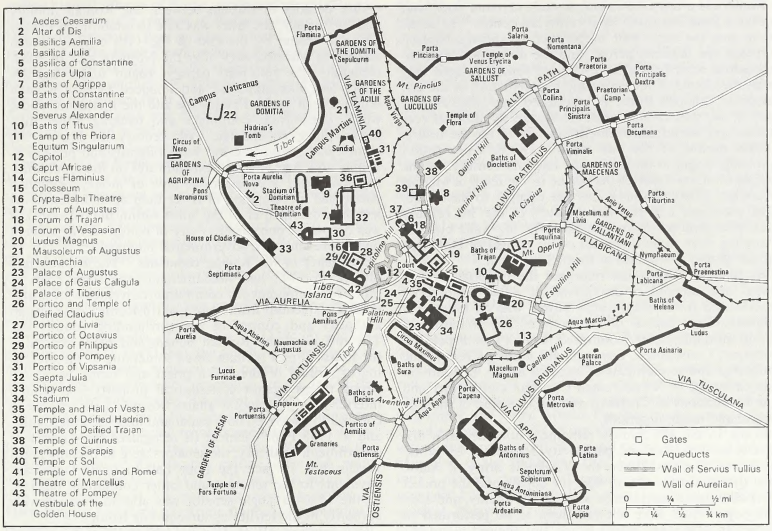
These and other projects, however, were inadequate to deal with the growing urban proletariat increasingly swollen with slaves and freedmen. Crowded into jerry-built apartment houses (*insulae*) and with only minimum employment opportunities in what was an essentially non-industrial city, the lower classes were surviving on the sporadic public-works projects of the state and the largess of the rich before the end of the 2nd century. Rome had, moreover, neither police nor fire protection.

Impact of the Gracchi on municipal squalor

The Gracchi—Tiberius and later Gaius—attempted to deal with the problems of urban unemployment and rising food prices, first by advocating the reestablishment of a small farmer class in Italy, then through the subsidization of the grain supply for the poor. Gaius Gracchus also encouraged public expenditure on roads and buildings. Coupled with currency reforms and heavy government spending, these measures partially restored prosperity to Rome in the late 2nd century, but the basic structural faults in the city's economy and political life remained.

During the civil strife that occupied most of the first half of the 1st century BC, both population and problems multiplied in Rome. The creation of private armies attached to the Roman nobility offered employment to some of the urban lower classes but contributed greatly to the political violence that eventually spelled the end of the republic. Securing an adequate supply of cheap grain offered possibilities for the political manipulation of the urban masses. By the middle of the century, perhaps as many as 500,000 persons were receiving free grain. The upper classes became more interested in luxurious living and their tastes were matched in the public sphere by the building programs of Sulla and Pompey. Public buildings and theatres paid for with tribute and booty enhanced Rome's beauty but did not make a more livable city. In addition, heavy migration to Rome, especially from the Hellenistic east, added to the burdens of the already overcrowded city.

Municipal reforms of Augustus. Julius Caesar, the first to try to deal with the problems of Rome in a systematic way, did not live long enough to carry out his plans, which included canalizing the Tiber and building up the Campus Martius. His adopted son and successor, Augus-



Rome during the imperial period.

tus, attempted to transform Rome into a worthy capital for the new empire. Although his claim that he found the city brick and left it marble is exaggerated, Augustus and his colleagues did provide it with many fine public buildings, baths, theatres, temples, and warehouses. Such construction projects, together with the restoration of old buildings, provided employment for the urban masses; but the lack of any overall city planning left them to live in the unsafe and unsanitary tenements amid the narrow, winding streets and alleys of old Rome. Agrippa, a friend and supporter of Augustus, used his own immense wealth to further enhance the city's beauty and improve its water supply.

Advent of a professional city management

Augustus' reorganization of the administration of the city and his institution of certain public services were a significant break with the republican past. In 7 BC he divided Rome into 14 *regiones* ("wards") and these into *vici* ("precincts"), each with officials who performed both administrative and religious functions. The office of urban prefect, which Augustus revived *c.* 26 BC, did not become permanent until later, but in the late empire the post became the most important in Rome.

In response to an obvious need, Augustus organized a fire brigade in 21 BC, placing a number of public slaves under the command of *aediles*, officials in charge of streets and markets; after a bad fire in AD 6, he established a corps of professional firemen (*vigiles*), comprising seven squads, or cohorts, of 1,000 freedmen apiece. The *vigiles* also had minor police duties, especially at night. He sought to impose order in the often violent streets by creating three cohorts under the command of the urban prefect; their main duty was to keep order in the city, and they could call on the Praetorian Guard for help if necessary. Altogether, Augustus saw to it that the amateur system of Roman municipal administration was replaced by a more professional and permanent set of institutions—a work that probably contributed more to making Rome a great city than all of his marble monuments.

Contributions of later emperors. For the most part, the successors to Augustus continued his administrative policies and building program, though with less innovation and more ostentation. Claudius began a great port near Ostia, at the mouth of the Tiber, to facilitate grain shipments directly to Rome. Commerce remained largely in private hands, with public officials acting to ensure a regular supply and to prevent speculation.

Nero can be credited with introducing the most up-to-date ideas on town planning, though at a terrible price. The great fire of AD 64 destroyed large sections of the city. In the devastated areas Nero built new streets and colonnades as well as his fabulous Golden House, and he encouraged private citizens to build more spacious and more fireproof houses and apartment buildings with better access to the public water supply. Although Nero made Rome a more pleasant city in which to live, his measures did not prevent other devastating fires such as the one in 191 that gave Septimius Severus the opportunity to rebuild the city.

Other emperors in the late 1st and early 2nd centuries added to the glory of the imperial house and the amenities of Roman life—grandiose imperial forums, temples, arches, baths, and stadiums. Trajan's Forum, with its complex of buildings and courtyards, and his market, with its tiers of shops and its great market hall, represent, in the judgment of many historians, the supreme achievement of city planning in Rome. Trajan's Column, which narrates his victories beyond the Danube, was recognized as without peer even in the Christian Middle Ages. Hadrian left two enduring structures in Rome: the great domed Pantheon and his mausoleum, which in AD 590 was renamed Castel Sant'Angelo.

Populace and economy at the height of the empire

In the late 1st and early 2nd centuries Rome was at the peak of its grandeur and population, which has been estimated at more than 1,000,000 persons but was probably less. It was kept at a high level by a steady stream of immigrants, both slave and free, from the provinces and beyond—although life expectancy in the city was probably lower than elsewhere in the empire. Rome's famous paved streets, water supply, and sewage system, however, should not be overestimated; even after the reforms of Nero, large numbers of the urban inhabitants continued to live in expensive, poorly built, overcrowded, and unheated slums without water or cooking facilities. The arena and the public bath relieved some pressures of high density and physical squalor, but Rome's refined technology was applied haphazardly to the problems of urban social organization. Garbage was usually dumped into the Tiber or pits on the city's outskirts.

Rome was a city of consumers, both rich and poor, and never a great industrial or commercial centre. The small shop was the basic unit of production and distribution through the imperial period, and the numerous trade associations served social and religious functions until they were enveloped in the economic regimentation of the late empire. Although Rome far surpassed any other ancient city in size and monumental splendour, its minimal economic and social achievement augured ill for the future.

Slow decline of the late empire. Rome's population probably began to decline in the late 2nd century. At the height of an outbreak of the plague in the reign of Marcus Aurelius, 2,000 persons a day are thought to have died. The economic and political disasters of the 3rd century did little good for Rome. In the 270s the walls built by Aurelian were more a symbol of the danger of barbarian attack than a restoration of Rome's grandeur.

By the time Diocletian reformed the imperial government and ushered in the period of relative prosperity symbolized in his great baths, Rome was no longer the administrative capital of the empire. The founding of Constantinople merely confirmed Rome's loss of political primacy. Constantine, however, did much to restore the buildings and monuments of imperial Rome. In addition, his patronage of Rome's small Christian community laid the foundations of Christian and papal Rome of the medieval and modern periods.

Rome in the 4th century remained, nonetheless, a distinctly conservative and pagan city dominated by proud senatorial families. When the Visigothic army of Alaric first threatened the city in 408, the Senate and the prefect proposed pagan sacrifices to ward off the enemy, and even the pope would have allowed them to be performed in secret. In 410 Alaric seized Rome and allowed his troops to pillage the city for three days; much booty was taken, and many Romans fled.

It is unlikely, however, that the monuments of Rome suffered extensive damage. Its churches, for the most part, were spared. Even the longer, 14-day sack of Rome by the Vandals in 455 did less damage than the Romans themselves. In the 4th and 5th centuries, the emperors repeatedly legislated against those who were stripping buildings and monuments for their materials, especially the marble. By the mid-5th century, the population had dropped to fewer than 250,000.

THE CITY OF THE POPES

Decay of imperial authority. Within a decade of Theodoric's death in 526, Justinian began his attempt to restore Roman imperial rule in the West. His ultimate success was disastrous for Italy and for Rome. Three times Rome was under siege; its aqueducts were cut, and once it was abandoned by its inhabitants. By the end of the century, with the urban population fewer than 50,000, civil authority and the responsibility for protecting the city were in the hands of the church. Pope Gregory I tried to provide an adequate urban administration, and for nearly two centuries his successors played a similar role.

In the middle of the 8th century, when the Byzantines were no longer able or willing to supply Rome with adequate military aid, the papacy turned to the Franks. The "Donation" of Pepin—who owed his new title as king of the Franks in part to the Pope—and that of his son Charlemagne were the theoretical foundations of the temporal power of the papacy. In 774 Charlemagne conquered the Lombard kingdom, and in 800 he was crowned emperor by Pope Leo III and acclaimed by the people of Rome. The period of the late 8th and early 9th centuries was one of vigorous building and restoration of churches in Rome.

Bases of the papacy's temporal power

Factional struggles: papacy and nobility. The decline of Carolingian authority in Italy led to the renewal of family and factional struggles. After the Muslims plundered St. Peter's and the outlying areas of Rome in 846, Pope Leo IV built a wall around the area of the Vatican, thus enclosing the suburb that came to be known as the Leonine City. From the late 9th through the mid-11th century, Rome and the papacy were controlled by various families from Rome's landed nobility, with brief interludes of intervention from the German emperors.

After decades of dispute between the Roman nobility and the papacy, the latter was able to establish an uneasy peace in Rome by the end of the 11th century. Much rebuilding was necessary after the Norman sack of 1084. Generally, the reformed papacy, begun under Leo IX (1049–54), was supported and financed by new Roman families such as the Frangipane and the Pierleone, whose wealth came from commerce and banking rather than landholdings. By the late 11th century the seat of the church had begun to draw many pilgrims and prelates to Rome, and their gifts and expenditures on food and housing stimulated a considerable flow of money. Although Rome had a population of fewer than 30,000 (occupying less than one-quarter of the lands within the old walls), it was becoming once again a city of consumers dependent on the presence of a governmental bureaucracy.

Emergence of the Roman commune. The Roman revolution in 1143 had fundamentally the same goals as other contemporaneous communal movements in northern Italy: freedom from episcopal (in Rome's case, papal) authority and control of the surrounding countryside. The revival of the Roman Senate and other echoes of the Classical past perhaps owed something to the preaching of Arnold of Brescia, a priest and monk, who said strong things against ecclesiastical property and church interference in temporal affairs. Rome's new republican constitution survived both papal and imperial attack alike, and in 1188 Pope Clement III recognized the communal government. In theory, the senators were to become papal vassals, but, in fact, the Pope had to make large cash payments to the senators and other communal officials. In the 1190s a single senator was able to exercise wide authority in the territories surrounding Rome.

Papal and communal authority in conflict

Pope Innocent III made it his first order of business to secure a firm papal position in Rome and in the Vatican. Only moderately successful, he found it expedient to support the Roman commune's expansionist policies. Territorial rivalry between Innocent's family and the Orsini led to rioting and finally open warfare in the streets of Rome in 1204, during which siege machines destroyed many ancient buildings. After a settlement, Innocent's many charitable projects won him Roman support. Gregory IX and the Roman commune clashed over Rome's expansionist policies and its claims to the right to tax the clergy and church property. Bitter struggles with the Hohenstaufen emperor Frederick II as well as the varied interests of Rome's leading families—the Orsini, the Savelli, the Annibaldi, and, above all, the Colonna—complicated the situation. After Frederick's death, an antipapal regime promoted a rising middle class and a resurgence of the commune.

Period of the Avignon papacy. Few popes in the second half of the 13th century were able to reside in Rome. In the 1280s and 1290s, Rome was torn by the bitter rivalries among the Colonna, the Orsini, and the Annibaldi families, a discord encouraged by Pope Boniface VIII. In 1309 Clement V moved the papal residence to Avignon in France; Rome was left to its factional strife and its economic impoverishment.

In spite of sharp rivalries, Roman and papal interests had often coincided throughout the 13th century. Since Rome was never an important industrial or commercial city, its citizens, from the small shopkeepers and innkeepers to the great banking families, had depended economically on the presence of the papal Curia and the large numbers of pilgrims, prelates, and litigants it brought to Rome. The many brick campaniles of its Romanesque churches and the fortress towers on the palaces of its leading families symbolized Rome's ecclesiastical character; but, with a population of never more than 30,000 in the 13th century, it retained a village air for all its urbanity and classical aspirations. Most of the populace was concentrated around St. Peter's and in the low-lying areas of the Campus Martius and Trastevere; large sections of the city within the old Aurelian walls were pastures, gardens, vineyards, and wastelands.

Ambience of the city in the late Middle Ages

The popes in Avignon were able to maintain a tenuous rule over the city, especially Benedict XII (1334–42). The brief popular revolution (1347) of Cola di Rienzo—who,

styling himself tribune of Rome, combined apocalyptic visions with ideas of a renewal of Rome's ancient glories—had more dramatic than political impact. The terrible mortality of the Black Death reduced Rome's population to less than 20,000, and the city staggered through the last half of the 14th century still racked by factional strife. The return of the papacy from Avignon in 1377 did not help. Around 1400, Rome was described as a city filled with huts, thieves, and vermin, and in the neighbourhood of St. Peter's wolves could be seen at night.

The city of the Renaissance. The entry of Pope Martin V (a member of the Colonna family) into Rome in 1420 marked the beginning of the Renaissance city and of the absolute papal rule that lasted until 1870. Although Martin was neither a builder nor a patron of the arts, he laid the foundations of government that made Rome the capital of a Renaissance state. From this period, the apostolic vice chamberlain, as governor of Rome, controlled municipal offices, communal finances, and the statutes of the city. The Roman commune was transformed into a unit of authoritarian papal rule.

Destruction of ancient buildings

From the pontificates of Nicholas V (1447–55) and, especially, Sixtus IV (1471–84), the squalid narrow streets of medieval Rome were widened and paved, and new Renaissance buildings replaced crumbling structures. At the same time, the monuments of ancient Rome suffered further damage as they were torn apart for their building materials. The popes attracted scholars and artists from across Italy, and, by the end of the 15th century, Rome was the principal centre of Renaissance culture. The high point was reached under Leo X (1513–21), with his plans for a new St. Peter's and his patronage of such artists as Michelangelo and Raphael. Rome flourished economically under the Renaissance popes. Banking and the exploitation of alum deposits near Civitavecchia by the popes (with the help of the Medici family of Florence) stimulated a flow of capital into the city. Although Rome once again had become a great consumer of imported luxuries, it still had little large-scale industry or commerce.

EVOLUTION OF THE MODERN CITY

Rebuilding and repopulation. The sack and pillage of Rome in 1527 by the armies of Emperor Charles V ended the city's preeminence as a Renaissance centre. But, even under the repressive rule of the Counter-Reformation papacy, Rome recovered; a new era of construction culminated in a vast program of city planning by Sixtus V (1585–90) and his architect Domenico Fontana. New streets and squares were laid out, obelisks raised, the Lateran and Vatican palaces rebuilt, and aqueducts repaired. Fortunately, his project to convert the Colosseum into a wool factory to provide employment for Rome's prostitutes came to nothing. By 1600, Rome was again a prosperous cosmopolitan city. A great influx of new inhabitants attracted by employment opportunities in the papal bureaucracy and related service industries increased Rome's population to more than 100,000.

Decline and fall of the papal empire. In the 17th and 18th centuries Rome's noble families built fine palaces and patronized the arts while manoeuvring to win high positions in the church hierarchy. The highest prize of all, the papal crown, brought wealth and status to the wearer's family. But as corruption and bribery within these circles became a way of life, the influence of the papacy and of Rome declined throughout Europe and even throughout the Papal States. Although Sixtus V had created one of the best planned cities in Europe, in the 18th century Rome was still a backward town, with poorly paved streets lacking road signs, public lighting, or adequate sanitation.

Continued backwardness and provinciality

The armies of Napoleon occupied Rome for the first time in 1798, and a republic was declared; but in 1809 Rome was annexed into the French Empire. The return of the pope to Rome in 1814 led to a long period of repression and reaction, though popes Leo XII and Gregory XVI promoted educational improvements and new public baths and hospitals. With the liberal attitude that characterized the early part of his reign, Pope Pius IX (1846–78) granted Rome a constitution in 1848; but, after the Revolution of 1848–49, he became an archconservative, attempting with French support to save the temporal power of the papacy and to stave off the modern world.

Capital of a united Italy. Most of the Papal States were included in the Kingdom of Italy, proclaimed in 1861, but France defended papal rule in Rome. Attempts by Garibaldi to capture the city in 1862 and 1867 were unsuccessful, but the withdrawal of the French garrison allowed Italian troops to enter Rome in September 1870, ending 13 centuries of ecclesiastical rule. After a plebiscite in October, Rome became the capital of a united Italy. Pius refused to accept the government's offer of settlement, choosing to style himself a prisoner in the Vatican. The situation was not resolved until 1929, when the Lateran Treaty between Pius XI and Mussolini recognized the pope's sovereignty within Vatican City.

The expansion of government employment and Rome's service sector brought a steady stream of migrants to the new capital. Rome's population grew rapidly after 1870, passing the 500,000 mark before World War I and reaching more than 1,000,000 by 1930. Its area of settlement also expanded for the first time well beyond the old walls of the ancient city.

After World War I, Rome went through a period of unrest as a centre of the socialist and Fascist movements. Benito Mussolini's March on Rome brought thousands of uniformed and often armed Fascists to the capital demanding power in 1922. King Victor Emmanuel III capitulated and pronounced Mussolini prime minister. Under Mussolini in the 1920s and '30s, Rome was transformed into a modern capital, with grandiose new avenues and pompous buildings. Mussolini's encouragement of archaeological excavation contributed to the revelation and preservation of many of the monuments of classical Rome.

Mussolini's transformations

During World War II, Rome underwent a brutal German occupation. Allied bombing destroyed many buildings, and the Germans deported the city's Jews to concentration camps. After the war, thousands of Romans sought shelter in squalid shantytowns on the outskirts.

As the centre of the new Italian republic, Rome began to rebuild but suffered from rampant speculative construction in the 1950s and '60s fostered by a series of corrupt city governments. Some developments, however, such as the Tiburtino housing development and the vast EUR (Esposizione Universale Roma) on the southern outskirts, manifested the best in architecture and planning. Rome also became a mecca for postwar filmmakers such as Federico Fellini and Pier Paolo Pasolini.

The corruption scandals of the 1980s and '90s rocked both the national government and Rome's city government, and a growing regionalist movement challenged and eroded Rome's political dominance in Italy. At the start of the new millennium, Rome had begun to seek a new role based not on its unstable bureaucratic power but instead on its enduring role as a centre of cultural, intellectual, and professional excellence.

BIBLIOGRAPHY

General works: ALTA MACADAM, *Rome and Environs*, 3rd ed. (1985), is a good illustrated guide. RICHARD R. and BARBARA G. MERTZ, *Two Thousand Years in Rome* (1968), is a popular outline, including suggestions for walks and other tourist information.

History: RAYMOND BLOCH, *The Origins of Rome* (1960; originally published in French, 1946), is a well-written survey, archaeologically oriented, with illustrations; H. and A. GELLER, *Jewish Rome* (1970), is a pictorial history of the Jews in Rome from 161 BC, text with plates and bibliography. Sources on the Roman Empire include BARRY BALDWIN, *The Roman Emperors* (1980), a study based on a variety of primary sources; RAMSEY MACMULLEN, *Paganism in the Roman Empire* (1981), a study in social history; CHESTER G. STARR, *The Roman Empire: 27 B.C.–A.D. 476* (1982), an informative analysis of administration and local government; TIM CORNELL and JOHN MATTHEWS, *Atlas of the Roman World* (1982), a comprehensive overview of geographical and cultural setting. JOHN F. D'AMICO, *Renaissance Humanism in Papal Rome* (1983), explores Roman intellectual life in the 15th and early 16th centuries.

Antiquities: RODOLFO LANCIANI, *Ancient Rome in the Light of Recent Discoveries* (1888, reprinted 1975), a fascinating account by one of the best of the old-school archaeologists;

ERNEST NASH, *A Pictorial Dictionary of Ancient Rome*, 2nd rev. ed., 2 vol. (1968), for the archaeologist, art historian, and interested nonspecialist; DONALD REYNOLDS DUDLEY (ed. and trans.), *Urbs Roma* (1967), classical texts on the city and its monuments, relating the literature of the period to its art and architecture—perhaps the best single book on the ancient city for the general reader. Later sources include C. WADE MEADE, *Ruins of Rome: A Guide to the Classical Antiquities* (1980); JAMES PHILLIPS, *Early Christian Architecture in the City of Rome*, 2 vol. (1982); D.P.S. PEACOCK, *Pottery in the Roman World: An Ethnoarchaeological Approach* (1982).

Art: RONALD BOTTRALL, *Art Centers of the World: Rome* (1968), a detailed guide to the principal museums, galleries, and free-standing monuments; ÉMILE MÂLE, *The Early Churches of Rome* (1960; originally published in French, 1942), churches to the 13th century related to the history of their times; MARIANO ARMELLINI, *Le chiese di Roma dal secolo IV al XIX*, rev. ed. by CARLO CECCHELLI, 2 vol. (1942); ROBERT PAYNE, *The Horizon Book of Ancient Rome* (1966). See also ANTHONY BLUNT, *Guide to Baroque Rome* (1982); JUDITH RICE MILLON, *St. Paul's Within the Walls, Rome: A Building History and Guide, 1870–1980* (1982); ROLOFF BENY and PETER GUNN, *The Churches of Rome* (1981); and RICHARD KRAUTHEIMER, *Rome, Profile of a City: 312–1308* (1980).

Daily life in ancient Rome: UGO ENRICO PAOLI, *Rome: Its People, Life and Customs* (1963, reprinted 1975; originally published in Italian, 1940); JEROME CARCOPINO, *Daily Life in Ancient Rome* (1940, reissued 1973; originally published in French, 1939); and J.P.V.D. BALSDON, *Life and Leisure in Ancient Rome* (1969, reissued 1974); three outstanding works on the subject. Later works include H.H. SCULLARD, *Festivals and Ceremonies of the Roman Republic* (1981); SABINE MACCORMACK, *Art and Ceremony in Late Antiquity* (1981); PETER BROWN, *Society and the Holy in Late Antiquity* (1982); BARBARA K. GOLD (ed.), *Literary and Artistic Patronage in Ancient Rome* (1982); and JOHN H. D'ARMS, *Commerce and Social Standing in Ancient Rome* (1981).

Special topics: BERNARD WALL, *A City and a World* (1962), Rome seen in its religious setting and significance; S.G.A. LUFF, *The Christian's Guide to Rome* (1967); GABRIEL FAURE, *Gardens of Rome* (1960; originally published in French, 1959); H.V. MORTON, *The Waters of Rome* (1966; reissued as *The Fountains of Rome*, 1970), an account of the aqueducts of Rome and their principal fountains. See also OLIVER KNOX, *From Rome to San Marino: A Walk in the Steps of Garibaldi* (1982); and RALEIGH TREVELYAN, *Rome '44: The Battle for the Eternal City* (1982).

Personal views: WILLIAM WETMORE STORY, *Roba di Roma*, 8th ed., 2 vol. (1887); STENDHAL, *A Roman Journal* (1957; originally published in French, 1829); AUGUSTUS HARE, *Walks in Rome* (1871; 22nd ed., 1925); ELEANOR CLARK, *Rome and a Villa* (1952, reissued 1982), a personal account of living in Rome and its suburbs and countryside; H.V. MORTON, *Traveller in Rome* (1957); ELIZABETH BOWEN, *A Time in Rome* (1960); AUBREY MENEN, *Rome Revealed* (1960); MAURICE ROWDON, *A Roman Street* (1964); PAUL HOFMANN, *Rome: The Sweet, Tempestuous Life* (1982).

Rome in photographs: MARTIN HÜRLIMANN, *Rome* (1954); WILLIAM KLEIN, *Rome: The City and Its People* (1960); R.S. MAGOWAN, *Rome* (1960).

(R.R.R./B.E./Ed.)

Franklin D. Roosevelt

Franklin Delano Roosevelt, popularly known as F.D.R., the 32nd president of the United States, held office from 1933 to 1945 during the New Deal era and World War II. The modern role of the United States government, in both its domestic and foreign policies, owes much to the changes that Roosevelt helped bring about. To counter the Great Depression of the 1930s he enlisted the powers of the federal government to promote the economic welfare of the U.S. people. He was a leader in the Allied struggle against the Axis powers in World War II, preparing the way for the United States to assume a continuing role in world security. He was the only president to be reelected three times.

EARLY LIFE AND POLITICAL GROWTH

Roosevelt was born at Hyde Park, New York, on January 30, 1882, the only son of James and Sara Delano Roosevelt. The Roosevelts lived in unostentatious and genteel luxury, dividing their time between the Hudson River Valley and European resorts. They often took young Franklin to Europe; he was taught privately at home and was reared to be a gentleman, responsible toward those less fortunate. At 14, Roosevelt, a rather shy youth, entered Groton School (Groton, Massachusetts), modeled after the great public schools of England, where wealthy young men were trained to exercise Christian stewardship through public service.

Schooling and training

After he entered Harvard in 1900, Franklin Roosevelt threw himself into undergraduate activities. His strenuous extracurricular and social life left him relatively little time for his academic studies, in which his record was undistinguished. He was, however, influenced by his economics professors, who modified traditional laissez-faire views with advocacy of government regulation of economic activities, but, even more, Roosevelt fell under the spell of the progressive president, his glamorous distant relative Theodore Roosevelt, a fifth cousin.

During his final year at Harvard, Franklin became engaged to Theodore Roosevelt's niece, Eleanor Roosevelt, who was then active in settlement work in New York City; they were married on March 17, 1905. Eleanor helped open young Roosevelt's eyes to the deplorable living conditions of the underprivileged in the slums.

New York social life interested Roosevelt more than did his studies at Columbia University School of Law. As soon as he passed the New York bar examination, he discontinued his schooling. This attitude of indifference toward the legal profession carried over into Roosevelt's years as a clerk with the distinguished Wall Street firm of Carter, Ledyard & Milburn, defense counsel in several spectacular antitrust cases.

UPI



Roosevelt, 1937.

Early political activities. His admiration for his cousin Theodore, who continued to urge young men of substance to enter public service, led Roosevelt toward politics. His opportunity came in 1910 when the Democratic leaders of Dutchess County, New York, persuaded him to undertake an apparently futile campaign for the state senate. Roosevelt, whose branch of the family had always been Democratic, hesitated only long enough to make sure his distinguished Republican relative would not speak against him.

State senator. He campaigned so strenuously that, with the aid of a Republican schism and his famous name, he won the election. Roosevelt, not quite 29, quickly won statewide and even some national attention by leading a small group of Democratic insurgents who refused to vote for the nominee of Tammany Hall, the New York City organization. For three months Roosevelt helped hold the insurgents firm, until Tammany switched to another candidate.

Roosevelt became the foremost champion in the New York Senate of the upstate farmers, and in the process he converted to the full program of progressive reform. From the New York City legislators, whom he had earlier scorned and now continued to fight, he learned much of the give-and-take of politics. Among them were James J. Walker, later mayor of New York City; Robert Ferdinand Wagner, who became a leading U.S. senator; and Alfred E. Smith, later governor of New York. Roosevelt gradually abandoned his patrician airs and attitude of superiority.

Before the end of 1911, Roosevelt supported the presidential boom for Gov. Woodrow Wilson of New Jersey, the leading Democratic progressive. An attack of typhoid fever kept Roosevelt from participating in the 1912 campaign, but, even without making a single public appearance, he was reelected to the state senate. This was because of publicity by an Albany newspaperman, Louis McHenry Howe, who saw in the tall, handsome young Roosevelt a promising politician. Howe served Roosevelt for the rest of his life with a jealous loyalty.

Assistant secretary of the navy. For his work on behalf of Wilson, Roosevelt was rewarded in March 1913 with an appointment as assistant secretary of the navy under Josephus Daniels. Roosevelt loved the sea and naval traditions, and he knew more about them than did his superior, with whom he was frequently impatient. Roosevelt tried with mixed success to bring reforms to the navy yards, which were under his jurisdiction, meanwhile learning to negotiate with labour unions among the civilian employees. After war broke out in Europe, Roosevelt became a vehement advocate of preparedness; following U.S. entrance, he built a reputation as an effective administrator. In the summer of 1918 he made an extended tour of naval bases and battlefields overseas. During much of his seven years as assistant secretary, he had been less than loyal to Daniels, but in the end he came to appreciate his superior's skill in dealing with Southern congressmen and his solid worth as an administrator.

Paralytic attack. At the 1920 Democratic convention Roosevelt was nominated for vice president. He campaigned vigorously with the presidential nominee, James M. Cox, on behalf of U.S. entrance into the League of Nations. After a Republican landslide, Roosevelt became a vice president of the Fidelity & Deposit Company of Md., a bonding company, entered into numerous business schemes (some of a speculative nature), and remained active in Democratic politics. Suddenly, in August 1921, while on vacation at Campobello Island, New Brunswick, Roosevelt was severely stricken with poliomyelitis. He suffered intensely and for some time was almost completely paralyzed, but he soon began predicting (as he did for some years) that he would quickly regain the use of his

Vice-presidential nomination

legs. His mother wished him to retire to Hyde Park, but his wife and his secretary, Louis Howe, felt it essential to his morale that he remain active in his career and in politics. Because Roosevelt could not himself go to political gatherings, his wife attended for him, acting as his eyes and ears (a service she frequently performed for him during the rest of his life). Under the tutelage of Howe, she overcame her shyness and became an effective political worker and speaker. Because he could not run for office for the time being, Roosevelt was able to function effectively as a sort of premature "elder statesman," trying to promote unity between the urban and rural wings of the Democratic Party. Himself a rural Democrat, he nominated Gov. Al Smith of New York, the favourite of the city faction, at the 1924 and 1928 Democratic conventions.

Smith urged Roosevelt to run for governor of New York in 1928 to strengthen the ticket. Roosevelt was reluctant; he still could not walk without braces and assistance. In the years since 1921 he had worked incessantly to try to regain the use of his legs—for several winters he swam in warm Florida waters and, beginning in 1924, in the mineralized water at Warm Springs, Georgia. Wishing to share with others the beneficent effect of the warm water and a systematic program of therapy, Roosevelt in 1927 established the Warm Springs Foundation, a nonprofit institution for the care of polio victims. He wished to develop Warm Springs further and to continue treatments in the hope of regaining full use of his legs.

Governor of New York. Nevertheless, despite these concerns and his feeling that 1928 was not a propitious year to run on the Democratic ticket, Roosevelt succumbed to strong persuasion and accepted the nomination. When he began campaigning by automobile, he demonstrated that he had retained his youthful buoyance and vitality; he also showed that he had matured into a more serious and human person. Opponents raised the question of his health, but his vigorous campaigning effectively disposed of the issue. Smith was defeated in Herbert Hoover's landslide, and he failed to carry New York state; but Roosevelt won by 25,000 votes.

Succeeding Smith as governor, Roosevelt decided he must establish his own type of administration. He did not keep Smith's closest adviser nor did he depend upon Smith for advice. Smith, already stung by his defeat for the presidency, was hurt and alienated. Whereas Smith had built his reputation on administrative reform, Roosevelt concentrated upon a program to give tax relief to farmers and to provide cheaper public utilities for consumers. The appeal of this program in upstate New York, coupled with the effects of the deepening Depression, led to Roosevelt's reelection in 1930 by the overwhelming plurality of 725,000 votes.

Program as governor

During his first term as governor, Roosevelt's policies, except on the power issue, were scarcely further to the left than those of President Hoover in Washington, D.C. But during Roosevelt's second term, as the Depression became more catastrophic in its effects, he acted to mobilize the machinery of the state government to aid the economy. In the fall of 1931 he obtained legislation establishing the Temporary Emergency Relief Administration, the first of the state relief agencies. Throughout his four years, he was successful in most of his bouts with the Republican legislature, sharpening skills that would prove vital in the future. And, increasingly, beginning with some slight speculation in November 1928, he was being talked of as the most likely Democratic presidential nominee in 1932. After his spectacular victory in 1930, he was so conspicuous a target for the Republicans and for rival Democratic aspirants that he had no choice but to begin immediately and quietly to obtain support for the convention. Because it then took a two-thirds vote in the Democratic convention to nominate, a leading contender could be stopped with relative ease. It soon became apparent that Roosevelt's strongest opposition would come from urban and conservative Eastern Democrats still loyal to Smith; his strongest support was in the South and the West.

Progressives and intellectuals found Roosevelt's overall program attractive, but many feared that he was weak because he sidestepped Republican challenges to oust corrupt Democratic officials in New York City. The opposition became stronger when John Nance Garner of Texas, speaker of the House of Representatives, won the California primary.

PRESIDENCY

At the 1932 convention Roosevelt had an early majority of the delegates but seemed blocked by a combination of the Smith and Garner forces. On the third ballot, Garner allowed his delegates to be thrown to Roosevelt; in return, Garner was nominated for the vice presidency.

First term. In the campaign of 1932 the Depression was the only issue of consequence. Roosevelt, displaying smiling confidence, campaigned throughout the country, outlining in general terms a program for recovery and reform that came to be known as the New Deal. In a series of addresses carefully prepared by a team of speech writers, popularly called the Brain Trust, he promised aid to farmers, public development of electric power, a balanced budget, and government policing of irresponsible economic power. He declared in his most notable speech in San Francisco: "Private economic power is . . . a public trust as well." His program appealed to millions who were nominally Republicans, especially Western progressives. Roosevelt received 22,822,000 popular votes in the election to Hoover's 15,762,000; the electoral vote was 472 to 59. The Democrats also won substantial majorities in both houses of Congress.

Election to the presidency

Inauguration as president. Roosevelt took office on March 4, 1933. Following the election, President Hoover had sought Roosevelt's cooperation in stemming the deepening economic crisis that culminated in the closing of banks in several states during February 1933. But Roosevelt refused either to accept responsibility without the accompanying power or to subscribe to Hoover's proposals for reassuring business; Hoover himself granted that his proposals would mean "the abandonment of 90 per cent of the so-called new deal."

When Roosevelt took office, most of the nation's banks were closed, industrial production was down to 56 percent of the 1929 level, 13,000,000 or more persons were unemployed, and farmers were in desperate straits. Even the congressional leaders were so shaken that for the time being they were ready to follow Roosevelt's recommendations.

In his inaugural address Roosevelt promised prompt, decisive action and somehow conveyed to the nation some of his own unshakable self-confidence. "This great Nation will endure as it has endured, will revive and will prosper," he asserted, ". . . the only thing we have to fear is fear itself." For the moment, people of all political views were Roosevelt's allies, and he acted swiftly to obtain enactment of the most sweeping peacetime legislative program in U.S. history.

The Hundred Days. Through a broad array of measures, Roosevelt first sought quick recovery and then reform of the malfunctions in the economic system that he thought had caused the collapse. He tried to aid each of the main interest groups in the U.S. economy and, at a time when the Democrats were the minority political party, to hold the backing of many who were previously Republicans. His choice of Cabinet members indicated his efforts to maintain a consensus; it was geographically and politically balanced, containing both liberal and conservative Democrats, three Republicans, and, for the first time, a woman—Secretary of Labor Frances Perkins.

Attempts to maintain a consensus

He also directed his legislative program toward a broad constituency. The prelude was the enactment of several conservative measures, to inspire confidence among businessmen and bankers. First Roosevelt ended depositors' runs on banks by closing all banks until Congress, meeting in special session on March 9, could pass a cautious measure allowing those in a sound condition to reopen (Roosevelt also strongly favoured banking reform, but it came later). In March, Roosevelt redeemed one of his most important campaign pledges by introducing a program of drastic government economy. He firmly believed in economy and never became a convert to the Keynesian views so often attributed to him. The emergency banking and economy acts brought him the enthusiastic support of an

overwhelming proportion of the electorate but, he pointed out at the time, could do little to bring real recovery.

Roosevelt was already preparing, and he soon sent to Congress, a series of messages and draft bills proposing the program that comprised the early New Deal. Roosevelt first obtained from Congress federal funds for the relief of human suffering. Congress established the Federal Emergency Relief Administration (FERA), which granted funds to state relief agencies for direct relief. It also established a Civilian Conservation Corps (CCC), which at its peak employed 500,000 young men in reforestation and flood-control work; it was a favourite project of Roosevelt's and remained popular through the New Deal. Mortgage relief aided other millions of persons, both farmers and homeowners. The key loan agency of the New Deal was the previously established Reconstruction Finance Corporation (RFC), the powers of which were broadened so that it could make loans to small enterprises as well as to large. Although at the time Roosevelt did not envisage public spending as the primary role of these relief agencies, the agencies poured so much money into the economy that within several years they were stimulating recovery.

Recovery measures. The two key recovery measures of the New Deal were acts to restore farm prosperity and to stimulate business enterprise. The first act, in 1933, established the Agricultural Adjustment Administration (AAA), the objective of which was to raise farm prices and increase the proportion of the national income going to farmers. The principal means was through subsidies given to growers of seven basic commodities in return for their willingness to reduce production. The subsidies were to be paid from a processing tax on the commodities. Roosevelt accepted this scheme as a temporary expedient, which Congress would enact because a majority of farm organization leaders favoured it. He also hoped to raise farm prices through mild inflation. Roosevelt envisaged a program, following farm recovery, of extensive rural planning—moving farmers from submarginal to better lands and luring some of the unemployed from metropolises to rural and village life. In 1935 Roosevelt obtained the Resettlement Administration, which gave some aid to smaller, poorer farmers. When the Supreme Court invalidated the processing tax in 1936, he switched the AAA program to one of soil conservation. Nevertheless, throughout the New Deal, farm leaders and Congress succeeded in maintaining an agricultural program the major emphasis of which was to raise farm prices. Thanks to this legislation and several years of drought, production fell, and farm income gradually improved. But not until 1941 did it reach even the inadequate level of 1929.

National Industrial Recovery Act

The demand of businessmen for government stabilization and of labour for a shorter workweek led Roosevelt to recommend to Congress the National Industrial Recovery Act (NIRA) of 1933. It was a two-pronged program. On one side was a $3,300,000,000 appropriation for public works. Had this money been poured into the economy rapidly, it would probably have done much to bring recovery, but Roosevelt wanted to be sure it would be spent soundly on self-liquidating public works, through the Public Works Administration (PWA). Because careful planning took time, the PWA did not become an important factor until late in the New Deal. On the other side of the NIRA was a National Recovery Administration (NRA), to administer codes of fair practice within given industries. At first under a "blanket code," then under specific codes negotiated by representatives of each industry and labour, minimum wages, maximum hours, and fair trade practices were established within each industry. The codes were designed to stabilize production, raise prices, and protect labour and consumers. Consumers received scant protection, but labour received guarantees on wages and hours and also the right to bargain collectively. During the summer of 1933 there was a quick flurry of recovery as manufacturers produced goods in anticipation of sale at higher prices under the codes; the boom collapsed by fall because prices had risen faster than purchasing power.

By February 1934 the code making was over, but far too many—557 basic codes and 208 supplementary ones—had come into existence, containing innumerable provisions that were difficult to enforce. By 1935 the business community, which had demanded the NRA at the outset, was becoming disillusioned with it and blaming Roosevelt for its ineffectiveness. In May the Supreme Court, in the Schechter decision, invalidated the code system. Despite shortcomings, however, the NRA had aided several highly competitive industries, such as textiles, and brought reforms that were re-enacted in other legislation: federal wages-and-hours regulation, collective-bargaining guarantees, and abolition of child labour in interstate commerce.

In the fall of 1933 Roosevelt had already turned to other expedients for bolstering the economy. He experimented with "managed currency," driving down the gold content of the dollar and tripling the price of silver through large purchases. These efforts brought only small price increases at home, but they improved the position of the United States in foreign trade by making dollars cheaper abroad. In January 1934 Roosevelt stabilized the gold content of the dollar at 59.06 percent of its earlier value. Managed currency created a significant precedent, even though it did little to bring recovery at the time.

Altogether, by the fall of 1934 Roosevelt's program was bringing a limited degree of recovery, but it was alienating conservatives, including many businessmen. They contended that much of the program was unconstitutional, that it created uncertainties for business that hampered recovery, and that the lowering of the gold content of the dollar had deprived holders of government obligations of their just return. At the same time, many of the underprivileged who were still in serious difficulties felt that the New Deal had not gone far enough. They were ready to listen to demagogic leaders offering still more. In the 1934 mid-term election they voted overwhelmingly for Democratic candidates for Congress; but there was a danger that in the 1936 presidential election they might vote for a third-party candidate to the left of Roosevelt.

Opposition from the right and left

Reform measures. To meet the threat to his political coalition from the left, Roosevelt emphasized reform in his annual message to Congress in January 1935. This was less a shift from a first to a second New Deal than it was a rush to enact reform measures that Roosevelt had long been planning. In 1933 he had obtained the Tennessee Valley Authority (TVA), to provide flood control, cheap hydroelectric power, and regional planning for an impoverished region. At his recommendation also, Congress had enacted two laws to protect investors: the Truth-in-Securities Act of 1933 and an act establishing the regulatory Securities and Exchange Commission (SEC) in 1934.

Additional legislation in 1935 did much to undermine the appeal of demagogues to the needy, especially the Social Security Act, which included unemployment insurance and old-age insurance. For workers still unemployed, Congress created the Works Progress Administration (WPA), to provide relief that would stem the erosion of their skills and self-respect (between 1935 and 1941 the WPA employed an average of 2,100,000 workers, and by the end of 1935 it was already bringing a marked measure of recovery by pouring billions of dollars into the economy). For workers who were employed, the National Labor Relations Act (the Wagner Act), only belatedly accepted by Roosevelt, strengthened the government guarantees of collective bargaining and created a National Labor Relations Board (NLRB) to adjudicate labour disputes. The Public Utility Holding Company Act, also of 1935, regulated the control holding companies had over operating public utility companies. A new 1935 tax measure, labelled by its opponents the "soak-the-rich" tax, raised the levies on persons with large incomes and on big corporations and became a significant factor in redistributing U.S. income.

Second term. These measures effectively undercut the left-wing opposition to Roosevelt, but they further alienated conservatives. He ran for reelection in 1936 with the firm support of farmers, labourers, and the underprivileged; and the epithets that the extreme right hurled at him merely helped unify his following. The Republican nominee, Gov. Alfred Mossman Landon of Kansas, a moderate, could do little to stem the Roosevelt tide. Roosevelt received 27,752,000 popular votes to Landon's 16,680,000 and carried every state except Maine and Vermont.

Reelection in 1936

Supreme Court fight. The only hope of conservatives to thwart the New Deal was for the Supreme Court to invalidate its key measures. Following the Schechter decision, the court in 1936 ruled against the AAA processing taxes, and cases challenging the Social Security Act and the Wagner Act were also pending. Roosevelt, beginning his second term with a massive mandate, was determined to remove this threat. Believing that the measures were well within the scope of the Constitution and that the reasoning of the justices was old-fashioned and at fault, he proposed early in 1937 the reorganization of the court, including the appointment of as many as six new justices. The proposal, labelled by opponents as a courtpacking scheme, touched off a vehement debate in which many of Roosevelt's previous supporters in and out of Congress expressed their opposition. Meanwhile, in the spring of 1937, the Supreme Court upheld both the Wagner Act and the Social Security Act. With the need for the court plan dissolving, its enemies managed by summer to bring about its defeat. This was a severe political blow for Roosevelt, even though the new decisions by the court opened the way for almost unlimited government regulation of the economy.

Growing opposition. Roosevelt's prestige dropped further in the summer of 1937, when much of the public blamed him for labour difficulties that grew out of organizing drives in the steel, automobile, and other mass-production industries. Operating under the protection of the Wagner Act, the unions engaged in strikes that often resulted in violence. Roosevelt himself preferred paternalistic government aid to all workers, such as the wages-and-hours guarantees of the Fair Labor Standards Act of 1938. But union membership jumped to about 9,500,000 by 1941, while most middle class people returned to the Republican Party.

A sharp economic recession in the fall of 1937 added to Roosevelt's troubles. There had been substantial recovery by 1937; but Roosevelt, wishing to balance the budget, had curtailed government spending drastically, sending the economy plummeting back toward 1932 levels. Businessmen blamed the New Deal spending policies; Roosevelt blamed the businessmen and inaugurated an antimonopoly program. In October 1937 massive government spending began again, and by June 1938 the crisis was past.

From 1938 on, many of the conservative Southern Democrats heading key congressional committees openly opposed the New Deal. In 1938 Roosevelt tried unsuccessfully to defeat several of them in the primaries and was inveighed against as a dictator trying to conduct a "purge." Democrats won the November elections, but the Republicans gained 80 seats in the House and seven in the Senate, permitting a coalition of Republicans and conservative Democrats that could thwart the President.

Nevertheless, the second Roosevelt administration saw the passage of some notable reform legislation, extending and improving earlier legislation and moving into some new fields. The development of soil conservation to stem erosion and the large-scale construction of public works, including public housing and slum clearance, also occurred during these years. Many New Deal innovations, such as social security, the agricultural program, the TVA, and the SEC, had now become accepted as permanent functions of the federal government.

Foreign policy. By 1939 foreign policy was overshadowing domestic policy. Even before taking office, Roosevelt had endorsed Hoover's refusal to recognize Japanese conquests in Manchuria. From the outset of his administration, Roosevelt was deeply involved in foreign-policy questions, mostly relating to the Depression. In the early summer of 1933 he refused to support international currency stabilization at the London Economic Conference, but by 1934 he had stabilized the dollar and had begun helping France and Great Britain to keep their currencies from being undermined by dictator nations. In November 1933 Roosevelt recognized the government of the Soviet Union in the mistaken hope that he could thus promote trade. Greater opportunities seemed to exist in negotiating reciprocal trade agreements with numerous nations—a program that began in 1935—and in fostering more cordial relations with Latin American nations. In his first inaugural address Roosevelt had pledged himself to the "policy of the good neighbor." Secretary of State Cordell Hull had interpreted this to mean no unilateral U.S. intervention in Latin America; but, gradually, as European war became imminent, the Good Neighbor Policy led to collective-security and mutual-defense agreements.

Good Neighbor Policy

In the early New Deal years, Roosevelt not only pursued programs of economic nationalism but, like most Americans, was also intent upon keeping the United States out of any impending war. He thus supported a series of neutrality laws, beginning with the Neutrality Act of August 1935. Roosevelt moved toward a new policy in 1937, after Japan began a major thrust into northern China. In October, speaking in Chicago, he proposed that peace-loving nations make concerted efforts to quarantine aggressors. He seemed to mean nothing more drastic than the breaking off of diplomatic relations, but the proposal created such national alarm that during ensuing months he was slow to develop a collective-security position. He quickly accepted Japanese apologies when the U.S. gunboat "Panay" was sunk on the Yangtze River in December 1937. Relations between the United States and Japan gradually worsened, but the rapid domination of Europe by Adolf Hitler of Germany was more threatening.

The outbreak of war. When World War II began in Europe in September 1939, Roosevelt called Congress into special session to revise the Neutrality Act to permit belligerents to buy arms on a "cash-and-carry" basis. With Hitler's aggressions and the fall of France in the spring and early summer of 1940, Roosevelt and Congress turned to defense preparations and "all aid short of war" to Great Britain. Roosevelt even gave Great Britain 50 over-age destroyers in exchange for eight Western Hemisphere bases. Isolationists, fearing U.S. involvement in the war, debated hotly with those who felt the national self-interest demanded aid to Britain.

The third and fourth terms. In the 1940 presidential campaign the Republicans nominated Wendell L. Willkie, who agreed with Roosevelt's foreign policy. Both candidates pledged to keep the nation out of foreign war; but isolationists tended to support Willkie, while those favouring strong measures against Hitler swung toward Roosevelt. By a closer margin than before—27,244,000 to 22,305,000 popular votes and 449 electoral votes to 82—Roosevelt was elected to an unprecedented third term.

Reelection in 1940

Through 1941 the nation moved gradually closer toward actual belligerency with Germany. After a bitter debate in Congress, Roosevelt in March 1941 obtained the Lend-Lease Act, enabling the United States to finance aid to Great Britain and its allies. Preventing submarines from sinking goods en route to Europe gradually involved more drastic protection by the U.S. Navy; in the fall Roosevelt authorized the navy to "shoot on sight" at German submarines. Meanwhile, in August, on a battleship off Newfoundland, Roosevelt met with Prime Minister Winston Churchill of Great Britain and signed a joint press release proclaiming an Atlantic Charter to provide national self-determination, greater economic opportunities, freedom from fear and want, freedom of the seas, and disarmament.

U.S. entry into the war. Yet it was in the Pacific that war came to the United States. Japan, bound in a treaty of alliance with Germany and Italy, the so-called Axis, extended its empire in East Asia. Roosevelt, viewing these moves as part of Axis world aggression, began to deny Japan supplies essential to its war making. Throughout 1941 the United States negotiated with Japan, but proposals by each side were unsatisfactory. Roosevelt did not want war with Japan in the fall of 1941, but he miscalculated in thinking the Japanese were bluffing. By the end of November he knew that Japanese fleet units and transports were at sea and that war was imminent; an attack in Southeast Asia and perhaps on the Philippines seemed likely. To Roosevelt's angered surprise, the Japanese, on December 7, 1941, struck Pearl Harbor, Hawaii. On December 8, at Roosevelt's request, Congress voted a war resolution within four hours; on December 11, Germany and Italy declared war on the United States.

Roosevelt made concessions to the conservatives in

Congress in order to obtain support in prosecuting the war. Several New Deal agencies were abolished. At a press conference Roosevelt asserted that "Dr. Win the War" had replaced "Dr. New Deal" but that this was to be only for the duration of the war. Roosevelt also fought resourcefully although not always successfully against inflationary pressures.

One of the immediate problems after Pearl Harbor was to build up massive production for war. Roosevelt had begun experimenting in 1939 with various defense agencies to mobilize the economy. Eventually, a workable organization had evolved. At the time of Pearl Harbor, U.S. war production was already nearly as great as that of Germany and Japan combined; by 1944 it was double the total of all Axis nations.

Relations with allies. During the war, Roosevelt concentrated upon problems of strategy, negotiations with the nation's allies, and the planning of the peace. From the outset, he took the lead in establishing a grand alliance among all countries fighting the Axis.

Wartime leader

Roosevelt met with Churchill in a number of wartime conferences at which differences were settled amicably. Debate at the earlier conferences centred upon the question of a landing in France, which the British succeeded in postponing repeatedly; the great Normandy invasion was finally launched in June 1944. Meanwhile, the United States had followed the British lead in invading North Africa in November 1942, Sicily in July 1943, and Italy in September 1943. At one of the most significant of the meetings, at Casablanca, Morocco, in January 1943, Roosevelt, after previous consultation with Churchill, proclaimed the doctrine of unconditional surrender of the Axis. He seemed to want to avoid the sort of differences of opinion among the Allies and misunderstanding by the Germans that had made trouble at the time of the 1918 Armistice. There is no tangible evidence that the doctrine in any way lengthened the war.

Relations with the Soviet Union posed a difficult problem for Roosevelt. Throughout the war the Soviet Union accepted large quantities of lend-lease supplies but seldom divulged its military plans or acted in coordination with its Western Allies. Roosevelt, feeling that the maintenance of peace after the war depended upon friendly relations with the Soviet Union, hoped to win Joseph Stalin's confidence. Roosevelt seemed to get along well with Stalin when he and Churchill first met with the Soviet leader at Tehrān, Iran, in November 1943. In their optimism, Roosevelt and Churchill seemed not to see realistically that the sort of peace being foreshadowed at Tehrān would leave the Soviet Union dominant in Europe.

Meanwhile, the Axis had been suffering serious defeats in both Europe and the Pacific. By February 1945, when the Big Three met again at Yalta in the Crimea, the war seemed almost over in Europe. As for Japan, the United States expected a last-ditch defense that might require another 18 months or more of fighting. Work in developing an atomic bomb was well advanced, but its power was expected to be only a fraction of what it actually turned out to be. Consequently, Roosevelt and his military advisers were eager to obtain Soviet aid in Asia; and, in return for Stalin's promise to enter the war against Japan, Roosevelt and Churchill offered concessions in the Far East. As for eastern Europe, earlier decisions were ratified, and plans were made for the establishment of democratic governments. Had the arrangements for eastern Europe been followed by Stalin in the manner expected by Roosevelt and Churchill, there would have been little room for criticism. But the understandings were not precise enough, and they received different interpretations in the Soviet Union. By mid-March false Soviet accusations against the United States led Roosevelt to send a sharp telegram to Stalin.

Declining health and death. Roosevelt hoped that the establishment of an effective international organization, the United Nations, could maintain the peace in years to come. He planned to attend a conference of 50 nations at San Francisco, opening April 25, 1945, to draft a United Nations charter. But, since January 1944, his health had been declining. His political opponents had tried to make much of this during the campaign of 1944, when he ran for a fourth term against Gov. Thomas E. Dewey of New York. A final burst of vigour on Roosevelt's part, however, seemed to refute the rumours. Roosevelt won by 25,602,000 to 22,006,000 in popular votes and 432 to 99 in electoral votes. But his address to Congress after he returned from Yalta had to be delivered sitting down. He went to Warm Springs for a rest, and there, on April 12, 1945, he died of a massive cerebral hemorrhage.

Reelection in 1944

EVALUATION

During Roosevelt's years as president, he had relatively little time for personal life. He continued his interest in his lands at Hyde Park. His zest for sailing and his enjoyment in collecting stamps and naval books and prints continued unabated. His tight schedule and the incessant publicity imposed upon him limited the time he could give to his wife, who became an important figure in her own right, and to his five children: Anna Eleanor, James, Elliott, Franklin D., Jr., and John A. As a public figure he was, at the same time, one of the most loved and most hated men in U.S. history. Opponents ascribed to him shallowness, incompetence, trickiness, and dictatorial ambitions. His supporters hailed him as the saviour of his nation's economy and the defender of democracy not only in the United States but throughout the world as well. It was generally conceded that as a political leader he was unexcelled in winning and holding popular support and in retaining, in his administration, leaders of diverse views. Many experts have expressed the opinion that despite occasional confusion and overlapping authority, his administration was unusually effective. He brought even more than this to the office: in 1932 he stated what remained his view through peace and war, "The Presidency . . . is pre-eminently a place of moral leadership." (F.Fr.)

BIBLIOGRAPHY. SAMUEL I. ROSENMAN (ed.), *The Public Papers and Addresses of Franklin D. Roosevelt,* 13 vol. (1938–50, reprinted 1969), contains official statements. RUSSELL D. BUHITE and DAVID W. LEVY (eds.), *FDR's Fireside Chats* (1992), offers transcripts of Roosevelt's radio broadcasts.

JAMES MACGREGOR BURNS, *Roosevelt,* 2 vol. (1956–70, reissued 1989); and FRANK FREIDEL, *Franklin D. Roosevelt* (1952–), are among the most detailed biographies. ARTHUR M. SCHLESINGER, JR., *The Age of Roosevelt,* 3 vol. (1957–60), is a brilliant survey. Other biographical studies include KENNETH S. DAVIS, *FDR* (1972–); JOSEPH ALSOP and ROLAND GELATT, *FDR, 1882–1945* (1982), highlighted by many photographs; NATHAN MILLER, *FDR* (1983); OTIS L. GRAHAM, JR., and MEGHAN ROBINSON WANDER (eds.), *Franklin D. Roosevelt: His Life and Times* (1985), an alphabetized collection of biographical and topical articles; TED MORGAN, *FDR* (1985); GEOFFREY C. WARD, *Before the Trumpet* (1985), and *A First-Class Temperament* (1989); and FRANK FREIDEL, *Franklin D. Roosevelt* (1990), which focuses on Roosevelt the president. RICHARD THAYER GOLDBERG, *The Making of Franklin D. Roosevelt: Triumph Over Disability* (1981), examines the effects of poliomyelitis on Roosevelt, the man and the politician. HUGH GREGORY GALLAGHER, *FDR's Splendid Deception,* rev. ed. (1994), is also of interest on the topic.

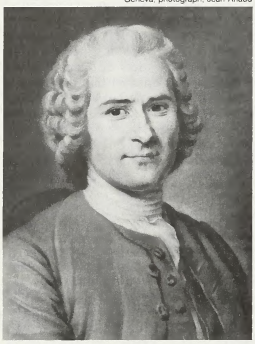
The New Deal is explored in ALBERT U. ROMASCO, *The Politics of Recovery* (1983); and in two works by WILLIAM E. LEUCHTENBURG, *Franklin D. Roosevelt and the New Deal, 1932–1940* (1963), and *The FDR Years: On Roosevelt and His Legacy* (1995), while his *The Supreme Court Reborn* (1995) deals with Roosevelt's plan to "pack" the court. Foreign relations are discussed in ROBERT DALLEK, *Franklin D. Roosevelt and American Foreign Policy* (1979); WAYNE S. COLE, *Roosevelt & the Isolationists, 1932–45* (1983); FREDERICK W. MARKS III, *Wind Over Sand* (1988); WARREN F. KIMBALL, *The Juggler: Franklin Roosevelt as Wartime Statesman* (1991); and JOHN LAMBERTON HARPER, *American Visions of Europe* (1994), which shows how Roosevelt helped shape American foreign policy. WALDO HEINRICHS, *Threshold of War: Franklin D. Roosevelt and American Entry into World War II* (1988), treats the events of 1941. The relationship between Roosevelt and Winston Churchill is the focus of KEITH SAINSBURY, *Churchill and Roosevelt at War* (1994). CORNELIS A. VAN MINNEN and JOHN F. SEARS (eds.), *FDR and His Contemporaries: Foreign Perceptions of an American President* (1992), is a collection of essays.

Among numerous studies dealing with Roosevelt's wife, Eleanor Roosevelt, are BLANCHE WIESEN COOK, *Eleanor Roosevelt* (1992–); DORIS KEARNS GOODWIN, *No Ordinary Time: Franklin and Eleanor Roosevelt* (1994); and ALLIDA M. BLACK, *Casting Her Own Shadow: Eleanor Roosevelt and the Shaping of Postwar Liberalism* (1996). (F.Fr./Ed.)

Jean-Jacques Rousseau

Jean-Jacques Rousseau, the least academic of modern philosophers, was in many ways the most influential. His thought marked the end of the Age of Reason and the birth of Romanticism. He propelled political and ethical thinking into new channels. His reforms revolutionized taste, first in music, then in the other arts. He had a profound impact on people's way of life; he taught parents to take a new interest in their children and to educate them differently; he furthered the expression of emotion rather than polite restraint in friendship and love. He introduced the cult of religious sentiment among people who had discarded religious dogma. He opened men's eyes to the beauties of nature, and he made liberty an object of almost universal aspiration.

By courtesy of the Musee d'Art et d'Histoire, Geneva, photograph, Jean Arlaud



Rousseau, drawing in pastels by Maurice-Quentin de La Tour, 1753. In the Musée d'Art et d'Histoire, Geneva.

Formative years. Rousseau was born in Geneva—the city of Calvin—on June 28, 1712. His mother died in childbirth and he was brought up by his father, who taught him to believe that the city of his birth was a republic as splendid as Sparta or ancient Rome. Rousseau senior had an equally glorious image of his own importance; after marrying above his modest station as a watchmaker, he got into trouble with the civil authorities by brandishing the sword that his upper-class pretentions prompted him to wear, and he had to leave Geneva to avoid imprisonment. Rousseau, the son, then lived for six years as a poor relation in his mother's family, patronized and humiliated, until he, too, at the age of 16, fled from Geneva to live the life of an adventurer and a Roman Catholic convert in the kingdoms of Sardinia and France.

Mme de Warens

Rousseau was fortunate in finding in the province of Savoy a benefactress named the Baronne de Warens, who provided him with a refuge in her home and employed him as her steward. She also furthered his education to such a degree that the boy who had arrived on her doorstep as a stammering apprentice who had never been to school developed into a philosopher, a man of letters, and a musician.

Mme de Warens, who thus transformed the adventurer into a philosopher, was herself an adventuress—a Swiss convert to Catholicism who had stripped her husband of his money before fleeing to Savoy with the gardener's son to set herself up as a Catholic missionary specializing in the conversion of young male Protestants. Her morals distressed Rousseau, even when he became her lover. But she was a woman of taste, intelligence, and energy, who brought out in Rousseau just the talents that were needed to conquer Paris at a time when Voltaire had made radical ideas fashionable.

Rousseau reached Paris when he was 30 and was lucky enough to meet another young man from the provinces seeking literary fame in the capital, Denis Diderot. The two soon became immensely successful as the centre of a group of intellectuals—or "Philosophes"—who gathered round the great French *Encyclopédie,* of which Diderot was appointed editor. The *Encyclopédie* was an important organ of radical and anticlerical opinion, and its contributors were as much reforming and even iconoclastic pamphleteers as they were philosophers. Rousseau, the most original of them all in his thinking and the most forceful and eloquent in his style of writing, was soon the most conspicuous. He wrote music as well as prose, and one of his operas, *Le Devin du village* (1752; *The Cunning-Man*), attracted so much admiration from the king and the court that he might have enjoyed an easy life as a fashionable composer, but something in his Calvinist blood rejected this type of worldly glory. Indeed, at the age of 37 Rousseau had what he called an "illumination" while walking to Vincennes to visit Diderot, who had been imprisoned there because of his irreligious writings. In the *Confessions,* which he wrote late in life, Rousseau says that it came to him then in a "terrible flash" that modern progress had corrupted instead of improved men. He went on to write his first important work, a prize essay for the Academy of Dijon entitled *Discours sur les sciences et les arts* (1750; *A Discourse on the Sciences and the Arts*), in which he argues that the history of man's life on earth has been a history of decay.

This *Discourse* is by no means Rousseau's best piece of writing, but its central theme was to inform almost everything else he wrote. Throughout his life he kept returning to the thought that man is good by nature but has been corrupted by society and civilization. He did not mean to suggest that society and civilization were inherently bad but rather that both had taken a wrong direction and become more harmful as they had become more sophisticated. This idea in itself was not unfamiliar when Rousseau published his *Discourse on the Sciences and the Arts.* Many Roman Catholic writers deplored the direction that European culture had taken since the Middle Ages. They shared the hostility toward progress that Rousseau had expressed. What they did not share was his belief that man was naturally good. It was, however, just this belief in man's natural goodness that Rousseau made the cornerstone of his argument.

The natural goodness of man

Rousseau may well have received the inspiration for this belief from Mme de Warens; for although that unusual woman had become a communicant of the Roman Catholic Church, she retained—and transmitted to Rousseau—much of the sentimental optimism about human purity that she had herself absorbed as a child from the mystical Protestant Pietists who were her teachers in the canton of Bern. At all events, the idea of man's natural goodness, as Rousseau developed it, set him apart from both conservatives and radicals. Even so, for several years after the publication of his first *Discourse,* he remained a close collaborator in Diderot's essentially progressive enterprise, the *Encyclopédie,* and an active contributor to its pages. His speciality there was music, and it was in this sphere that he first established his influence as reformer.

Controversy with Rameau. The arrival of an Italian opera company in Paris in 1752 to perform works of *opera buffa* by Pergolesi, Scarlatti, Vinci, Leo, and other such composers suddenly divided the French music-loving public into two excited camps, supporters of the new Italian opera and supporters of the traditional French opera. The

Italian vs. French music

Philosophes of the *Encyclopédie*—d'Alembert, Diderot, and d'Holbach among them—entered the fray as champions of Italian music, but Rousseau, who had arranged for the publication of Pergolesi's music in Paris and who knew more about the subject than most Frenchmen after the months he had spent visiting the opera houses of Venice during his time as secretary to the French ambassador to the doge in 1743–44, emerged as the most forceful and effective combatant. He was the only one to direct his fire squarely at the leading living exponent of French operatic music, Jean-Philippe Rameau.

Rousseau and Rameau must at that time have seemed unevenly matched in a controversy about music. Rameau, already in his 70th year, was not only a prolific and successful composer but was also, as the author of the celebrated *Traité de l'harmonie* (1722; *Treatise on Harmony*) and other technical works, Europe's leading musicologist. Rousseau, by contrast, was 30 years younger, a newcomer to music, with no professional training and only one successful opera to his credit. His scheme for a new notation for music had been rejected by the Academy of Sciences, and most of his musical entries for Diderot's *Encyclopédie* were as yet unpublished. Yet the dispute was not only musical but also philosophical, and Rameau was confronted with a more formidable adversary than he had realized. Rousseau built his case for the superiority of Italian music over French on the principle that melody must have priority over harmony, whereas Rameau based his on the assertion that harmony must have priority over melody. By pleading for melody, Rousseau introduced what later came to be recognized as a characteristic idea of Romanticism, namely, that in art the free expression of the creative spirit is more important than strict adhesion to formal rules and traditional procedures. By pleading for harmony, Rameau reaffirmed the first principle of French Classicism, namely, that conformity to rationally intelligible rules is a necessary condition of art, the aim of which is to impose order on the chaos of human experience.

In music, Rousseau was a liberator. He argued for freedom in music, and he pointed to the Italian composers as models to be followed. In doing so he had more success than Rameau; he changed people's attitudes. Gluck, who succeeded Rameau as the most important operatic composer in France, acknowledged his debt to Rousseau's teaching, and Mozart based the text for his one-act operetta *Bastien und Bastienne* on Rousseau's *Devin du village.* European music had taken a new direction. But Rousseau himself composed no more operas. Despite the success of *Le Devin du village,* or rather because of its success, Rousseau felt that, as a moralist who had decided to make a break with worldly values, he could not allow himself to go on working for the theatre. He decided to devote his energies henceforth to literature and philosophy.

Major works of political philosophy. As part of what Rousseau called his "reform," or improvement of his own character, he began to look back at some of the austere principles that he had learned as a child in the Calvinist republic of Geneva. Indeed he decided to return to that city, repudiate his Catholicism, and seek readmission to the Protestant church. He had in the meantime acquired a mistress, an illiterate laundry maid named Thérèse Levasseur. To the surprise of his friends, he took her with him to Geneva, presenting her as a nurse. Although her presence caused some murmurings, Rousseau was readmitted easily to the Calvinist communion, his literary fame having made him very welcome to a city that prided itself as much on its culture as on its morals.

Rousseau had by this time completed a second *Discourse* in response to a question set by the Academy of Dijon: "What is the origin of the inequality among men and is it justified by natural law?" In response to this challenge he produced a masterpiece of speculative anthropology. The argument follows on that of his first *Discourse* by developing the proposition that natural man is good and then tracing the successive stages by which man has descended from primitive innocence to corrupt sophistication.

Rousseau begins his *Discours sur l'origine de l'inegalité* (1755; *Discourse on the Origin of Inequality*) by distinguishing two kinds of inequality, natural and artificial, the first arising from differences in strength, intelligence, and so forth, the second from the conventions that govern societies. It is the inequalities of the latter sort that he sets out to explain. Adopting what he thought the properly "scientific" method of investigating origins, he attempts to reconstruct the earliest phases of man's experience of life on earth. He suggests that original man was not a social being but entirely solitary, and to this extent he agrees with Hobbes's account of the state of nature. But in contrast to the English pessimist's view that the life of man in such a condition must have been "poor, nasty, brutish and short," Rousseau claims that original man, while admittedly solitary, was healthy, happy, good, and free. The vices of men, he argues, date from the time when men formed societies.

Argument of the *Discourse on the Origin of Inequality*

Rousseau thus exonerates nature and blames society for the emergence of vices. He says that passions that generate vices hardly exist in the state of nature but begin to develop as soon as men form societies. Rousseau goes on to suggest that societies started when men built their first huts, a development that facilitated cohabitation of males and females; this in turn produced the habit of living as a family and associating with neighbours. This "nascent society," as Rousseau calls it, was good while it lasted; it was indeed the "golden age" of human history. Only it did not endure. With the tender passion of love there was also born the destructive passion of jealousy. Neighbours started to compare their abilities and achievements with one another, and this "marked the first step towards inequality and at the same time towards vice." Men started to demand consideration and respect; their innocent self-love turned into culpable pride, as each man wanted to be better than everyone else.

The introduction of property marked a further step toward inequality since it made it necessary for men to institute law and government in order to protect property. Rousseau laments the "fatal" concept of property in one of his more eloquent passages, describing the "horrors" that have resulted from men's departure from a condition in which the earth belonged to no one. These passages in his second *Discourse* excited later revolutionaries such as Marx and Lenin, but Rousseau himself did not think that the past could be undone in any way; there was no point in men dreaming of a return to the golden age.

Civil society, as Rousseau describes it, comes into being to serve two purposes: to provide peace for everyone and to ensure the right to property for anyone lucky enough to have possessions. It is thus of some advantage to everyone, but mostly to the advantage of the rich, since it transforms their de facto ownership into rightful ownership and keeps the poor dispossessed. It is a somewhat fraudulent social contract that introduces government since the poor get so much less out of it than do the rich. Even so, the rich are no happier in civil society than are the poor because social man is never satisfied. Society leads men to hate one another to the extent that their interests conflict, and the best they are able to do is to hide their hostility behind a mask of courtesy. Thus Rousseau regards the inequality between men not as a separate problem but as one of the features of the long process by which men become alienated from nature and from innocence.

In the dedication Rousseau wrote for the *Discourse,* in order to present it to the republic of Geneva, he nevertheless praises that city-state for having achieved the ideal balance between "the equality which nature established among men and the inequality which they have instituted among themselves." The arrangement he discerned in Geneva was one in which the best men were chosen by the citizens and put in the highest positions of authority. Like Plato, Rousseau always believed that a just society was one in which everyone was in his right place. And having written the *Discourse* to explain how men had lost their liberty in the past, he went on to write another book, *Du Contrat social* (1762; *The Social Contract*), to suggest how they might recover their liberty in the future. Again Geneva was the model; not Geneva as it had become in 1754 when Rousseau returned there to recover his rights as a citizen, but Geneva as it had once been; *i.e.,* Geneva as Calvin had designed it.

Geneva as a model for a just society

The Social Contract begins with the sensational opening sentence: "Man was born free, but he is everywhere in chains," and proceeds to argue that men need not be in chains. If a civil society, or state, could be based on a genuine social contract, as opposed to the fraudulent social contract depicted in the *Discourse on the Origin of Inequality,* men would receive in exchange for their independence a better kind of freedom, namely true political, or republican, liberty. Such liberty is to be found in obedience to a self-imposed law.

Rousseau's definition of political liberty raises an obvious problem. For while it can be readily agreed that an individual is free if he obeys only rules he prescribes for himself, this is so because an individual is a person with a single will. A society, by contrast, is a set of persons with a set of individual wills, and conflict between separate wills is a fact of universal experience. Rousseau's response to the problem is to define his civil society as an artificial person united by a general will, or *volonté générale.* The social contract that brings society into being is a pledge, and the society remains in being as a pledged group. Rousseau's republic is a creation of the general will—of a will that never falters in each and every member to further the public, common, or national interest—even though it may conflict at times with personal interest.

The concept of the general will

Rousseau sounds very much like Hobbes when he says that under the pact by which men enter civil society everyone totally alienates himself and all his rights to the whole community. Rousseau, however, represents this act as a form of exchange of rights whereby men give up natural rights in return for civil rights. The bargain is a good one because what men surrender are rights of dubious value, whose realization depends solely on an individual man's own might, and what they obtain in return are rights that are both legitimate and enforced by the collective force of the community.

There is no more haunting paragraph in *The Social Contract* than that in which Rousseau speaks of "forcing a man to be free." But it would be wrong to interpret these words in the manner of those critics who see Rousseau as a prophet of modern totalitarianism. He does not claim that a whole society can be forced to be free but only that an occasional individual, who is enslaved by his passions to the extent of disobeying the law, can be restored by force to obedience to the voice of the general will that exists inside of him. The man who is coerced by society for a breach of the law is, in Rousseau's view, being brought back to an awareness of his own true interests.

For Rousseau there is a radical dichotomy between true law and actual law. Actual law, which he describes in the *Discourse on the Origin of Inequality,* simply protects the status quo. True law, as described in *The Social Contract,* is just law, and what ensures its being just is that it is made by the people in its collective capacity as sovereign and obeyed by the same people in their individual capacities as subjects. Rousseau is confident that such laws could not be unjust because it is inconceivable that any people would make unjust laws for itself.

Rousseau is, however, troubled by the fact that the majority of a people does not necessarily represent its most intelligent citizens. Indeed, he agrees with Plato that most people are stupid. Thus the general will, while always morally sound, is sometimes mistaken. Hence Rousseau suggests the people need a lawgiver—a great mind like Solon or Lycurgus or Calvin—to draw up a constitution and system of laws. He even suggests that such lawgivers need to claim divine inspiration in order to persuade the dim-witted multitude to accept and endorse the laws it is offered.

Comparison with Machiavelli

This suggestion echoes a similar proposal by Machiavelli, a political theorist Rousseau greatly admired and whose love of republican government he shared. An even more conspicuously Machiavellian influence can be discerned in Rousseau's chapter on civil religion, where he argues that Christianity, despite its truth, is useless as a republican religion on the grounds that it is directed to the unseen world and does nothing to teach citizens the virtues that are needed in the service of the state, namely, courage, virility, and patriotism. Rousseau does not go so far as Machiavelli in proposing a revival of pagan cults, but he does propose a civil religion with minimal theological content designed to fortify and not impede (as Christianity impedes) the cultivation of martial virtues. It is understandable that the authorities of Geneva, profoundly convinced that the national church of their little republic was at the same time a truly Christian church and a nursery of patriotism, reacted angrily against this chapter in Rousseau's *Social Contract.*

By the year 1762, however, when *The Social Contract* was published, Rousseau had given up any thought of settling in Geneva. After recovering his citizen's rights in 1754, he had returned to Paris and the company of his friends around the *Encyclopédie.* But he became increasingly ill at ease in such worldly society and began to quarrel with his fellow Philosophes. An article for the *Encyclopédie* on the subject of Geneva, written by d'Alembert at Voltaire's instigation, upset Rousseau partly by suggesting that the pastors of the city had lapsed from Calvinist severity into unitarian laxity and partly by proposing that a theatre should be erected there. Rousseau hastened into print with a defense of the Calvinist orthodoxy of the pastors and with an elaborate attack on the theatre as an institution that could only do harm to an innocent community such as Geneva.

Years of seclusion and exile. By the time his *Lettre à d'Alembert sur les spectacles* (1758; *Letter to Monsieur d'Alembert on the Theatre*) appeared in print, Rousseau had already left Paris to pursue a life closer to nature on the country estate of his friend Mme d'Épinay near Montmorency. When the hospitality of Mme d'Épinay proved to entail much the same social round as that of Paris, Rousseau retreated to a nearby cottage, called Montlouis, under the protection of the Maréchal de Luxembourg. But even this highly placed friend could not save him in 1762 when his treatise on education, *Émile,* was published and scandalized the pious Jansenists of the French Parlements even as *The Social Contract* scandalized the Calvinists of Geneva. In Paris, as in Geneva, they ordered the book to be burned and the author arrested; all the Maréchal de Luxembourg could do was to provide a carriage for Rousseau to escape from France. Rousseau spent the rest of his life as a fugitive moving from one refuge to another.

The years at Montmorency had been the most productive of his literary career; besides *The Social Contract* and *Émile, Julie: ou, la nouvelle Héloïse* (1761; *Julie: or, The New Eloise*) came out within 12 months, all three works of seminal importance. *The New Eloise,* being a novel, escaped the censorship to which the other two works were subject; indeed of all his books it proved to be the most widely read and the most universally praised in his lifetime. It develops the Romanticism that had already informed his writings on music and perhaps did more than any other single work of literature to influence the spirit of its age. It made the author at least as many friends among the reading public—and especially among educated women—as *The Social Contract* and *Émile* made enemies among magistrates and priests. If it did not exempt him from persecution, at least it ensured that his persecution was observed, and admiring femmes du monde intervened from time to time to help him so that Rousseau was never, unlike Voltaire and Diderot, actually imprisoned.

The New Eloise

The theme of *The New Eloise* provides a striking contrast to that of *The Social Contract.* It is about people finding happiness in domestic as distinct from public life, in the family as opposed to the state. The central character, Saint-Preux, is a middle-class preceptor who falls in love with his upper-class pupil, Julie. She returns his love and yields to his advances, but the difference between their classes makes marriage between them impossible. Baron d'Étange, Julie's father, has indeed promised her to a fellow nobleman named Wolmar. As a dutiful daughter, Julie marries Wolmar and Saint-Preux goes off on a voyage around the world with an English aristocrat, Bomston, from whom he acquires a certain stoicism. Julie succeeds in forgetting her feelings for Saint-Preux and finds happiness as wife, mother, and chatelaine. Some six years later Saint-Preux returns from his travels and is engaged as tutor to the Wolmar children. All live together in harmony,

and there are only faint echoes of the old affair between Saint-Preux and Julie. The little community, dominated by Julie, illustrates one of Rousseau's political principles: that while men should rule the world in public life, women should rule men in private life. At the end of *The New Eloise,* when Julie has made herself ill in an attempt to rescue one of her children from drowning, she comes face-to-face with a truth about herself: that her love for Saint-Preux has never died.

The novel was clearly inspired by Rousseau's own curious relationship—at once passionate and platonic—with Sophie d'Houdetot, a noblewoman who lived near him at Montmorency. He himself asserted in the *Confessions* (1781–88) that he was led to write the book by "a desire for loving, which I had never been able to satisfy and by which I felt myself devoured." Saint-Preux's experience of love forbidden by the laws of class reflects Rousseau's own experience; and yet it cannot be said that *The New Eloise* is an attack on those laws, which seem, on the contrary, to be given the status almost of laws of nature. The members of the Wolmar household are depicted as finding happiness in living according to an aristocratic ideal. They appreciate the routines of country life and enjoy the beauties of the Swiss and Savoyard Alps. But despite such an endorsement of the social order, the novel was revolutionary; its very free expression of emotions and its extreme sensibility deeply moved its large readership and profoundly influenced literary developments.

Émile is a book that seems to appeal alternately to the republican ethic of *The Social Contract* and the aristocratic ethic of *The New Eloise.* It is also halfway between a novel and a didactic essay. Described by the author as a treatise on education, it is not about schooling but about the upbringing of a rich man's son by a tutor who is given unlimited authority over him. At the same time the book sets out to explore the possibilities of an education for republican citizenship. The basic argument of the book, as Rousseau himself expressed it, is that vice and error, which are alien to a child's original nature, are introduced by external agencies, so that the work of a tutor must always be directed to counteracting those forces by manipulating pressures that will work with nature and not against it. Rousseau devotes many pages to explaining the methods the tutor must use. These methods involve a noticeable measure of deceit, and although corporal punishment is forbidden, mental cruelty is not.

Whereas *The Social Contract* is concerned with the problems of achieving freedom, *Émile* is concerned with achieving happiness and wisdom. In this different context religion plays a different role. Instead of a civil religion, Rousseau here outlines a personal religion, which proves to be a kind of simplified Christianity, involving neither revelation nor the familiar dogmas of the church. In the guise of *La Profession de foi du vicaire savoyard* (1765; *The Profession of Faith of a Savoyard Vicar*) Rousseau sets out what may fairly be regarded as his own religious views, since that book confirms what he says on the subject in his private correspondence. Rousseau could never entertain doubts about God's existence or about the immortality of the soul. He felt, moreover, a strong emotional drive toward the worship of God, whose presence he felt most forcefully in nature, especially in mountains and forests untouched by the hand of man. He also attached great importance to conscience, the "divine voice of the soul in man," opposing this both to the bloodless categories of rationalistic ethics and to the cold tablets of biblical authority.

A minimal creed

This minimal creed put Rousseau at odds with the orthodox adherents of the churches and with the openly atheistic Philosophes of Paris, so that despite the enthusiasm that some of his writings, and especially *The New Eloise,* excited in the reading public, he felt himself increasingly isolated, tormented, and pursued. After he had been expelled from France, he was chased from canton to canton in Switzerland. He reacted to the suppression of *The Social Contract* in Geneva by indicting the regime of that city-state in a pamphlet entitled *Lettres écrites de la montagne* (1764; *Letters Written from the Mountain*). No longer, as in the *Discourse on the Origin of Inequality,* was Geneva depicted as a model republic but as one that had been taken over by "twenty-five despots"; the subjects of the king of England were said to be free men by comparison with the victims of Genevan tyranny.

It was in England that Rousseau found refuge after he had been banished from the canton of Bern. The Scottish philosopher David Hume took him there and secured the offer of a pension from King George III; but once in England, Rousseau became aware that certain British intellectuals were making fun of him, and he suspected Hume of participating in the mockery. Various symptoms of paranoia began to manifest themselves in Rousseau, and he returned to France incognito. Believing that Thérèse was the only person he could rely on, he finally married her in 1768, when he was 56 years old.

Paranoia and marriage

The last decade. In the remaining 10 years of his life Rousseau produced primarily autobiographical writings, mostly intended to justify himself against the accusations of his adversaries. The most important was his *Confessions,* modeled on the work of the same title by St. Augustine and achieving something of the same classic status. He also wrote *Rousseau juge de Jean-Jacques* (1780; "Rousseau, Judge of Jean-Jacques") to reply to specific charges by his enemies and *Les Rêveries du promeneur solitaire* (1782; *Reveries of the Solitary Walker*), one of the most moving of his books, in which the intense passion of his earlier writings gives way to a gentle lyricism and serenity. And indeed, Rousseau does seem to have recovered his peace of mind in his last years, when he was once again afforded refuge on the estates of great French noblemen, first the Prince de Conti and then the Marquis de Girardin, in whose park at Ermenonville he died on July 2, 1778.

MAJOR WORKS

NOVELS: *Julie: ou, la nouvelle Héloïse* (1761; *Julie: or, The New Eloise,* trans. by Judith H. McDowell, 1968); *Émile: ou, de l'éducation* (1762; *Emile: or, On Education,* trans. by Allan Bloom, 1979).

AUTOBIOGRAPHICAL WORKS: *Rousseau juge de Jean-Jacques: dialogue* (1780); *Les Rêveries du promeneur solitaire* (1782; *The Reveries of the Solitary Walker,* trans. by Charles E. Butterworth, 1979); *Les Confessions* (1782–89; *The Confessions,* trans. by J.M. Cohen, 1953).

ESSAYS: *Discours qui a remporté le prix à l'Académie de Dijon en l'année 1750; sur cette question proposée par la même académie si le rétablissement des sciences et des arts a contribué à épurer les moeurs* (1750; "Discourse on the Sciences and Arts," trans. by Roger D. Masters and Judith R. Masters, in *The First and Second Discourses,* 1964); *Discours sur l'origine et les fondements de l'inégalité parmi les hommes* (1755; *Discourse on Inequality,* trans. by Maurice Cranston, 1984); *Du Contrat social* (1762; *The Social Contract,* trans. by Maurice Cranston, 1968); *Considérations sur le gouvernement de Pologne* (1782; *The Government of Poland,* trans. by Willmoore Kendall, 1972); *Lettres élémentaires sur la botanique* (1780; *Letters on the Elements of Botany,* trans. by Thomas Martyn, 1785).

LETTERS: *J.J. Rousseau, citoyen de Genève, à M. d'Alembert, sur son article Genève dans le septième volume de l'Encyclopédie, et particulièrement sur le projet d'établir un Théâtre de Comédie en cette ville* (1758; *Politics and the Arts: Letter to M. d'Alembert on the Theatre,* trans. by Allan Bloom, 1960); *Lettres écrites de la montagne* (1764).

COLLECTED WORKS: *Oeuvres complètes,* ed. by Bernard Gagnebin and Marcel Raymond (1959–), will eventually be the definitive collected edition. Four carefully annotated volumes have been published so far. *Oeuvres complètes,* ed. by Michel Launay, 3 vol. (1967–71), is the most comprehensive contemporary edition, but far from complete. In some earlier editions of Rousseau's collected works, published at the end of the 18th and the beginning of the 19th centuries, material can be found that has not been reprinted in the 20th-century collections.

The *Correspondance complète de Jean Jacques Rousseau: édition critique,* ed. by R.A. Leigh (1965–), of which 43 volumes have so far appeared, wholly supersedes the *Correspondance générale de J.-J. Rousseau,* 20 vol., ed. by Théophile Dufour and Pierre P. Plan (1924–34).

BIBLIOGRAPHY

Bibliography: JEAN SÉNELIER, *Bibliographie générale des oeuvres de Jean-Jacques Rousseau* (1950), is still the best available source. THÉOPHILE DUFOUR, *Recherches bibliographiques sur les oeuvres imprimées de J.J. Rousseau,* 2 vol. (1925, reprinted in 1 vol., 1971), is not entirely superseded by Sénelier's work. ALBERT SCHINZ, *État present des travaux sur J.-J. Rousseau* (1941, reprinted 1971), includes publications in languages other than

French. PETER GAY, *The Party of Humanity* (1963, reissued 1971), contains a critical bibliography in English of Rousseau and his contemporaries. SOCIÉTÉ JEAN-JACQUES ROUSSEAU, GENEVA, *Annales* (irregular), published since 1905, contains reviews of all important publications in several languages, concerning Rousseau. HERMINE DE SOUSSURE, *Rousseau et les manuscrits des Confessions* (1958), and *Étude sur le sort des manuscrits de J.-J. Rousseau* (1974), provide information on the whereabouts of Rousseau's manuscripts.

Biographies: JEAN GUÉHENNO, *Jean-Jacques Rousseau,* 2 vol. (1966; originally published in French, 1948–52; new ed. 1983), is still the most comprehensive biography. LESTER G. CROCKER, *Jean-Jacques Rousseau,* 2 vol. (1968–73), is a detailed but somewhat hostile biographical study; as is FREDERICK C. GREEN, *Jean-Jacques Rousseau: A Critical Study of His Life and Writings* (1955, reprinted 1970). BERNARDIN DE SAINT-PIERRE, *La Vie et les ouvrages de Jean-Jacques Rousseau,* ed. from the author's unfinished manuscript by MAURICE SOURIAU (1907), is the only biography by an author who knew Rousseau personally. LOUIS J. COURTOIS, *Chronologie critique de la vie et des oeuvres de Jean-Jacques Rousseau* (1924, reprinted 1973), sets out the events of Rousseau's life in chronological order; as does, on a smaller scale and in English, GEORGE R. HAVENS, *Jean-Jacques Rousseau* (1978). MAURICE CRANSTON, *Jean-Jacques: The Early Life and Work of Jean-Jacques Rousseau, 1712–1754* (1983), is based on original manuscript sources but covers only the first 42 years of Rousseau's life. WILLIAM H. BLANCHARD, *Rousseau and the Spirit of Revolt* (1967); and JACQUES BOREL, *Génie et folie de Jean-Jacques Rousseau* (1966), are both Freudian biographies; while RONALD GRIMSLEY, *Jean-Jacques Rousseau: A Study in Self-Awareness,* 2nd ed. (1969), discusses the psychological aspects of Rousseau's *Confessions* from a more philosophical perspective. DANIEL MORNET, *Rousseau, l'homme et l'oeuvre,* 5th ed. (1967), sets out to correct many popular misconceptions about Rousseau's life and work. GASPARD VALLETTE, *Jean-Jacques Rousseau, Genevois* (1911); and J.S. SPINK, *Jean-Jacques Rousseau et Genève* (1934), investigate Rousseau's origins in Geneva. RENÉ HUBERT, *Rousseau et l'Encyclopédie: essai sur la formation des idées politiques de Rousseau, 1742–1756* (1928), examines Rousseau's relations as a young man with the Philosophes of Paris. JULIEN TIERSOT, *J.-J. Rousseau,* 2nd ed. (1920, reprinted 1978), is one of the rare studies of Rousseau's career as a reformer of music. HENRI GUILLEMIN, *Un Homme, deux ombres: (Jean-Jacques, Julie, Sophie)* (1943), discusses Rousseau's relationships with women as reflected in his novels. ELIZABETH A. FOSTER, *Le Dernier Séjour de J.J. Rousseau à Paris* (1921), is an account of Rousseau's last years.

Philosophy: RONALD GRIMSLEY, *The Philosophy of Rousseau* (1973), provides a clear scholarly introduction to Rousseau's philosophical ideas. Other useful introductory commentaries are ERNEST HUNTER WRIGHT, *The Meaning of Rousseau* (1929, reissued 1963); and J.H. BROOME, *Rousseau: A Study of His Thought* (1963). ERNST CASSIRER, *The Question of Jean-Jacques Rousseau* (1954, reprinted 1963; originally published in German, 1932), is an influential study, written from a Kantian perspective; and CHARLES WILLIAM HENDEL, *Jean-Jacques Rousseau, Moralist,* 2 vol. (1934, reissued 1962), is a longer study reaching much the same conclusions. ROBERT DERATHÉ, *Le Rationalisme de Jean-Jacques Rousseau* (1948, reprinted 1979), which opened a new phase in Rousseau's scholarship, reaffirms Rousseau's place in the Cartesian tradition. PIERRE BURGELIN, *La Philosophie de l'existence de J.-J. Rousseau,* 2nd ed. (1973), places Rousseau between Pascal and Kierkegaard. BERNHARD GROETHUYSEN, *J.-J. Rousseau* (1949), demonstrates Rousseau's importance from the point of view of 20th-century philosophy. MARC F. PLATTNER, *Rousseau's State of Nature: An Interpretation of the "Discourse on Inequality"* (1979), is a scholarly though brief study of Rousseau's concepts.

Literature: LÉO LAUNAY and MICHEL LAUNAY, *Le Vocabulaire littéraire de Jean-Jacques Rousseau* (1979), provides a linguistic key to Rousseau's literary work. JEAN STAROBINSKI, *Jean-Jacques Rousseau: la transparence et l'obstacle,* new ed. (1971, reprinted 1976), is a seminal work by an academic psychologist turned literary critic. MARCEL RAYMOND, *Jean-Jacques Rousseau: la quête de soi et la rêverie* (1962), provides a subtle analysis of Rousseau's literary achievement. PHILIP E.J. ROBINSON, *Jean-Jacques Rousseau's Doctrine of the Arts* (1984), is a pioneering attempt to depict Rousseau's ideas on literature and the other arts as a coherent system. HENRI GOUHIER, *Rousseau et Voltaire: portraits dans deux miroirs* (1983), is an impartial appraisal of the two literary giants of the French Enlightenment. ALBERT SCHINZ, *La Pensée de Jean-Jacques Rousseau* (1929), is an important study of Rousseau's Romanticism.

Religion: The most substantial study of Rousseau's religious ideas is still PIERRE MAURICE MASSON, *La Religion de J.-J. Rousseau,* 3 vol. (1916, reprinted 1970). The best introduction to the subject in English is by RONALD GRIMSLEY, *Rousseau and the Religious Quest* (1968). PIERRE BURGELIN, *Jean-Jacques Rousseau et la religion de Genève* (1962), examines Rousseau's debt to Calvinism; and J.F. THOMAS, *Le Pélagianisme de J.-J. Rousseau* (1956), studies his links to Roman Catholic philosophy. ALBERT SCHINZ, *La Pensée religieuse de Rousseau et ses récents interprètes* (1927), relates Rousseau's theological views to those of his contemporaries. JEAN-JACQUES ROUSSEAU, *Religious Writings of Rousseau,* ed. by RONALD GRIMSLEY (1970), contains key passages in English translation.

Political and social theory: ROBERT DERATHÉ, *Jean-Jacques Rousseau et la science politique de son temps,* 2nd ed. (1970), interprets Rousseau's political ideas within the tradition of the natural law school. A similar view is taken by ALFRED COBBAN, *Rousseau and the Modern State,* 2nd ed. (1964). The suggestion that Rousseau must be seen as a forerunner of totalitarianism is put forward unambiguously by J.L. TALMON, *The Origins of Totalitarian Democracy* (1952, reissued 1970; U.S. title, *The Rise of Totalitarian Democracy*); and in a modified form by JUDITH N. SHKLAR, *Men and Citizens: A Study of Rousseau's Social Theory* (1969, reprinted 1985). JOHN W. CHAPMAN, *Rousseau—Totalitarian or Liberal?* (1956, reprinted 1968), considers arguments for and against Talmon's interpretation. A commentary that stays close to the text is ROGER D. MASTERS, *The Political Philosophy of Rousseau* (1968, reprinted 1976); and an equally exacting study is JOHN CHARVET, *The Social Problem in the Philosophy of Rousseau* (1974). JAMES MILLER, *Rousseau: Dreamer of Democracy* (1984), stresses the democratic elements in Rousseau's political thought; while DAVID CAMERON, *The Social Thought of Rousseau and Burke* (1973), draws attention to resemblances between Rousseau's political ideas and those of the Irish conservative. STEPHEN ELLENBURG, *Rousseau's Political Philosophy: An Interpretation from Within* (1976), studies different interpretations of Rousseau's views. MICHEL LAUNAY, *Jean-Jacques Rousseau, écrivain politique, 1712–1762* (1971), depicts Rousseau both as champion of the popular classes in Geneva and as a theorist of the left; while GALVANO DELLA VOLPE, *Rousseau and Marx* (1978; originally published in Italian, 4th ed., 1964), presents Rousseau as a prophet of Communism. RAYMOND POLIN, *La Politique de la solitude: essai sur la philosophie politique de Jean-Jacques Rousseau* (1971), considers Rousseau as a philosopher rather than an ideologue; and BRONISLAW BACZKO, *Rousseau, solitude et communauté* (1974; originally published in Polish, 1970), gives new grounds for regarding Rousseau as one of the greatest social thinkers of modernity. JOEL SCHWARTZ, *The Sexual Politics of Jean-Jacques Rousseau* (1984), analyzes Rousseau's views on the role of sexuality in social politics and morals.

Shorter writings on his political thought are in SIMON HARVEY *et al.* (eds.), *Reappraisals of Rousseau* (1980); MAURICE CRANSTON and RICHARD S. PETERS (eds.), *Hobbes and Rousseau* (1982); R.A. LEIGH (ed.), *Rousseau After Two Hundred Years* (1982); COMITÉ NATIONAL POUR LA COMMÉMORATION DE J.-J. ROUSSEAU, *Jean-Jacques Rousseau et son oeuvre: problèmes et recherches* (1964); MICHEL LAUNAY *et al., Jean-Jacques Rousseau et son temps: politique et littérature au XVIIIe siècle* (1969).

(M.C.)

Russia

Once the preeminent republic of the Union of Soviet Socialist Republics (U.S.S.R.), Russia (officially Russian Federation; Rossiyskaya Federatsiya) has been an independent country since the dissolution of the union in December 1991. Under the Soviet system it was called the Russian Soviet Federated Socialist Republic (R.S.F.S.R.). With an area of 6,592,800 square miles (17,075,400 square kilometres), Russia is the world's largest country, covering almost twice the territory of either the United States or China. It ranks sixth in the world in population, following China, India, the United States, Indonesia, and Brazil. The great majority of the people are Russians, but there also are some 70 smaller national groups living within its borders. Most of the population is concentrated in a great triangle in the western, or European, part of the country, although over the past three centuries—and particularly during the early and mid-20th century—there was a steady flow of people eastward to the Asiatic section commonly referred to as Siberia.

On its northern and eastern sides Russia is bounded by the Arctic and Pacific oceans, and it has small frontages in the northwest on the Baltic Sea at St. Petersburg and at the detached Russian *oblast* (province) of Kaliningrad. On the south it borders North Korea, China, Mongolia, and the former Soviet republics of Kazakstan, Azerbaijan, and Georgia. On the southwest and west it borders the former Soviet republics of Ukraine, Belarus, Latvia, and Estonia, as well as Finland and Norway; in addition, Kaliningrad (formerly a part of what was once East Prussia annexed in 1945) abuts Poland and Lithuania.

Extending nearly halfway around the Northern Hemisphere and covering much of eastern and northeastern Europe as well as the whole of northern Asia, Russia has a maximum east-west extent, along the Arctic Circle, of some 4,800 miles (7,700 kilometres) and a north-south width of 1,250 to 1,850 miles. There is an enormous variety of landforms and landscapes, which occur mainly in a series of broad latitudinal belts. Arctic deserts lie in the extreme north, giving way southward to the tundra and then to the forest zones, which cover about half of the country and give it much of its character. South of the forest zone lie the wooded steppe and steppe, beyond which are small sections of semidesert along the northern shore of the Caspian Sea. Much of the federation lies in latitudes where the winter cold is intense and where evaporation can barely keep pace with the accumulation of moisture, engendering abundant rivers, lakes, and swamps.

The capital of the Russian Federation is Moscow, which was also the capital of the R.S.F.S.R. and of the Soviet Union. The republic itself had been established immediately after the Russian Revolution of October (November, New Style) 1917 and became a union republic on Dec. 30 (Dec. 17, Old Style), 1922. Following the termination of the U.S.S.R. in 1991, Russia joined with other former Soviet republics in forming the Commonwealth of Independent States (CIS).

Historically, the territory of European Russia was the core of the expanding Russian state and suffered onslaughts ranging from that of the Mongol hordes in the 13th century to the Nazi invasion of World War II. This historical heritage, together with the country's vast area and natural wealth, which permitted the development of a large-scale industrial economy, gave Russia a unique place of leadership among the former Soviet republics. Its brooding landscapes and the complexities of the prerevolutionary society inspired the prose and music of such giants of world culture as Anton Chekhov, Aleksandr Pushkin, Leo Tolstoy, and Peter Ilich Tchaikovsky, while the October Revolution (of 1917) and the changes it brought were reflected in the works of such noted figures as the novelists Maksim Gorky, Mikhail Sholokhov, and Aleksandr Solzhenitsyn, the poet Vladimir Mayakovsky, and the composers Dimitry Shostakovich and Sergey Prokofiev.

For the geography and history of Russia's two largest cities, see the *Macropædia* articles MOSCOW and SAINT PETERSBURG. For the history of the Soviet Union as a whole, from the Revolution to 1991, see UNION OF SOVIET SOCIALIST REPUBLICS. For the geography and history of the other former Soviet republics, see BALKAN STATES (Moldova), BALTIC STATES (Estonia, Latvia, Lithuania), BELARUS, CENTRAL ASIA (Kazakstan, Kyrgyzstan, Tajikistan, Turkmenistan, Uzbekistan), TRANSCAUCASIA (Armenia, Azerbaijan, Georgia), and UKRAINE.

This article is divided into the following sections:

Physical and human geography 944
- The land 944
  - Relief
  - Drainage
  - Climate
  - Soils and plant and animal life
  - Settlement patterns
- The people 954
  - Ethnic and linguistic composition
  - Religion
  - Demography
- The economy 956
  - Resources
  - Industry
  - Agriculture, forestry, and fishing
  - Trade
  - Transportation
- Administration and social conditions 958
  - Government
  - Education
  - Social conditions
- Cultural life 960
  - The development of Russian culture
  - The arts
  - Cultural and educational institutions
  - Sports and recreation

History 965
- From the beginnings to *c.* 1700 965
  - Prehistory and the rise of the Rus
  - Kiev
  - The lands of Rus
  - The Mongol period
  - Rurikid Muscovy
  - Romanov Muscovy
- The 18th century 974
  - The reign of Peter I the Great (1689–1725)
  - Peter I's successors (1725–62)
  - The reign of Catherine II the Great (1762–96)
  - Education and social change in the 18th century
  - The reign of Paul I (1796–1801)
- Russia from 1801 to 1917 982
  - The reigns of Alexander I and Nicholas I
  - From Alexander II to Nicholas II
  - The last years of tsardom
- Soviet and post-Soviet Russia 995
  - Russia after the monarchy
  - The October (November) Revolution
  - The Civil War and War Communism (1918–21)
  - New economic policy (1921–28)
  - Russia under Stalin (1928–53)
  - Russia under Khrushchev (1953–64)
  - Russia under Brezhnev (1964–82)
  - Gorbachev: *perestroika* and *glasnost*
  - Collapse of the Soviet Union
  - Post-Soviet Russia

Bibliography 1000

PHYSICAL AND HUMAN GEOGRAPHY

The land

RELIEF

On the basis of geologic structure and relief, Russia can be divided into two main parts—western and eastern—roughly along the line of the Yenisey River. In the western section, occupying some two-fifths of the total area, lowland plains predominate over vast areas broken only by low hills and plateaus. In the eastern section the bulk of the terrain is mountainous, although there are some extensive lowlands. Given these topological factors, the republic may be subdivided into six main relief regions: the Kola-Karelian region, the Russian Plain, the Ural Mountains, the West Siberian Plain, the Central Siberian Plateau, and the mountains of the south and east.

Main relief regions

The Kola-Karelian region. The smallest of these regions, Kola-Karelia, lies in the northwestern part of European Russia between the Finnish border and the White Sea. Karelia is a low, ice-scraped plateau with a maximum height of 1,896 feet (578 metres), but for the most part it is below 650 feet; low ridges and knolls alternate with lake- and marsh-filled hollows. The Kola Peninsula is similar, but the small Khibiny mountain range reaches 3,907 feet. Mineral-rich ancient rocks lie at or near the surface in many places.

The Russian Plain. Western Russia makes up the largest part of one of the great lowland areas of the world, the Russian (or East European) Plain, which extends into Russia from the western border eastward for 1,000 miles to the Ural Mountains and from the Arctic Ocean more than 1,500 miles to the Caucasus and the Caspian Sea. About half this vast area is less than 650 feet above sea level, and the highest point (in the Valdai Hills northwest of Moscow) reaches only 1,125 feet. Nevertheless, the detailed topography is quite varied. North of the latitude of Moscow, features of lowland glacial deposition predominate, and morainic ridges, of which the most pronounced are the Valdai Hills and the Smolensk-Moscow Upland (1,030 feet), stand out above low, ill-drained hollows interspersed with lakes and marshes. South of Moscow there is a west-east alternation of rolling plateaus and extensive plains. In the west the Central Russian Upland, with a maximum height of 961 feet, separates the lowlands of the upper Dnieper River valley from those of the Oka and Don rivers, beyond which the Volga Hills rise gently to 1,230 feet before descending abruptly to the Volga River. Small river valleys are sharply incised into these uplands, whereas the major rivers cross the lowlands in broad, shallow floodplains. East of the Volga is the large Caspian Depression, the lowest parts of which are more than 90 feet below sea level. The Russian Plain also extends southward through the Azov-Caspian isthmus (in the North Caucasus economic region) to the foot of the Caucasus Mountains, the crest line of which forms the boundary between Russia and the Transcaucasian states of Georgia and Azerbaijan; just inside this border stands Mount Elbrus, which at 18,510 feet (5,642 metres) is the highest point in Russia. The large Kuban and Kuma plains of the North Caucasus are separated by the Stavropol Upland at heights of 1,000 to 2,000 feet.



Typical wooden buildings in a village in the Central Ural Mountains near Kungur.

© Wolfgang Kaehler

The Ural Mountains. Along the eastern edge of the Russian Plain a belt of low mountains and plateaus 1,150 to 1,500 feet high flanks the Ural Mountains proper. The north-south spine of the Urals extends about 1,300 miles from the Arctic coast to the Kazak border and is extended an additional 600 miles into the Arctic Ocean by Novaya Zemlya, the largest of the Arctic islands. Although they form the traditional boundary between Europe and Asia, the Urals are no great barrier to movement. The highest peak, Mount Narodnaya, reaches 6,217 feet, but the system is largely composed of a series of broken, parallel ridges with summits generally between 3,000 and 5,000 feet; there are several low passes through these, particularly in the central section between Perm and Yekaterinburg, which carry the main routes from Europe into Siberia. Mineral-rich rocks occur in many districts.

The West Siberian Plain. Russia's most extensive region, the West Siberian Plain, is the most striking single relief feature of the country and quite possibly of the world. Covering an area well in excess of one million square miles, one-seventh of Russia's territory, it stretches about 1,200 miles from the Urals to the Yenisey and 1,500 miles from the Arctic Ocean to the foothills of the Altai Mountains. Only in the extreme south do heights exceed 650 feet, and more than half the plain is less than 330 feet above sea level. Vast floodplains and some of the world's largest swamps are characteristic features, particularly of the plain's northern half. Slightly higher and drier territory is located south of latitude 55° N, and it is there that the bulk of the region's population is concentrated.

Floodplains and swamps

The Central Siberian Plateau. Occupying most of the area between the Yenisey and Lena rivers, the Central Siberian Plateau comprises a series of sharply dissected plateau surfaces at heights between 1,000 and 2,300 feet. Toward its northern edge the Putoran Mountains rise to 5,581 feet. On its southern side the plateau is bounded by the Eastern Sayan Mountains and the mountains of Baikal (Baikalia); on the north it descends to the North Siberian Lowland, an eastward extension of the West Siberian Plain. Farther north the Byrranga Mountains reach 3,760 feet on the Taymyr (Taimyr) Peninsula, which extends into the Arctic Ocean. On its eastern side the Central Siberian Plateau gives way to the low-lying Central Yakut Lowland.

The mountains of the south and east. The remaining territory of Russia, to the south and east, makes up about



A stand of birch trees and conifers in the taiga of the West Siberian Plain, near Nizhnevartovsk.
N. Gyngazov/© TRIP

one-fourth of the republic's total area and is dominated by a complex series of high mountain systems. Of varied geologic origin, these mountains may be considered as a single major relief region, which forms part of the mountain barrier that encloses Russia on its southern and eastern sides.

In the section to the west of Lake Baikal, the mountain barrier is relatively narrow. The Altai Mountains, with a maximum height of 14,783 feet, lie on Russia's borders with Kazakstan and Mongolia; they are succeeded eastward by the V-shaped system of the Western Sayan and Eastern Sayan mountains (10,240 feet and 9,990 feet, respectively), which enclose the high Tuva Basin. Subsidiary ranges extend northward, enclosing the Kuznetsk and Minusinsk basins.

The area around Lake Baikal is one of massive block faulting in which major faults separate high plateaus and mountain ranges from deep valleys and basins. The scale of relief in this area is indicated by the fact that the floor of the lake at its deepest is more than 3,800 feet below sea level, while the mountains rising from its western shore reach elevations of 8,400 feet above sea level, a vertical difference of some 12,200 feet.

East of Lake Baikal, mountain ranges fan out to occupy most of the territory between the Lena River and the Pacific coast. This section is conventionally divided into northeastern and southeastern Siberia along the line of the Stanovoy Mountains. With a maximum height of 7,913 feet, the Stanovoy run some 400 miles eastward to the Pacific coast and separate the Lena and Amur drainage systems, which flow to the Arctic and Pacific oceans, respectively. Branching northeastward from the eastern end of the Stanovoy, the Dzhugdzhur Range rises to 6,253 feet along the coast, and its line is continued toward the Chukchi Peninsula by the Kolyma Mountains. From this chain, major ranges branch off to the northwest, notably the Verkhoyansk Mountains, which rise to 7,838 feet immediately east of the Lena, and the Chersky Mountains, having a maximum height of 10,325 feet. North of this system the low-lying, swampy Kolyma Lowland fronts the Arctic Ocean, extending for some 460 miles to the Chersky Mountains.

A narrow lowland corridor from the Sea of Okhotsk to the Bering Sea separates these complex fold mountain systems from the Kamchatka-Kuril region, where the Koryak and Sredinny mountains rise to 8,405 feet and 11,880 feet, respectively, forming a northeast-southwest chain that extends along the Pacific-rimmed Kamchatka Peninsula. The peninsula contains numerous volcanic peaks, many of them still active, of which the tallest is Klyuchevskaya Volcano, at 15,584 feet the highest point in the whole of the Russian Far East; several others rise well above 10,000 feet. This volcanic zone, part of the great circum-Pacific ring of seismic activity, continues southeastward through the Kuril Islands chain and into Japan.

Volcanic peaks

Southeastern Siberia is a land of high mountain ranges and extensive lowland plains. The most prominent mountains are the Badzhalsky Mountains (8,661 feet), to the west of the lower Amur, and the Sikhote-Alin (6,814 feet), between the Amur-Ussuri lowlands and the Pacific.

Separated from the Siberian mainland by the Tatar Strait, only about 4 miles wide at its narrowest, is Sakhalin Island. Some 600 miles from north to south but only 25 to 95 miles across, Sakhalin comprises a lowland plain in the north and, in the south, the parallel Eastern and Western Sakhalin mountain ranges, which reach 5,279 feet and 4,347 feet, respectively.

DRAINAGE

Rivers. The vast lowland plains that dominate the Russian landscape carry some of the world's longest rivers. Five main drainage basins may be distinguished: the Arctic, Pacific, Baltic, Black Sea, and Caspian. Of these by far the most extensive is the Arctic basin, which lies mostly in Siberia but also includes the northern part of the Russian Plain. The greater part of this basin is drained by three gigantic rivers: the Ob (2,268 miles, which with its main tributary, the Irtysh, extends for a continuous 3,362 miles), the Yenisey (2,540 miles), and the Lena (2,734 miles). Their catchments cover a total area of 3,127,000 square miles in West Siberia, East Siberia, and the Far East north of the Stanovoy Mountains, and their combined discharge into the Arctic averages 1,750,000 cubic feet (49,554 cubic metres) per second. Smaller, but still impressive, rivers make up the remainder of the Arctic drainage: in the European section these include the Northern Dvina (with its tributaries the Vychegda and Sukhona) and the Pechora, and in Siberia the Indigirka and Kolyma. The Siberian


FEDERAL DISTRICTS OF THE RUSSIAN FEDERATION
St. Petersburg
NORTHWEST
Moscow
CENTRAL
Nizhny Novgorod
VOLGA
URALS
SIBERIA
FAR EAST
Rostov-na-Donu
Yekaterinburg
NORTH CAUCASUS
Novosibirsk
Khabarovsk
NORTH SEA
NORWEGIAN SEA
BARENTS SEA
KARA SEA
ARCTIC
BALTIC SEA
BLACK SEA
CASPIAN SEA
ARAL SEA
CASPIAN DEPRESSION
UNITED KINGDOM
FAROE ISLANDS (Dem.)
SHETLAND ISLANDS (U.K.)
SVALBARD (Norway)
FRANZ JOSEF LAND
NOVAYA ZEMLYA
BELY ISLAND
YAMAL PENINSULA
VAYGACH ISLAND
GYDAN PENINSULA
KOLGUYEV ISLAND
KANIN PENINSULA
PECHORA SEA
WHITE SEA
TIMAN RIDGE
NORTHERN HILLS
URAL MOUNTAINS
WEST SIBERIAN PLAIN
KAZAKHSTAN
UZBEKISTAN
KYRGYZSTAN
TAJIKISTAN
CHINA
IRAN
TURKMENISTAN
ARMENIA
AZERBAIJAN
TURKEY
UKRAINE
BELARUS
POLAND
LITHUANIA
LATVIA
ESTONIA
FINLAND
SWEDEN
NORWAY
DENMARK
Oslo
Stockholm
Helsinki
Tallinn
Riga
Vilnius
Minsk
Kiev
Warsaw
Moscow
St. Petersburg
Novgorod
Kazan
Samara
Ufa
Perm
Chelyabinsk
Omsk
Novosibirsk
Tomsk
Krasnoyarsk
Kemerovo
Barnaul
Biysk
Rubtsovsk
Astana
Qaraghandy
Almaty
Bishkek
Tashkent
Dushanbe
Urumchi
Norilsk
Vorkuta
Nizhnevartovsk
Surgut
Nizhnyy Tagil
Yekaterinburg
Kurgan
Magnitogorsk
Orsk
Orenburg
Saratov
Volgograd
Rostov
Krasnodar
Sochi
Tbilisi
Yerevan
Baku
Astrakhan
Makhachkala
Volgodonsk
Novocherkassk
Kharkov
Sevastopol
Kaliningrad
Lake Balqash
Scale 1:23,700,000
1 inch equals approx. 374 miles
Conic Equal-Area Projection
© 2002 Encyclopædia Britannica, Inc.


Cities over 2,000,000
Cities 800,000 to 2,000,000
Cities 150,000 to 800,000
Cities under 150,000
Other localities
National capitals
Political subdivision capitals
TUVA Political subdivision names
International boundaries
Political subdivision boundaries
Canals
Intermittent rivers
Dams
Intermittent lakes
Salt lakes
Glaciers
Swamps and marshes
Salt flats
Sand areas
National parks
Historical sites
Spot elevations in metres (1 m = 3.28 ft)
Key to political subdivision names (shown by number on map):
1 ADYGEA
2 KARACHAY-CHERKESSIA
3 KABARDINO-BALKARIA
4 NORTH OSSETIA-ALANIA
5 CHECHNIA
6 DAGESTAN
7 INGUSHETIA
8 KALMYKIA-KHALMG TANGCHI
9 MORDVINIA
10 CHUVASHIA
11 MARI EL
12 TATARSTAN
13 UDMURTIA
14 BASHKORTOSTAN
15 KOMI-PERMYAK
16 UST-ORDA BURYAT
17 AGA-BURYAT


RUSSIA
NORWAY
SWEDEN
FINLAND
ESTONIA
LATVIA
BELARUS
UKRAINE
POLAND
LITHUANIA
KAZAKHSTAN
GEORGIA
AZERBAIJAN
BARENTS SEA
WHITE SEA
BALTIC SEA
BLACK SEA
SEA OF AZOV
CASPIAN SEA
KOLA PENINSULA
KANIN PENINSULA
KOLGUYEV ISLAND
RYBACHY PENINSULA
SOLOVETSKIE ISLANDS
KHIBINY MOUNTAINS
TIMAN RIDGE
PECHORA SEA
NORTHERN HILLS
VALDAI HILLS
SMOLENSK-MOSCOW UPLAND
CENTRAL RUSSIAN UPLAND
VOLGA HILLS
PRECASPIAN DEPRESSION
KUMA-MANYCH DEPRESSION
VOLGA DELTA
CAUCASUS MOUNTAINS
URAL MOUNTAINS
URALO-PERMYAN ECONOMIC REGION
KARELIAN PLATEAU
PECHORA-ILYCH NATURE RESERVE
LAPLAND NATURE RESERVE
DARWINSKY NATURE RESERVE
CAUCASUS NATURE RESERVE
KOMI
KARELIA
KALMYKIA
CHECHNYA
DAGESTAN
Vasyuganye Swamp
Moscow
St. Petersburg
Nizhny Novgorod
Kazan
Samara
Perm
Ufa
Chelyabinsk
Yekaterinburg
Volgograd
Saratov
Voronezh
Rostov-na-Donu
Krasnodar
Stavropol
Astrakhan
Tbilisi
Baku
Kiev
Minsk
Cities over 2,000,000
Cities 800,000 to 2,000,000
Cities 150,000 to 800,000
Cities under 150,000
Other localities
National capitals
Political subdivision capitals
Political subdivision names
International boundaries
Political subdivision boundaries
Canals
Intermittent lakes
Glaciers
Swamps and marshes
Salt flats
Sand areas
National parks
Historical sites
Spot elevations in metres (1 m = 3.28 ft)
Conic Equal-Area Projection
© 1991 Encyclopædia Britannica, Inc.

MAP INDEX

Political subdivisions

Minority republics

Adygea ..... 45 00 N 40 00 E
Altay (Gorno-Altay) ... 51 00 N 86 00 E
Bashkortostan ... 54 00 N 56 00 E
Buryatia ..... 53 00 N 109 00 E
Chechnia ..... 43 15 N 45 30 E
Chuvashia ..... 55 30 N 47 00 E
Dagestan ..... 43 00 N 47 00 E
Ingushetia ..... 43 20 N 45 00 E
Kabardino-Balkaria ..... 43 30 N 43 30 E
Kalmykia (Khalmg Tangch) ..... 46 30 N 45 30 E
Karachay-Cherkessia .... 44 00 N 42 00 E
Karelia ..... 64 00 N 32 30 E
Khakassia ..... 53 00 N 90 00 E
Komi ..... 64 00 N 54 00 E
Mari El ..... 56 30 N 48 00 E
Mordvinia ..... 54 30 N 44 00 E
North Ossetia (Alania) ..... 43 00 N 44 00 E
Sakha (Yakutia) .. 65 00 N 130 00 E
Tatarstan ..... 55 00 N 51 00 E
Tuva (Tyva) ..... 51 30 N 95 00 E
Udmurtia ..... 57 00 N 53 00 E

Autonomous *oblast* (region)

Yevreyskaya (Jewish) ..... 48 30 N 132 00 E

Autonomous *okruga* (districts)

Aga-Buryat ..... 51 00 N 114 30 E
Chukchi (Chukotka) .... 67 30 N 170 00 E
Evenk ..... 65 00 N 95 00 E
Khanty-Mansi .... 62 00 N 72 00 E
Komi-Permyak ... 60 00 N 54 30 E
Koryak ..... 62 00 N 166 00 E
Nenets ..... 67 30 N 54 00 E
Taymyr (Dolgano-Nenets) ..... 72 00 N 95 00 E
Ust-Orda Buryat ..... 53 30 N 104 00 E
Yamalo-Nenets .... 66 00 N 76 00 E

Cities and towns

Abakan ..... 53 43 N 91 26 E
Achinsk ..... 56 17 N 90 30 E
Aginskoye ..... 51 06 N 114 32 E
Akademgorodok .. 54 49 N 83 09 E
Aldan ..... 58 37 N 125 24 E
Aleksandrovsk-Grushevsky, see Shakhty
Aleksandrovsk-Sakhalinsky .... 50 54 N 142 10 E
Alekseyevsk, see Svobodny
Aleksin ..... 54 31 N 37 05 E
Almetyevsk ..... 54 53 N 52 20 E
Amursk ..... 50 14 N 136 54 E
Anadyr (Novo-Mariinsk) .. 64 45 N 177 29 E
Andropov, see Rybinsk
Angarsk ..... 52 34 N 103 54 E
Anzhero-Sudzhensk .... 56 07 N 86 00 E
Apatity ..... 67 34 N 33 22 E
Arkhangelsk (Archangel) .... 64 34 N 40 32 E
Armavir ..... 45 00 N 41 08 E
Arsenyev ..... 44 10 N 133 15 E
Artyom (Artëm) .. 43 22 N 132 13 E
Arzamas ..... 55 23 N 43 50 E
Asbest ..... 57 00 N 61 30 E
Asino ..... 57 00 N 86 09 E
Astrakhan ..... 46 21 N 48 03 E
Azov ..... 47 07 N 39 25 E
Balakovo ..... 52 02 N 47 47 E
Balashov ..... 51 32 N 43 08 E
Baley ..... 51 36 N 116 38 E
Baltiysk (Baltiisk, or Pillau) ..... 54 39 N 19 55 E
Barnaul ..... 53 22 N 83 45 E
Bataysk (Bataisk) ..... 47 10 N 39 44 E
Belgorod ..... 50 36 N 36 34 E
Belogorsk ..... 50 55 N 128 28 E
Beloretsk ..... 53 58 N 58 24 E
Belovo ..... 54 25 N 86 18 E
Berdsk ..... 54 47 N 83 02 E
Berezniki ..... 59 24 N 56 46 E
Bikin ..... 46 48 N 134 16 E
Bilibino ..... 68 03 N 166 20 E
Birobidzhan ..... 48 48 N 132 57 E
Biysk (Bilisk) ..... 52 34 N 85 15 E
Blagoveshchensk . 50 16 N 127 32 E
Bobriki, see Novomoskovsk
Bodaybo ..... 57 51 N 114 10 E
Bogotol ..... 56 12 N 89 33 E
Bolshoy Kamen .. 43 06 N 132 21 E
Borisoglebsk .... 51 23 N 42 06 E
Borovichi ..... 58 24 N 33 55 E
Borzya ..... 50 24 N 116 31 E
Bratsk ..... 56 21 N 101 55 E
Brezhnev, see Naberezhnye Chelny
Bryansk ..... 53 15 N 34 22 E
Bugulma ..... 54 33 N 52 48 E
Buguruslan ..... 53 39 N 52 26 E
Buzuluk ..... 52 47 N 52 18 E
Chapayevsk (Ivashchenkovo) . 52 58 N 49 41 E
Cheboksary ..... 56 09 N 47 15 E
Chelyabinsk ..... 55 10 N 61 24 E
Cheremkhovo .... 53 09 N 103 05 E
Cherepovets ..... 59 08 N 37 54 E
Cherkessk ..... 44 14 N 42 03 E
Chernogorsk ..... 53 49 N 91 18 E
Chernyakhovsk .. 54 36 N 21 48 E
Chita ..... 52 03 N 113 30 E
Chkalov, see Orenburg
Derbent ..... 42 03 N 48 18 E
Dimitrovgrad (Melekess) ..... 54 15 N 49 33 E
Dudinka ..... 69 25 N 86 15 E
Dzaudzhikau, see Vladikavkaz
Dzerzhinsk ..... 56 15 N 43 24 E
Ekaterinodar, see Krasnodar
Elektrostal ..... 55 47 N 38 28 E
Elista ..... 46 16 N 44 14 E
Engels ..... 51 30 N 46 07 E
Essentuki, see Yessentuki
Gatchina (Krasnogvardeysk) ..... 59 34 N 30 08 E
Georgiu-Dezh, see Liski
Glazov ..... 58 09 N 52 40 E
Gorky, see Nizhny Novgorod
Gorno-Altaysk (Ulala, or Oyrot-Tura) .... 51 58 N 85 58 E
Grozny ..... 43 20 N 45 42 E
Gubkin ..... 51 17 N 37 32 E
Gukovo ..... 48 03 N 39 56 E
Gusev ..... 54 36 N 22 12 E
Igarka ..... 67 28 N 86 35 E
Ilansky ..... 56 14 N 96 03 E
Inta ..... 66 05 N 60 08 E
Ioshkar-Ola, see Yoshkar-Ola
Irbit ..... 57 41 N 63 03 E
Irkutsk ..... 52 16 N 104 20 E
Ishim ..... 56 09 N 69 27 E
Ishimbay ..... 53 28 N 56 02 E
Iskitim ..... 54 37 N 83 24 E
Ivanovo ..... 57 00 N 40 59 E
Ivashchenkovo, see Chapayevsk
Izhevsk (Ustinov) ..... 56 51 N 53 14 E
Kalinin, see Tver
Kaliningrad ..... 55 55 N 37 49 E
Kaliningrad (Königsberg) ... 54 43 N 20 30 E
Kaluga ..... 54 31 N 36 16 E
Kamen-na-Obi ... 53 47 N 81 20 E
Kamensk-Shakhtinsky ... 48 21 N 40 19 E
Kamensk-Uralsky . 56 25 N 61 54 E
Kamyshin ..... 50 06 N 45 24 E
Kansk ..... 56 13 N 95 41 E
Karabash ..... 55 29 N 60 14 E
Karasuk ..... 53 44 N 78 02 E
Kazan ..... 55 45 N 49 08 E
Kemerovo ..... 55 20 N 86 05 E
Khabarovsk ..... 48 30 N 135 06 E
Khanty-Mansiysk (Ostyako-Vogulsk) ..... 61 00 N 69 06 E
Khasavyurt ..... 43 15 N 46 35 E
Khem-Beldyr, see Kyzyl
Kholmogory ..... 64 14 N 41 39 E
Kimry ..... 56 52 N 37 21 E
Kineshma ..... 57 28 N 42 07 E
Kirov, see Vyatka
Kirovo-Chepetsk .. 58 33 N 50 01 E
Kirovograd ..... 57 26 N 60 04 E
Kirovsk (Khibinogorsk) .. 67 37 N 33 40 E
Kiselevsk (Kiselyovsk) .... 54 00 N 86 39 E
Kislovodsk ..... 43 55 N 42 43 E
Kizel ..... 59 03 N 57 40 E
Klintsy ..... 52 45 N 32 14 E
Klyuchi ..... 56 18 N 160 51 E
Kolchugino, see Leninsk-Kuznetsky
Kolomna ..... 55 05 N 38 47 E
Kolpashevo ..... 58 20 N 82 50 E
Kolpino ..... 59 45 N 30 36 E
Komsomolsk-na-Amure ..... 50 35 N 137 02 E
Königsberg, see Kaliningrad
Kopeysk (Kopeisk, or Ugolnye Kopi) .. 55 07 N 61 37 E
Korf ..... 60 19 N 165 50 E
Korkino ..... 54 54 N 61 23 E
Korsakov ..... 46 38 N 142 46 E
Kostroma ..... 57 46 N 40 55 E
Kotlas ..... 61 16 N 46 35 E
Kovrov ..... 56 25 N 41 18 E
Kozlov, see Michurinsk
Krasnodar (Ekaterinodar, or Yekaterinodar) .. 45 02 N 39 00 E
Krasnogvardeysk, see Gatchina
Krasnokamensk .. 50 03 N 118 01 E
Krasnokamsk .... 58 04 N 55 48 E
Krasnoturinsk .... 59 46 N 60 12 E
Krasnouralsk .... 58 21 N 60 03 E
Krasnoyarsk ..... 56 01 N 92 50 E
Kropotkin ..... 45 26 N 40 34 E
Kstovo ..... 56 11 N 44 11 E
Kudymkar ..... 59 01 N 54 39 E
Kulebaki ..... 55 24 N 42 32 E
Kumertau, see Yermolaevo
Kungur ..... 57 25 N 56 57 E
Kurgan ..... 55 26 N 65 18 E
Kursk ..... 51 42 N 36 12 E
Kushva ..... 58 18 N 59 45 E
Kuybyshev, see Samara
Kuznetsk ..... 53 07 N 46 36 E
Kuznetsk, see Novokuznetsk
Kyakhta ..... 50 20 N 106 30 E
Kyzyl (Khem-Beldyr) .. 51 42 N 94 27 E
Leningrad, see St. Petersburg
Leninogorsk ..... 54 36 N 52 30 E
Leninsk-Kuznetsky (Kolchugino) ... 54 38 N 86 10 E
Lensk ..... 60 43 N 114 55 E
Lesosibirsk ..... 58 16 N 92 29 E
Lesozavodsk .... 45 28 N 133 27 E
Lipetsk ..... 52 37 N 39 35 E
Liski (Georgiu-Dezh) .. 50 59 N 39 30 E
Lomonosov (Oranienbaum) .. 59 55 N 29 46 E
Lysva ..... 58 07 N 57 47 E
Lyubertsy ..... 55 41 N 37 53 E
Magadan ..... 59 34 N 150 48 E
Magnitogorsk .... 53 27 N 59 04 E
Makhachkala ..... 42 58 N 47 30 E
Maykop ..... 44 35 N 40 10 E
Mednogorsk ..... 51 24 N 57 37 E
Megion ..... 61 03 N 76 06 E
Melekess, see Dimitrovgrad
Mezhdurechensk . 53 42 N 88 03 E
Miass ..... 54 59 N 60 06 E
Michurinsk (Kozlov) ..... 52 54 N 40 30 E
Mikhaylovka ..... 50 04 N 43 15 E
Mirny ..... 62 33 N 113 53 E
Molotov, see Perm
Molotovsk, see Severodvinsk
Monchegorsk .... 67 56 N 32 58 E
Moscow (Moskva) ..... 55 45 N 37 35 E
Murmansk ..... 68 58 N 33 05 E
Murom ..... 55 34 N 42 02 E
Mys Shmidta .... 68 56 N 179 27 E
Naberezhnye Chelny (Brezhnev) ..... 55 42 N 52 19 E
Nadezhdinsk, see Serov
Nadym ..... 65 35 N 72 42 E
Nakhodka ..... 42 47 N 132 52 E
Nalchik ..... 43 29 N 43 37 E
Naryan-Mar ..... 67 39 N 53 00 E
Nazran ..... 43 13 N 44 46 E
Neftekamsk ..... 56 05 N 54 16 E
Nefteyugansk .... 61 05 N 72 42 E
Nerchinsk ..... 51 58 N 116 35 E
Neryungri ..... 56 39 N 124 38 E
Nevinnomyssk .... 44 38 N 41 57 E
Nikolayevsk-na-Amure ..... 53 08 N 140 44 E
Nizhnekamsk ..... 55 36 N 51 47 E
Nizhnekolymsk .... 68 32 N 160 56 E
Nizhneudinsk .... 54 54 N 99 03 E
Nizhnevartovsk ... 60 56 N 76 38 E
Nizhny Novgorod (Gorky) ..... 56 20 N 44 00 E
Nizhny Tagil ..... 57 55 N 59 57 E
Norilsk ..... 69 20 N 88 06 E
Novgorod ..... 58 31 N 31 17 E
Novo-Mariinsk, see Anadyr
Novocherkassk .. 47 25 N 40 06 E
Novokuybyshevsk ..... 53 07 N 49 58 E
Novokuznetsk (Kuznetsk, or Stalinsk) ..... 53 45 N 87 06 E
Novomoskovsk (Bobriki, or Stalinogorsk) ... 54 05 N 38 13 E
Novorossiysk .... 44 43 N 37 47 E
Novoshakhtinsk .. 47 47 N 39 56 E
Novosibirsk ..... 55 02 N 82 55 E
Novotroitsk ..... 51 12 N 58 20 E
Novy Urengoy (Urengoy) ..... 65 58 N 78 25 E
Noyabrsk ..... 63 10 N 75 37 E
Nyandoma ..... 61 40 N 40 12 E
Nyurba ..... 63 17 N 118 20 E
Ob ..... 55 00 N 82 43 E
Oktyabrsky ..... 54 28 N 53 28 E
Olyokminsk ..... 60 24 N 120 24 E
Omsk ..... 55 00 N 73 24 E
Onega ..... 63 54 N 38 08 E
Oranienbaum, see Lomonosov
Ordzhonikidze, see Vladikavkaz
Orenburg (Chkalov) ..... 51 45 N 55 06 E
Orsk ..... 51 12 N 58 34 E
Oryol ..... 52 55 N 36 05 E
Ostyako-Vogulsk, see Khanty-Mansiysk
Oymyakon ..... 63 28 N 142 49 E
Oyrot-Tura, see Gorno-Altaysk
Palana ..... 59 07 N 159 58 E
Partizansk (Suchan) ..... 43 07 N 133 05 E
Pechenga ..... 69 33 N 31 12 E
Pechora ..... 65 25 N 57 02 E
Penza ..... 53 13 N 45 00 E
Perm (Molotov) ... 58 00 N 56 15 E

Pervouralsk ..... 56 54 N 59 58 E
Petrodvorets (Peterhof) ..... 59 53 N 29 54 E
Petropavlovsk-Kamchatsky ... 53 01 N 158 39 E
Petrozavodsk .... 61 49 N 34 20 E
Pevek ..... 69 42 N 170 17 E
Pillau, see Baltiysk
Podolsk ..... 55 26 N 37 33 E
Primorsko-Akhtarsk ..... 46 03 N 38 10 E
Prokopyevsk .... 53 53 N 86 45 E
Pskov ..... 57 50 N 28 20 E
Pushkin (Tsarskoye Selo) ..... 59 43 N 30 25 E
Pyatigorsk ..... 44 01 N 43 05 E
Revda ..... 56 48 N 59 57 E
Roslavl ..... 53 57 N 32 52 E
Rostov ..... 57 11 N 39 25 E
Rostov-na-Donu (Rostov-on-Don) . 47 14 N 39 42 E
Rubtsovsk ..... 51 30 N 81 15 E
Ryazan (Riazan) ..... 54 38 N 39 44 E
Rybinsk (Andropov, or Shcherbakov) ... 58 03 N 38 50 E
Rzhev ..... 56 15 N 34 20 E
St. Petersburg (Leningrad, or Sankt Peterburg) ..... 59 55 N 30 15 E
Salavat ..... 53 21 N 55 55 E
Salekhard ..... 66 33 N 66 40 E
Salsk ..... 46 28 N 41 33 E
Samara (Kuybyshev) .... 53 12 N 50 09 E
Sangar ..... 63 55 N 127 31 E
Sankt Peterburg, see St. Petersburg
Saransk ..... 54 11 N 45 11 E
Sarapul ..... 56 28 N 53 48 E
Saratov ..... 51 34 N 46 02 E
Sergiyev Posad .. 56 18 N 38 08 E
Serov (Nadezhdinsk) .. 59 36 N 60 35 E
Serpukhov ..... 54 55 N 37 25 E
Severodvinsk (Molotovsk) .... 64 34 N 39 50 E
Severomorsk .... 69 05 N 33 27 E
Shadrinsk ..... 56 05 N 63 38 E
Shakhty (Aleksandrovsk-Grushevsky) ... 47 42 N 40 13 E
Shcherbakov, see Rybinsk
Shchyokino (Shchekino) .... 54 00 N 37 31 E
Shimanovsk ..... 52 00 N 127 42 E
Shiisseltburg, see Petrokrepost
Simbirsk, see Ulyanovsk
Skovorodino .... 53 59 N 123 55 E
Smolensk ..... 54 47 N 32 03 E
Sochi ..... 43 35 N 39 45 E
Sokol ..... 59 28 N 40 10 E
Solikamsk ..... 59 39 N 56 47 E
Sovetsk (Tilsit) .... 55 05 N 21 53 E
Sovetskaya Gavan ..... 48 58 N 140 18 E
Spassk-Dalniy .... 44 37 N 132 48 E
Srednekolymsk ... 67 27 N 153 41 E
Stalingrad, see Volgograd
Stalinogorsk, see Novomoskovsk
Stalinsk, see Novokuznetsk
Staraya Russa ... 58 00 N 31 23 E
Stary Oskol ..... 51 17 N 37 51 E
Stavropol, see Tolyatti
Stavropol (Voroshilovsk) .... 45 03 N 41 58 E
Sterlitamak ..... 53 37 N 55 58 E
Strezhevoy ..... 60 42 N 77 34 E
Suchan, see Partizansk
Surgut ..... 61 14 N 73 20 E
Susuman ..... 62 47 N 148 10 E
Sverdlovsk, see Yekaterinburg
Svobodny (Alekseyevsk) ... 51 24 N 128 08 E
Syktyvkar ..... 61 40 N 50 48 E
Syzran ..... 53 11 N 48 27 E
Taganrog ..... 47 12 N 38 56 E
Talnakh ..... 69 30 N 88 11 E
Tambov ..... 52 43 N 41 27 E
Tara ..... 56 54 N 74 22 E
Tatarsk ..... 55 13 N 75 58 E
Tavda ..... 58 03 N 65 15 E
Tayshet ..... 55 57 N 98 00 E
Tikhoretsk ..... 45 51 N 40 07 E
Tikhvin ..... 59 39 N 33 31 E
Tiksi ..... 71 36 N 128 48 E
Tilsit, see Sovetsk
Tobolsk ..... 58 12 N 68 16 E
Tolyatti (Stavropol) ..... 53 31 N 49 26 E
Tomsk ..... 56 30 N 84 58 E
Torzhok ..... 57 03 N 34 58 E
Troitsk ..... 54 06 N 61 35 E
Tsaritsyn, see Volgograd
Tsarskoye Selo, see Pushkin
Tuapse ..... 44 05 N 39 06 E
Tula ..... 54 12 N 37 37 E
Tulun ..... 54 36 N 100 32 E
Tura ..... 64 17 N 100 15 E
Turukhansk ..... 65 49 N 87 59 E
Tver (Kalinin) .... 56 52 N 35 55 E
Tynda ..... 55 10 N 124 43 E
Tyumen ..... 57 09 N 65 26 E
Udachny ..... 62 33 N 113 53 E
Ufa ..... 55 45 N 55 56 E
Ugolnye Kopi, see Kopeysk
Ukhta ..... 63 33 N 53 40 E
Ulala, see Gorno-Altaysk
Ulan-Ude ..... 51 50 N 107 37 E
Ulyanovsk (Simbirsk) ..... 54 20 N 48 24 E
Urengoy, see Novy Urengoy
Usolye-Sibirskoye ..... 52 45 N 103 41 E
Ussuriysk ..... 43 48 N 131 59 E
Ust-Ilimsk ..... 58 03 N 102 39 E
Ust-Kut ..... 56 46 N 105 40 E
Ust-Omchuk .... 61 09 N 149 38 E
Ust-Ordinsky .... 52 48 N 104 45 E
Ustinov, see Izhevsk
Ulyar ..... 55 48 N 94 20 E
Uzhur ..... 55 18 N 89 50 E
Uzlovaya ..... 53 59 N 38 10 E
Velikiye Luki .... 56 20 N 30 32 E
Veliky Ustyug .... 60 48 N 46 18 E
Velsk ..... 61 05 N 42 08 E
Verkhoyansk .... 67 35 N 133 27 E
Viipuri, see Vyborg
Vladikavkaz (Dzaudzhikau, or Ordzhonikidze) .. 43 40 N 44 40 E
Vladimir ..... 56 10 N 40 25 E
Vladivostok ..... 43 08 N 131 54 E
Volgodonsk ..... 47 32 N 42 09 E
Volgograd (Stalingrad, or Tsaritsyn) ..... 48 45 N 44 25 E
Vologda ..... 59 13 N 39 54 E
Volsk ..... 52 02 N 47 23 E
Volzhsky ..... 48 49 N 44 44 E
Vorkuta ..... 67 30 N 64 00 E
Voronezh ..... 51 38 N 39 12 E
Voroshilovsk, see Stavropol
Votkinsk ..... 57 03 N 53 39 E
Vyatka (Viatka, or Kirov) ..... 58 33 N 49 42 E
Vyazemsky ..... 47 32 N 134 45 E
Vyazma ..... 55 13 N 34 18 E
Vyborg (Viipuri) .. 60 42 N 28 45 E
Vyshny Volochyok ..... 57 35 N 34 34 E
Yagodnaye ..... 62 33 N 149 40 E
Yakutsk ..... 62 00 N 129 40 E
Yalutorovsk ..... 56 40 N 66 18 E
Yaroslavl ..... 57 37 N 39 52 E
Yasnaya Polyana . 54 04 N 37 31 E
Yefremov ..... 53 09 N 38 07 E
Yekaterinburg (Sverdlovsk) .... 56 51 N 60 36 E
Yekaterinodar, see Krasnodar
Yelets ..... 52 37 N 38 30 E
Yelizovo ..... 53 11 N 158 23 E
Yermolaevo (Kumertau) ..... 52 46 N 55 47 E
Yessentuki (Essentuki) .... 44 03 N 42 51 E
Yeysk ..... 46 42 N 38 17 E
Yoshkar-Ola (Ioshkar-Ola) .... 56 40 N 47 55 E
Yurga ..... 55 42 N 84 51 E
Yuzhno-Sakhalinsk .... 46 57 N 142 44 E
Zavitinsk ..... 50 07 N 129 27 E
Zelenodolsk ..... 55 51 N 48 33 E
Zeya ..... 53 45 N 127 16 E
Zhatay ..... 62 10 N 129 49 E
Zheleznogorsk ... 52 19 N 35 21 E
Zlatoust ..... 55 10 N 59 40 E

Physical features and points of interest

Adycha, *river* .... 68 13 N 134 41 E
Adzva, *river* .... 66 36 N 59 22 E
Akhtuba, *river* .... 46 42 N 48 00 E
Aldan, *river* .... 63 28 N 129 35 E
Aldan Plateau .... 57 00 N 127 00 E
Ailakh-Yun, *river* .. 60 36 N 136 54 E
Altai (Altay) Mountains ..... 48 00 N 90 00 E
Altai (Altay) Nature Reserve . 51 20 N 88 00 E
Amga, *river* ..... 62 38 N 134 32 E
Amur, *river* ..... 52 56 N 141 10 E
Anabar, *river* .... 73 08 N 113 35 E
Anabar Plateau .. 70 00 N 108 00 E
Anadyr, *river* .... 64 54 N 176 13 E
Anadyr, Gulf of .. 64 00 N 178 00 W
Angara, *river* .... 58 06 N 93 00 E
Arctic Ocean .... 85 00 N 100 00 E
Ay, *river* ..... 56 08 N 57 40 E
Ayon Island ..... 69 50 N 168 40 E
Azov, Sea of .... 46 00 N 36 00 E
Badzhalsky Mountains .... 50 40 N 134 50 E
Baikal (Baykal), Lake ..... 54 00 N 109 00 E
Baikal-Lena Nature Reserve .. 54 20 N 108 15 E
Baikal Mountains . 54 00 N 108 00 E
Baikal Nature Reserve ..... 51 20 N 105 00 E
Baltic Sea .... 56 00 N 18 00 E
Balygychan, *river* . 63 58 N 154 12 E
Barents Sea .... 74 00 N 36 00 E
Barguzin Nature Reserve ..... 54 20 N 110 02 E
Belaya, *river* ..... 55 54 N 53 33 E
Belaya, *river* ..... 65 31 N 173 17 E
Beloye (White) Lake ..... 60 26 N 36 27 E
Beloye, see White Sea
Belukha, Mount .. 49 48 N 86 35 E
Bely Island ..... 73 10 N 70 45 E
Bering Sea ..... 60 00 N 175 00 W
Black Sea ..... 43 00 N 35 00 E
Bolgary, *historical site* ..... 54 58 N 49 04 E
Bolshevik Island .. 78 40 N 102 30 E
Bolshoy Anyuy, *river* ..... 68 30 N 160 49 E
Bolshoy Begichev Island ..... 74 20 N 112 30 E
Bolshoy Irgiz, *river* ..... 52 01 N 47 24 E
Bratsk Reservoir .. 56 05 N 101 50 E
Bureya, *river* .... 49 27 N 129 47 E
Byrranga Mountains ..... 75 00 N 104 00 E
Caspian Depression .... 48 00 N 52 00 E
Caspian Sea .... 42 00 N 50 00 E
Caucasus Mountains ..... 42 00 N 45 00 E
Caucasus Nature Reserve ..... 43 50 N 40 30 E
Central Russian Upland ..... 52 00 N 38 00 E
Central Siberian Plateau ..... 66 00 N 106 00 E
Central Yakut Lowland ..... 65 00 N 125 00 E
Chany, Lake .... 54 50 N 77 30 E
Cheboksary Reservoir ..... 56 00 N 46 10 E
Chelyuskin, Cape ..... 77 45 N 104 20 E
Chersky Mountains .... 65 00 N 144 00 E
Chukchi Mountains .... 67 00 N 176 00 E
Chukchi Peninsula..... 66 00 N 174 00 W
Chukchi Sea .... 69 00 N 174 00 E
Chulym, *river* .... 57 43 N 83 51 E
Chunya, *river* .... 61 36 N 96 30 E
Chusovaya, *river* . 58 13 N 56 22 E
Commander Islands, see Komandor Islands
Courland Lagoon . 55 00 N 21 00 E
Courland Spit .... 55 18 N 21 00 E
Crimean Peninsula ..... 45 00 N 34 00 E
Darvinsky Nature Reserve ..... 58 50 N 37 40 E
Daugava, see Western Dvina
Denezhkin, Mount ..... 60 25 N 59 32 E
Desna, *river* ..... 50 33 N 30 32 E
Dezhnev, Cape .. 66 05 N 169 40 W
Dmitriya Lapteva Strait ..... 73 00 N 142 00 E
Dnieper (Dnepr), *river* ..... 46 30 N 32 18 E
Don, *river* ..... 47 04 N 39 18 E
Donets, *river* .... 47 35 N 40 54 E
Donets Basin (Donbass) ..... 48 30 N 38 30 E
Dyoma, *river* .... 54 42 N 55 59 E
Dzhugdzhur Range ..... 58 00 N 136 00 E
East Siberian Sea ..... 74 00 N 166 00 E
Eastern Sakhalin Mountains ..... 51 00 N 143 10 E
Eastern Sayan Mountains ..... 53 00 N 97 00 E
Elbrus, Mount .... 43 21 N 42 26 E
Elton, Lake ..... 49 16 N 46 35 E
Franz Josef Land, *islands* .. 81 00 N 55 00 E
Gdańsk, Gulf of .. 54 40 N 19 15 E
Gonarn, *river* .... 57 21 N 131 14 E
Gydan Peninsula ..... 70 50 N 79 00 E
Ik, *river* ..... 55 55 N 52 36 E
Ilmen, Lake ..... 58 17 N 31 20 E
Indigirka, *river* .... 70 48 N 148 54 E
Iniklinsky Reservoir ..... 51 45 N 58 45 E
Irtysh, *river* ..... 61 04 N 68 52 E
Iset, *river* ..... 56 36 N 66 24 E
Ishim, *river* ..... 57 42 N 71 12 E
Izhma, *river* ..... 65 19 N 52 54 E
Kama, *river* ..... 55 25 N 50 40 E
Kama Reservoir .. 58 30 N 56 30 E
Kamchatka, *river* .. 55 15 N 162 30 E
Kamchatka Peninsula ..... 56 00 N 160 00 E
Kamchatsky Gulf .. 55 35 N 162 21 E
Karnen, Mount .... 69 06 N 94 48 E
Kanin Peninsula .. 68 00 N 45 00 E
Kara Sea ..... 76 00 N 80 00 E
Karagin Island ... 58 50 N 164 00 E
Karelian Isthmus .. 60 25 N 30 00 E
Karelian Plateau .. 63 20 N 32 00 E
Kazym, *river* .... 63 54 N 65 50 E
Kerch Strait ..... 45 20 N 36 30 E
Ket, *river* ..... 58 55 N 81 32 E
Khanka, Lake .... 45 00 N 132 24 E
Khantayskoye Reservoir ..... 68 20 N 88 00 E
Khatanga, *river* ... 72 55 N 106 00 E
Kheta, *river* .... 71 54 N 102 06 E
Khibiny Mountains ..... 67 45 N 33 45 E
Khingansk Nature Reserve ..... 49 30 N 129 40 E

Khopyor, *river* . . . 49 35 N 42 17 E
Klyuchevskaya Volcano . . . . . 56 04 N 160 38 E
Kola Peninsula . . . 67 20 N 37 00 E
Kolguyev Island . . 69 05 N 49 15 E
Kolva, *river* . . . . . 65 55 N 57 15 E
Kolyma, *river* . . . . 69 30 N 161 00 E
Kolyma Lowland . . 68 30 N 154 00 E
Kolyma Mountains . . . . . 63 00 N 160 00 E
Komandor (Commander) Islands . . . . . 55 00 N 167 00 E
Komsomolets Island . . . . . 80 30 N 95 00 E
Komsomolsk Nature Reserve . 50 35 N 137 02 E
Konzhakovsky, Mount . . . . . 59 38 N 59 08 E
Korkodon, *river* . . 64 44 N 154 00 E
Koryak Mountains . 62 30 N 172 00 E
Kosva, *river* . . . . . 58 47 N 56 45 E
Kotelny Island . . . 75 45 N 138 44 E
Kotuy, *river* . . . . . 71 55 N 102 05 E
Krasnoyarsk Reservoir . . . . . 55 00 N 91 05 E
Kronotsky Nature Reserve . . . . . 55 00 N 160 05 E
Kuban, *river* . . . . 45 20 N 37 22 E
Kuban Plain . . . . . 46 00 N 39 30 E
Kuloy, *river* . . . . . 66 03 N 43 22 E
Kulunda Steppe . . 53 00 N 79 00 E
Kuma, *river* . . . . . 44 55 N 46 32 E
Kuma-Manych Depression . . . . 46 00 N 43 00 E
Kuril Islands . . . . 46 10 N 152 00 E
Kuril Nature Reserve . . . . . 44 20 N 146 08 E
Kuyto, Lake . . . . . 65 08 N 31 24 E
La Perouse Strait . . . . . 45 45 N 142 00 E
Lacha, Lake . . . . . 61 20 N 38 48 E
Ladoga, Lake . . . . 61 00 N 31 30 E
Lapland Nature Reserve . . . . . 67 50 N 32 10 E
Laptev Sea . . . . . 76 00 N 126 00 E
Lazovsky Nature Reserve . . . . . 43 10 N 134 10 E
Lena, *river* . . . . . 72 25 N 126 40 E
Little Ob, *channel* . . . . . 66 05 N 65 40 E
Long (Longa) Strait . . . . . 70 20 N 178 00 E
Lopatka, Cape . . . 50 52 N 156 40 E
Lower Karna Reservoir . . . . . 55 10 N 52 50 E
Lower Tunguska, *river* . . . . . 65 48 N 88 04 E
Luza, *river* . . . . . 60 35 N 46 26 E
Lyakhovsky Islands . . . . . 73 30 N 141 00 E
Lyapin, *river* . . . . 63 36 N 61 54 E
Magadan Nature Reserve . . . . . 59 40 N 146 30 E
Manych, *river* . . . 47 15 N 40 15 E
Markha, *river* . . . . 63 28 N 118 50 E
Maya, *river* . . . . . 60 24 N 134 30 E
Medveditsa, *river* . 49 35 N 42 41 E
Mezen, *river* . . . . . 66 11 N 43 59 E
Moscow (Moskva), *river* . . 55 05 N 38 51 E
Moscow Canal . . . 56 43 N 37 08 E
Muass, *river* . . . . . 58 00 N 64 00 E
Muna, *river* . . . . . 67 52 N 123 06 E
Mus-Khaya, Mount . . . . . 62 35 N 140 50 E
Nadym, *river* . . . . 66 12 N 72 00 E
Narodnaya, Mount . . . . . 65 04 N 60 09 E
Neman (Nemunas), *river* . 55 18 N 21 23 E
New Siberia (Novaya Sibir) Island . . . . . 75 00 N 149 00 E
New Siberian (Novosibirskie) Islands . . . . . 75 00 N 142 00 E
Nizhny Novgorod Reservoir . . . . . 57 00 N 43 00 E
North Siberian Lowland (Taymyr Depression) . . . 72 00 N 104 00 E
Northern (Severnaya) Dvina, *river* . . . . . 64 32 N 40 30 E
Northern Hills . . . . 60 00 N 50 00 E
Northern Sosva, *river* . . . . . 64 10 N 65 28 E
Novaya Zemlya, *islands* . . . . . 74 00 N 57 00 E
Novosibirsk Reservoir . . . . . 54 35 N 82 35 E
Ob, *river* . . . . . 66 45 N 69 30 E
Ob, Gulf of . . . . . 69 00 N 73 00 E
Oka, *river* . . . . . 56 20 N 43 59 E
Oka-Don Lowland . 53 00 N 40 00 E
Okhotsk, Sea of . . 55 00 N 150 00 E
Olenyok, *river* . . . . 73 00 N 119 55 E
Olkhon Island . . . . 53 09 N 107 24 E
Olyokma, *river* . . . 60 22 N 120 42 E
Olyutorsky, Cape . 59 55 N 170 27 E
Om, *river* . . . . . 54 59 N 73 22 E
Omolon, *river* . . . . 68 42 N 158 36 E
Onega, *river* . . . . . 63 58 N 38 02 E
Onega, Lake . . . . . 61 30 N 35 45 E
Oskol, *river* . . . . . 49 06 N 37 25 E
Pacific Ocean . . . 45 00 N 160 15 E
Palyavaam, *river* . . 68 50 N 170 45 E
Paramushir Island . . . . . 50 25 N 155 50 E
Patience (Terpeniya), Cape . . . . . 48 39 N 144 44 E
Patom Plateau . . . 59 20 N 116 20 E
Payer, Mount . . . . 66 42 N 64 25 E
Pechora, *river* . . . . 68 13 N 54 15 E
Pechora Sea . . . . 70 00 N 54 00 E
Pechoro-Ilychsky Nature Reserve . 62 20 N 59 00 E
Peipus, Lake . . . . 58 45 N 27 30 E
Penzhin Bay . . . . . 61 00 N 162 00 E
Penzhina, *river* . . . 62 28 N 165 18 E
Peter The Great (Petra Velikogo) Bay . . . . . 42 40 N 132 00 E
Peza, *river* . . . . . 65 36 N 44 35 E
Pinega, *river* . . . . . 64 08 N 41 54 E
Pobeda, Mount . . 65 12 N 146 12 E
Poluy, *river* . . . . . 66 31 N 66 33 E
Poncy, *river* . . . . . 66 59 N 41 17 E
Pregolya, *river* . . . 54 41 N 20 22 E
Priangarskoye Plateau . . . . . 58 00 N 98 00 E
Prilenskoye Plateau . . . . . 60 45 N 125 00 E
Pur, *river* . . . . . 67 31 N 77 55 E
Putoran (Putorana) Mountains . . . . . 69 00 N 95 00 E
Pyshma, *river* . . . . 57 08 N 66 18 E
Rabocheostrovsk Island . . . . . 79 30 N 97 00 E
Russian Plain . . . . . 55 00 N 45 00 E
Rybachy Peninsula . . . . . 69 42 N 32 36 E
Rybinsk Reservoir . 58 30 N 38 25 E
Saiano-Shushenskoye Reservoir . . . . . 52 10 N 92 30 E
Sakhalin Island . . . 51 00 N 143 00 E
Sal, *river* . . . . . 47 31 N 40 45 E
Samara, *river* . . . . 53 10 N 50 04 E
Samara Reservoir . 55 00 N 48 20 E
Saratov Reservoir . 52 45 N 48 30 E
Seg, Lake . . . . . 63 18 N 33 45 E
Seliger, Lake . . . . 57 13 N 33 05 E
Severnaya Dvina, see Northern Dvina
Severnaya Zemlya, *islands* . . . . . 79 30 N 98 00 E
Shantar Islands . . . 55 00 N 137 36 E
Shapkina, *river* . . . 66 44 N 52 25 E
Shelikhova, Gulf of . . . . . 59 45 N 158 00 E
Shilka, *river* . . . . . 53 20 N 121 26 E
Shuya, *river* . . . . . 61 54 N 34 15 E
Siberia (Sibir), *region* . . . . . 60 00 N 100 00 E
Sikhote-Alin Mountains . . . . . 48 00 N 138 00 E
Sikhote-Alin Nature Reserve . 45 12 N 136 30 E
Smolensk-Moscow Upland . . . . . 55 00 N 35 00 E
Solovetskie Islands . . . . . 65 07 N 35 53 E
Sredinny Mountains . . . . . 56 00 N 158 00 E
Stanovoy Mountains . . . . . 56 20 N 126 00 E
Stanovoy Plateau . 56 00 N 114 00 E
Stavropol Upland . . . . . 45 00 N 43 00 E
Stony Tunguska, *river* . . . . . 61 36 N 90 18 E
Sukhona, *river* . . . 60 46 N 46 24 E
Suntar-Khayata Mountains . . . . . 62 00 N 143 00 E
Sura, *river* . . . . . 56 06 N 46 00 E
Svir, *river* . . . . . 60 30 N 32 48 E
Sysola, *river* . . . . . 61 42 N 50 53 E
Tannu-Ola Mountains . . . . . 51 00 N 94 00 E
Tanyu, *river* . . . . . 66 08 N 64 05 E
Tardoki-Yani, Mount . . . . . 48 55 N 138 04 E
Tatar Strait . . . . . 50 00 N 141 00 E
Tavda, *river* . . . . . 57 47 N 67 16 E
Taymura, *river* . . . 63 46 N 98 01 E
Taymyr, Lake . . . . 74 30 N 102 30 E
Taymyr Depression, see North Siberian Lowland
Taymyr Nature Reserve . . . . . 74 00 N 100 00 E
Taymyr Peninsula . 76 00 N 104 00 E
Taz, *river* . . . . . 67 32 N 78 40 E
Telpos-Iz, Mount . . 63 54 N 59 10 E
Terpeniya, see Patience, Cape
Timan Ridge . . . . 65 00 N 51 00 E
Tobol, *river* . . . . . 58 10 N 68 12 E
Top, Lake . . . . . 65 40 N 32 00 E
Tsimlyansk Reservoir . . . . . 48 00 N 43 00 E
Tsna, *river* . . . . . 54 28 N 42 02 E
Tura, *river* . . . . . 57 12 N 66 56 E
Uchur, *river* . . . . 58 48 N 130 35 E
Uda, *river* . . . . . 54 42 N 135 14 E
Ufa, *river* . . . . . 54 40 N 56 00 E
Ufimskoye Plateau . . . . . 56 45 N 57 30 E
Unzha, *river* . . . . . 57 43 N 43 43 E
Ural, *river* . . . . . 47 00 N 51 48 E
Ural Mountains . . . 60 00 N 60 00 E
Usa, *river* . . . . . 65 57 N 56 55 E
Ussuri, *river* . . . . . 48 28 N 135 02 E
Uy, *river* . . . . . 54 17 N 64 58 E
Vaga, *river* . . . . . 62 48 N 42 56 E
Valdai (Valday) Hills . . . . . 57 00 N 33 30 E
Vashka, *river* . . . . 64 53 N 45 47 E
Vasyuganye Swamp . . . . . 58 00 N 77 00 E
Vaygach Island . . 70 00 N 59 30 E
Velikaya, *river* . . . . 57 48 N 28 20 E
Velikaya, *river* . . . . 64 40 N 176 24 E
Verkhoyansk Mountains . . . . . 67 00 N 129 00 E
Vetluga, *river* . . . . 56 18 N 46 24 E
Vilyuy, *river* . . . . . 64 24 N 126 26 E
Vilyuy Mountains . 66 45 N 114 00 E
Vilyuy Reservoir. . . 62 48 N 111 00 E
Vishera, *river* . . . . 59 55 N 56 25 E
Vistula Lagoon . . . 54 20 N 19 30 E
Vitim, *river* . . . . . 59 26 N 112 34 E
Vitim Plateau . . . . 54 00 N 113 30 E
Volga, *river* . . . . . 45 55 N 47 52 E
Volga-Baltic Waterway . . . . . 59 58 N 30 10 E
Volga Delta . . . . . 45 55 N 47 52 E
Volga-Don Canal . 48 40 N 43 37 E
Volga Hills . . . . . 52 00 N 46 00 E
Volgograd Dam . . 48 46 N 44 26 E
Volgograd Reservoir . . . . . 50 00 N 45 25 E
Vorona, *river* . . . . 51 22 N 42 03 E
Votkinsk Reservoir . . . . . 57 10 N 55 00 E
Voykar, *river* . . . . . 65 40 N 64 39 E
Vozhe, Lake . . . . 60 40 N 39 00 E
Vrangelya, see Wrangel Island
Vyatka, *river* . . . . . 55 36 N 51 30 E
Vychegda, *river* . . 61 18 N 46 36 E
Vychegda Valley . 61 18 N 46 36 E
Vyg, Lake . . . . . 63 40 N 34 40 E
Vygozero, Lake . . 63 40 N 34 40 E
Vym, *river* . . . . . 62 13 N 50 25 E
West Siberian Plain . . . . . 60 00 N 75 00 E
Western Dvina (Daugava), *river* . 57 00 N 24 00 E
Western Sakhalin Mountains . . . . . 51 00 N 142 25 E
Western Sayan Mountains . . . . . 53 00 N 94 00 E
White, see Beloye Lake
White (Beloye) Sea . . . . . 65 30 N 38 00 E
White Sea–Baltic Canal . . . . . 64 48 N 34 48 E
Wrangel (Vrangelya) Island . . . . . 71 00 N 179 30 W
Wrangel Island Nature Reserve . 71 00 N 179 30 W
Yablonovy Range . . . . . 53 30 N 115 00 E
Yamal Peninsula . . 70 00 N 70 00 E
Yamantau, Mount . 54 15 N 58 06 E
Yana, *river* . . . . . 59 45 N 149 15 E
Yana, Gulf of . . . . 71 50 N 136 00 E
Yelizavety, Cape . 54 26 N 142 42 E
Yenisey, *river* . . . . 71 50 N 82 40 E
Yenisey Ridge . . . 59 00 N 93 00 E
Yuzhny, Cape . . . 57 45 N 156 45 E
Zeya, *river* . . . . . 50 15 N 127 35 E
Zeya Nature Reserve . . . . . 54 05 N 127 10 E
Zeya Reservoir . . . 54 30 N 127 45 E

rivers provide transport arteries from the interior to the Arctic sea route, although these are blocked by ice for long periods every year. The Siberian rivers have extremely gentle gradients—the Ob, for example, falls only 650 feet in more than 1,250 miles—as a result of which they meander slowly across immense floodplains. Owing to their northward flow, the upper reaches thaw before the lower parts, and floods occur over vast areas, leading to the development of huge swamps. The Vasyuganye Swamp at the Ob-Irtysh confluence covers 19,000 square miles.

The rest of Siberia, some 1,815,000 square miles, is drained into the Pacific. In the north, where the watershed is close to the coast, numerous small rivers descend abruptly from the mountains, but the bulk of southeastern Siberia is drained by the large Amur system. Over much of its 1,755-mile length, the Amur forms the boundary that divides Russia and China. The Amur's tributary, the Ussuri, forms another considerable length of the border.

Three drainage basins cover European Russia south of the Arctic basin. The Dnieper, of which only the upper reaches are in Russia, and the 1,160-mile-long Don flow south to the Black Sea, and a small northwestern section drains to the Baltic. By far the largest European river is the Volga. Rising in the Valdai Hills northwest of Moscow, it follows a course of 2,193 miles to the Caspian Sea. Outranked only by the Siberian rivers, the Volga drains an area of 533,000 square miles. Separated only by short overland portages and supplemented by several canals, the rivers of the Russian Plain long have been important transport arteries. The Volga system carries two-thirds of all Russian waterway traffic.

The Volga

Lakes. Russia contains an immense number of lakes. In the European section the largest are Lakes Ladoga and Onega in the northwest, with surface areas of 6,826 and 3,753 square miles, respectively; Lake Peipus (1,370 square miles) on the Estonian border; and the Rybinsk Reservoir on the Volga north of Moscow. Narrow lakes 100 to 200 miles long are located behind barrages (dams) on the Don, Volga, and Kama. In Siberia similar man-made lakes are located on the upper Yenisey and its tributary, the Angara, where the 340-mile-long Bratsk Reservoir is among the world's largest. All of these are dwarfed by Lake Baikal, the largest body of fresh water not only in Russia but also in the world. Some 395 miles in length and with an average width of 30 miles, Baikal has a surface area of 12,200 square miles and a maximum depth of 5,314 feet.

There are innumerable smaller lakes found mainly in the ill-drained, low-lying parts of the Russian and West Siberian plains, especially in their more northerly parts. Some of these reach considerable size, notably Beloye (White) Lake and Lakes Top, Vyg, and Ilmen, each occupying more than 400 square miles, in the European northwest and Lake Chany (768 square miles) in southwestern Siberia.

CLIMATE

Several basic factors determine the climates of the Russian Federation. The country's vast size and compact shape—the great bulk of the land is more than 250 miles from the sea, while certain parts lie as much as 1,500 miles away—produce a dominance of continental regimes. Its northerly latitude ensures that these are cold continental regimes: only southwestern Russia (the North Caucasus region and the lower Don and Volga basins), small sections of southern Siberia, and the maritime region of southeastern Siberia are below latitude 50° N, and more than half the federation is north of latitude 60° N. The great mountain barriers to the south and east prevent the ingress of ameliorating influences from the Indian and Pacific oceans, but the absence of relief barriers on the western and northern sides leaves the country open to Atlantic and Arctic influences. In effect there are only two seasons, winter and summer; spring and autumn are brief periods of rapid change from one extreme to the other.

Atmospheric pressure and winds. In winter the cooling of the Eurasian landmass leads to the development of an intense high-pressure cell over the interior: mean January pressures range above 1,040 millibars along the southern boundary of Siberia, from which a ridge of high pressure runs westward along Russia's borders with Kazakstan and Ukraine. Movement of air outward from these high-pressure zones ensures that winds are mainly from the southwest in European Russia, from the south over much of Siberia, and from the northwest along the Pacific coast. In summer the situation is reversed: the landmass heats up, low pressure develops over the Asian interior, and air moves inward, from the northwest in the European section, from the north in Siberia, and from the southeast along the Pacific.

Prevailing winds

Temperature. One effect of these air movements is to even out the north-south contrasts in winter temperatures, which might be expected to occur as a result of latitude. Thus, in the Russian Plain isotherms have a north-south trend, and temperatures at each latitude decline from the west toward a cold pole in northeastern Siberia. Taking reports west to east from a series of stations with a narrow latitudinal range, the January mean is 18° F (−8° C) at St. Petersburg, −17° F (−27° C) at Turukhansk in the West Siberian Plain, −46° F (−43° C) at Yakutsk, and −58° F (−50° C) at Verkhoyansk. The mean along the Mongolian border is only a degree or two above that along the Arctic coast 1,500 miles farther north. Out-blowing winds also depress temperatures along the Pacific coast: Vladivostok, at the same latitude as the French Riviera, has a January mean of 7° F (−14° C). In summer, temperatures are more closely connected with latitude; July means range from 39° F (4° C) in the Arctic islands to 68° F (20° C) along the federation's southern border. Extreme temperatures diverge greatly from these means. A world record low (outside Antarctica) is the January minimum of −96° F (−71° C) recorded at Oymyakon southeast of Verkhoyansk, while July maxima above 100° F (38° C) have occurred at several stations. The net result is a vast seasonal range that increases toward the interior. January and July means differ by a range of 52° F (29° C) at Moscow, 76° F (42° C) at Turukhansk, and 115° F (64° C) at Yakutsk. Extreme winter cold is the outstanding weather characteristic of most of Russia: the frost-free period exceeds six months only in the North Caucasus and varies with latitude from five to three months in the European section and from three months to less than two in Siberia.

Precipitation. The main characteristics of precipitation throughout Russia are the modest to low total amounts and the pronounced summer maximum. Across the plains of the European section and western Siberia the total declines from northwest to southeast. In these regions, except in a few places close to the Baltic, precipitation generally remains below 24 inches (600 millimetres), falling from 21 inches at Moscow to about 8 inches along the Kazak border. In eastern Siberia totals are generally less than 16 inches and as little as 5 inches along the Arctic coast. Precipitation increases again along the Pacific (24 inches in Vladivostok), where the moisture-laden onshore summer monsoon brings a precipitation maximum. Amounts vary with altitude: the higher parts of the Urals receive more than 28 inches, and the mountains of Kamchatka and the Sikhote-Alin receive well over 40 inches. Snow is a pronounced feature over the whole of Russia, its depth and duration having important effects on agriculture. The duration of snow cover varies with latitude as well as altitude, ranging from 40 to 200 days across the Russian Plain and from 120 to 250 in Siberia.

SOILS AND PLANT AND ANIMAL LIFE

Climate, soils, vegetation, and animal life are closely interrelated, and variations among these within Russia form a series of broad latitudinal environmental belts that sweep across the country's plains and plateaus from the western border to the Lena River. In the mountain zones of the south and east the pattern is more complex because altitude rather than latitude is the dominant factor, and there are striking changes over relatively short distances. Within Russia there are six main environmental belts (some with subdivisions): Arctic desert, tundra, taiga, mixed and deciduous forest, wooded steppe, and steppe.

Environmental belts

Arctic desert. Arctic desert—confined to the islands of Franz Josef Land, much of the Novaya Zemlya and

Severnaya Zemlya archipelagoes, and the New Siberian Islands—is completely barren land with little or no vegetation. Considerable areas are ice-covered.

Tundra. Almost 10 percent of Russian territory is tundra, a treeless, marshy plain. Occupying a narrow coastal belt in the extreme north of the European Plain, the tundra widens to a maximum of about 300 miles in Siberia. Tundra soils are extremely poor: low temperatures result in a moisture surplus so that they are poorly drained, and the limited and discontinuous vegetation cover provides little organic matter; moreover, this matter decays only slowly, and the soils are highly acidic. Tundra soils are frozen for much of the year, and during the summer thaw drainage is inhibited by the presence of permafrost beneath the thawed surface layer. A typical tundra soil has a shallow surface layer of raw humus, beneath which there is a horizon (soil layer) of gley (sticky, clayey soil) resting on the permafrost. Vegetation changes from north to south, and three subdivisions are recognized: Arctic tundra, with much bare ground and extensive areas of mosses and lichens; shrubby tundra, with mosses, lichens, herbaceous plants, dwarf Arctic birch, and shrub willow; and wooded tundra, with more extensive areas of stunted birch, larch, and spruce. There are considerable stretches of sphagnum bog. Apart from reindeer, which are herded by the indigenous population, the main animal species are the Arctic fox, musk ox, beaver, lemming, snowy owl, and ptarmigan.

Taiga. South of the tundra the vast taiga (boreal forest) zone, the largest of the environmental regions, occupies the Russian and West Siberian plains north of 56°–58° N together with most of the territory east of the Yenisey. A distinction is commonly made between the western taiga, where the climate is less extreme, and the eastern taiga beyond the Yenisey. In the western section, dense forests of spruce and fir in moister areas alternate with pine, shrubs, and grasses on the lighter soils. These species also are present in the east, but the larch becomes dominant. Only small areas have been cleared for agriculture, mainly in the European part, and the taiga remains the world's largest timber reserve. Coniferous forest cover is, however, by no means continuous: there are large stands of birch, alder, and willow and, in poorly drained areas, huge stretches of swamp and peat bog. The taiga is rich in fur-bearing animals, such as the sable, squirrel, marten, fox, and ermine, and is the home of the elk, bear, muskrat, and wolf.

Podzol

Throughout the taiga zone the dominant soil type is the podzol, a product of the intense leaching characteristic of this area of moisture surplus. The forest vegetation provides a surface layer of highly acidic raw humus that decomposes slowly, producing humic acids. Percolating downward, acidic groundwater removes iron and calcium compounds from the upper layers, which, as a result, are pale in colour and acidic. Soluble materials are redeposited at lower levels, often resulting in an iron-rich hardpan that impedes the drainage of the upper horizons so that gley podzols form. For agriculture to be successful, applications of lime and fertilizer are required.

Mixed and deciduous forest. As conditions become warmer with decreasing latitude, deciduous species appear in greater numbers and eventually become dominant. The mixed and deciduous forest belt is triangular in shape, widest along the western border and narrowing toward the Urals. Oak and spruce are the main trees, but there also are growths of ash, aspen, birch, elm, hornbeam, maple, and pine. East of the Urals as far as the Altai, a narrow belt of birch and aspen woodland separates the taiga from the wooded steppe. Much of the mixed and deciduous forest zone has been cleared for agriculture, particularly in the European section. As a result, the wildlife is less plentiful, but the roe deer, wolf, fox, and squirrel are common. Soils also show a north-south gradation. As the moisture surplus diminishes, leaching becomes less intense, and true podzols give way to gray and brown forest soils, which are less acidic and have a much greater organic content and a higher natural fertility. A second zone of mixed forest occurs in the Amur-Ussuri-Zeya lowlands of the Far East and includes Asiatic species of oak, hornbeam, elm, and hazel.

Wooded steppe and steppe. The southward succession is continued by the wooded steppe, which, as its name suggests, is transitional between the forest zone and the steppe proper. Forests of oak and other species (now largely cleared for agriculture) in the European section and birch and aspen across the West Siberian Plain alternate with areas of open grassland that become increasingly extensive toward the south. The wooded steppe eventually gives way to the true steppe, which occupies a belt some 200 miles across and extends from southern Ukraine through northern Kazakstan to the Altai Mountains. Russia has a relatively small share of the Eurasian Steppe, mainly in the North Caucasus and lower Volga regions, though pockets of wooded steppe and steppe also occur in basins among the mountains of southern Siberia.

The natural steppe vegetation is composed mainly of turf grasses such as bunchgrass, fescue, bluegrass, and agropyron (a type of bunchgrass). Perennial grasses, mosses, and lichens also grow on the steppe, and drought-resistant species become common in the south, where the sequence continues in Kazakstan through dry steppe and semidesert to the great deserts of Central Asia. Woodland is by no means wholly absent, occurring in damper areas in river valleys and depressions. Much of the steppe vegetation, particularly in the west, has been replaced by grain cultivation. The absence of natural shelter on the open steppe has conditioned the kind of animals that inhabit it. Typical rodents of the zone include the marmot and other such burrowing animals and various mouse species. The skunk, fox, and wolf are common, and antelope inhabit the south. The most common birds are bustards, eagles, kestrels, larks, and gray partridges.

Effect of absence of natural shelter

The distinctive soil of the steppe is chernozem (black earth). This soil type takes its name from the very dark upper horizon, often more than three feet thick, which is rich in humus derived from the thick grass cover. Winter frost and summer drought inhibit the decomposition of organic matter, and high evaporation rates prevent leaching so that humus accumulates. Calcium compounds are leached downward by the spring snowmelt but are drawn upward in summer and become concentrated in a lime-rich horizon beneath the humus layer. Low acidity and a high humus content combine to give the chernozems a high natural fertility, which has helped make the steppe the country's main source of grain.

SETTLEMENT PATTERNS

A major trend in Russia from the 1890s to the late 20th century was the movement of people from the European portion into Siberia and the Far East, those latter regions constituting three-fourths of the nation's territory but containing only about one-fifth of its population. The bulk of the population lives in the main settled belt of European Russia, which includes the southern fringes of the taiga and the whole of the mixed and deciduous forest, wooded steppe, and steppe zones. Rural densities in this section range from 25 to 250 persons per square mile, the highest concentrations occurring in the wooded steppe. East of the Urals, across the southern part of the West Siberian Plain, rural densities are considerably lower, rarely exceeding 65 persons per square mile. Beyond the Yenisey the settled zone breaks up into a series of pockets in the extreme south, along the line of the Trans-Siberian Railroad, of which the largest is that in the Amur-Ussuri-Zeya lowlands of the Far East. In the second half of the 20th century, rural depopulation has been a pronounced characteristic in the European section, where rural population fell by nearly 30 percent between 1960 and 1990. Many rural areas now are inhabited mainly by the elderly, and at the local level a natural population decrease sometimes occurs.

The bulk of the rural population lives in large villages associated with the collective and state farms (*kolkhozy* and *sovkhozy*) established by the erstwhile Soviet regime. These farms have carried on the long-established Russian tradition of communal farming from nucleated settlements. North of the main settled belt there are vast stretches of thinly settled and empty territory. Typical of this zone is Sakha (Yakutia), a minority republic that, with an area of about 1,200,000 square miles and about



Street scene in Khabarovsk, a major transportation hub in the Russian Far East.
© Bryan and Cherry Alexander

1,000,000 inhabitants, has a density of less than one person per square mile.

Urbanization

Since the mid-19th century, in association with massive economic development, there has been an enormous increase in the number of city dwellers. Some three-fourths of Russia's population is now classified as urban. Russia has about 160 towns with more than 100,000 inhabitants, of which roughly half have more than 250,000 and some dozen have more than 1,000,000. Moscow, the largest metropolis, has twice the population of its nearest rival, St. Petersburg, which, in turn, is at least three times the size of any of Russia's other major cities: Chelyabinsk, Kazan, Nizhny Novgorod (formerly Gorky), Novosibirsk, Omsk, Perm, Rostov-na-Donu, Samara (formerly Kuybyshev), Ufa, Volgograd, and Yekaterinburg (formerly Sverdlovsk). A number of major urban concentrations have developed in the main industrial regions. St. Petersburg (the tsarist capital) stands alone as the northernmost metropolis, whereas Moscow and Nizhny Novgorod are part of the large urbanized central industrial region, which has a score of large cities, numerous smaller towns, and an urban population that is about one-fifth the total population of Russia. In the Ural Mountains region the towns are more widely spaced and include numerous small mining and industrial centres as well as a number of towns with more than 250,000 inhabitants, altogether amounting to an urban population about half that of the Moscow region. The only slightly less populous Volga region has towns strung out along the riverbanks, with a particularly dense concentration in the vicinity of Samara. European Russia also includes a portion of the Donets Basin (Donbass) industrial zone, arbitrarily split by the Russia-Ukraine boundary; the largest city is Rostov-na-Donu, but there are numerous smaller centres.

East of the Urals the main urban concentration is in the Kuznetsk Basin (Kuzbass), with its mining and industrial complex. Major cities also occur at widely separated points along the length of the Trans-Siberian Railroad, including, from west to east, Omsk, Novosibirsk, Krasnoyarsk, Irkutsk, Ulan-Ude, Chita, Khabarovsk, and Vladivostok. A few very isolated cities are located in the far north, notably the ports of Murmansk and Arkhangelsk and mining centres such as Vorkuta and Norilsk. Resort towns are a feature of the North Caucasus region, including Sochi (on the Black Sea), Pyatigorsk, and Mineralnyye Vody. Elsewhere the capitals of provinces and other administrative divisions are the main towns, having grown to considerable size as the organizing centres for their territories.

The people

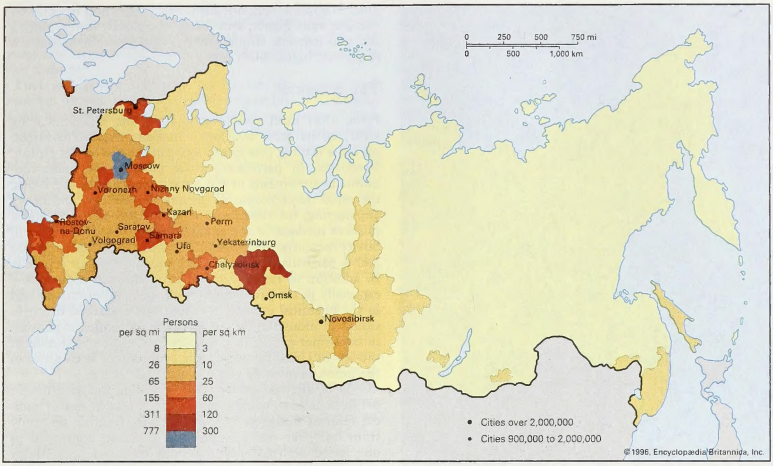
ETHNIC AND LINGUISTIC COMPOSITION

Of all of the 15 former union republics of the U.S.S.R., Russia displays the greatest ethnic diversity, with censuses recognizing more than 70 distinct nationalities. Many of these are extremely small—in some cases consisting of only a few thousand individuals—and, in addition to Russians, only a handful of groups have more than a million members each: Tatars, Ukrainians, Chuvash, Bashkir, Belarusians, and Mordvins. Russians, the overwhelming majority, constitute about four-fifths of the total. The multiplicity of peoples is reflected in 21 minority republics, and, within the Russian republic, there are 10 autonomous districts and an autonomous region. In most of these divisions, the eponymous nationality (which gives its name to the division) is outnumbered by Russians. Since the early 1990s, ethnicity has underlain numerous conflicts within and between these units; many have demanded more autonomy and, in a few cases, even complete independence. Those parts of the republic that do not form autonomous territories are divided into 6 regions and 49 provinces.

Language groups

Linguistically, the population of Russia can be divided into the Indo-European group, comprising East Slavic speakers and smaller numbers speaking several other languages; the Altaic group, including Turkic, Mongolian, and Manchu-Tungus; the Uralic group, including Finno-Ugric and Samoyedic; and the Caucasian group, comprising Abkhazo-Adyghian and Nakho-Dagestanian.

The Indo-European group. East Slavs—mainly Russians but including some Ukrainians and Belarusians—are about 85 percent of the total population and are prevalent throughout the federation. The Slavs emerged as a recognizable group in eastern Europe between the 3rd and 8th centuries AD, and the first Slav state, Kievan Rus, arose in the 9th century. Following the Mongol invasions, the centre of gravity shifted to Moscow, and the Russian Empire



Population density of Russia.

expanded to the Baltic, Arctic, and Pacific, numerically overwhelming the indigenous peoples. Despite its wide dispersal, the Russian language's main features are homogeneous throughout Russia. Indo-Iranian speakers include the Ossetes of the Caucasus. In addition, there are sizable contingents of German speakers (mainly in southwestern Siberia) and Jews (mainly in European Russia); the numbers of both groups have declined through emigration.

The Altaic group. Turkic speakers dominate the Altaic group. They live mainly in the Central Asian republics, but there is an important cluster of Turkic speakers between the middle Volga and southern Urals, comprising the Bashkir, Chuvash, and Tatars. A second cluster, in the North Caucasus region, includes the Balkar, Karachay, Kumyk, and Nogay. There also are numerous Turkic-speaking groups in southern Siberia between the Urals and Lake Baikal: the Altai, Khakass, Shor, Tofalar (or Karagasy), and Tuvans (who inhabit the area once known as Tannu Tuva, annexed by the U.S.S.R. in 1944). The Yakut live mainly in the middle Lena basin, the Dolgan in the Arctic.

Manchu-Tungus languages are spoken by the Evenk, Even, and other small groups who are widely dispersed throughout eastern Siberia. The Buryat, who live in the Lake Baikal region, and the Kalmyk, to the west of the lower Volga, speak Mongolian tongues.

The Uralic group. Widely disseminated in the Eurasian forest and tundra zones, the Uralic group has complex origins. Finnic peoples inhabit the European section: the Mordvins, Mari (formerly Cheremis), Udmurt (Votyak) and Komi (Zyryan), and the closely related Komi-Permyaks live around the upper Volga and in the Urals, while Karelians, Finns, and Veps inhabit the northwest. The Mansi (Vogul) and Khanty (Ostyak) are spread thinly over the lower Ob basin.

The Samoyedic group also has few members dispersed over a vast area: the Nenets in the tundra and forest tundra from the Kola Peninsula to the Yenisey, the Selkup around the middle Ob, and the Nganasan mainly in the Taymyr Peninsula.

The Caucasian group. The main Caucasian groups are located south of the Caucasus Mountains, but there are numerous smaller ones in the North Caucasus region of Russia. Abaza, Adyghian, and Kabardian (Circassian) are similar languages but differ sharply from the languages of the Nakh group (Chechen and Ingush) and of the Dagestanian group (Avar, Lezgian, Dargwa, Lakk, Tabasaran, and a dozen more).

Other groups. In far eastern Siberia are several Paleo-Siberian groups that share a common mode of life but differ linguistically. The Chukchi, Koryak, and Itelmen (Kamchadal) belong to a group known as Luorawetlan, which is distinct from the Eskimo-Aleut group. The languages of the Nivkh (Gilyak) along the lower Amur and on Sakhalin Island, of the Yukaghir of the Kolyma Lowland, and of the Ket of the middle Yenisey are completely isolated.

Paleo-Siberian groups

RELIGION

While the ethnic distinctiveness of the various peoples of Russia has long contained a religious element, the position of religious organizations and of their individual adherents has varied with political circumstances. The constitution of the former Soviet Union guaranteed religious freedom, but religious activities were greatly constrained, and membership in religious organizations was held to be incompatible with membership in the Communist Party. More open expression of Christian beliefs was permitted during World War II when the government sought the support of believers in the fight against fascism, but restrictions were reimposed when the war ended. With the declaration of *glasnost* ("openness") under the Soviet premier Mikhail Gorbachev, and especially since the dissolution of the U.S.S.R., religious freedom has become a reality, and the continuing adherence of large sections of the population to a variety of faiths has been revealed. Indeed, emerging Russian nationalism has again identified the Russian Orthodox church as a major element of Russian culture.

By far the largest denomination is Russian Orthodoxy. Its origin is usually dated to the closing years of the 10th century, when the ruler of Kievan Rus was converted to Christianity by missionaries from Byzantium, and it has remained the dominant faith of the Russian people for 1,000 years. Other Christian denominations are much smaller and include the Old Believers, who separated from Orthodoxy in the 17th century, and Baptist and Evangelical groups, which have grown some in membership during the 20th century. Catholics, both Western rite (Roman) and Eastern rite (Uniate), and Lutherans were numerous in the former U.S.S.R., but mainly outside the present Russian republic, where there are few adherents.

Religious affiliation does not entirely correspond to linguistic groupings. The Slavs are overwhelmingly Orthodox Christian. Turkic speakers are predominantly Muslim, although several Turkic groups in Russia are not: the Chuvash are mainly Christian; the Altai, Khakass, and Tuvans



Church officials being greeted outside the Trinity–St. Sergius monastery, the centre of Russian Orthodoxy, in Sergiyev Posad, north of Moscow.
© Boyd Norton

are Buddhists; and Turkic speakers east of the Yenisey have retained their shamanistic beliefs, although some have converted to Christianity. The Mongolian speakers, the Buryat and Kalmyk, are Buddhists.

Jews have been recognized as an ethnolinguistic group rather than a religious one. They suffered persecution under Stalin and during the Nazi occupation. Since the late 1980s, emigration to Israel and elsewhere has been permitted on an increasing scale, and the number of Jews living in the former U.S.S.R. has decreased. Prior to the breakup of the U.S.S.R., about one-third of the Soviet Jews lived in the Russian republic; by no means all of these practiced Judaism.

DEMOGRAPHY

Rate of population growth

The Russian Federation as a whole now has a rate of population growth well below that of previous decades, owing mainly to the decline in the birth rate of the Russian majority. Higher rates of natural increase continue among some of the minority peoples, particularly those of Islāmic background. Migration from the European sector to Siberia and the Far East is the main factor behind regional variations in population growth rates. Between the 1979 and 1989 censuses, when the then Soviet republic's population increased by about 7 percent, growth exceeded 15 percent in Siberia and the Far East but reached barely 5 percent in the Central and Ural regions and less than 2 percent in the Volga-Vyatka region.

The long-declining Russian birth rate has led to a progressive aging of the population, and this, in turn, has produced an increase in the crude death rate. A declining one-fourth of the total population of the Russian Federation is below the age of 15, while the group aged 60 or older is approaching one-fifth. The proportion of children was generally higher, and that of the elderly lower, among the non-Russian ethnic groups, who have maintained a somewhat higher birth rate.

The economy

From 1917 until 1991 Russia (then the R.S.F.S.R.), by virtue of its great size and abundant natural resources, played a leading role in the economy of the Soviet Union. Resources, in particular, made possible the main economic achievements of the Soviet regime until the 1950s: the rapid development of mining, metallurgy, and heavy engineering, the expansion of the railway network, and a massive increase in the energy supply. A second phase in Soviet industrial development began in the 1960s and also had a particularly strong effect on the Russian republic. In addition to further growth in established industries—especially in the production of oil, gas, and electricity and in the chemical industries—there was a marked diversification in industrial output, including a limited expansion in consumer goods. In the years before the dissolution of the Soviet Union, however, the economy of Russia and of the union as a whole was in a state of decline.

Reforms

In 1992, after the collapse of the union, the government of the Russian Federation implemented a series of radical reforms designed to transform the Russian economy from one that was centrally planned and controlled to one based on free enterprise and market forces. Major components of the reforms included establishing privately owned industrial and commercial ventures, with foreign as well as Russian investment, and privatizing state-owned enterprises. Vouchers were issued to each Russian citizen that were to be used to purchase shares in firms being privatized, which often were sold at auction; in practice, these vouchers frequently were sold for cash and were accumulated by entrepreneurs. A commodity- and stock-exchange system also was set up.

The privatization process was slow, however, and many firms—particularly in the heavy industries—remained under state ownership. In rural areas, land-reform legislation offered farmers the opportunity to gain title to parcels of land. This conversion also was slow, as many clung to the old, familiar system.

At the outset, these measures caused considerable hardships for the average Russian citizen. The monetary system was in disarray: the removal of price controls caused a huge escalation in inflation and prices, the value of the ruble plummeted, and real incomes fell dramatically. Industrial and agricultural production declined, and unemployment rose, causing shortages of consumer goods and (in urban areas) foodstuffs. These difficulties were reduced to some extent by the rapid development of a large informal (*i.e.*, black market) economy. A result of these trends was the appearance of a widening gap between the successful entrepreneur and average worker. Conditions began to improve by the mid-1990s, but both production and consumption remained below the peak levels achieved under the Soviet regime.

In addition to these difficulties, Russia and other republics had been subjected to serious, long-term environmental degradation during the Soviet period, the full extent of which became apparent only in the 1990s. The most visible aspects of this situation—such as the Chernobyl accident and its aftermath and the drastic reduction in the water volume of the Aral Sea through inflow diversions—were only symptomatic of decades of wasteful agricultural practices and resource exploitation and of widespread industrial pollution. Addressing environmental concerns placed another burden on Russia's already overwhelmed economic structure.

The economic foundation of the country itself remained similar to that which had been developed during the Soviet period. For purposes of description, it is convenient to refer to the official set of 11 economic regions into which Russia is divided. In Europe the regions are the North, Northwest, Central, Volga-Vyatka, Central Black Earth, North Caucasus, Volga, and Ural, and in Asia they are West Siberia, East Siberia, and the Far East.

RESOURCES

Russia has enormous energy resources and deposits of many different minerals. Most, if not all, of the raw materials required by modern industry are found within the country.

Fuel and power. Russia has by far the largest coal reserves among the former Soviet republics. The biggest fields lie in the remote Tunguska and Lena basins of East Siberia and the Far East, but these are largely untapped, and the bulk of output comes from more southerly fields along the Trans-Siberian Railroad. About three-fourths of Russia's coal is produced in Siberia—some two-fifths from the Kuznetsk Basin alone and the remainder from the Kansk-Achinsk, Cheremkhovo, and South Yakut basins and numerous smaller sources. The production of hard coal in the European section is mainly in the eastern Donets Basin and, in the Arctic, in the Pechora Basin around Vorkuta; the large Moscow Basin (entirely) and the small Urals fields (mainly) are sources of lignite.

Siberian oil fields

The Russian Federation is one of the world's leading producers of oil and natural gas. The great bulk of the supply comes from the huge fields that underlie the northern part of the West Siberia region. Another significant source is from the Volga-Ural zone, and the remainder is derived mainly from the Komi-Ukhta field (North region); the North Caucasus region, once the U.S.S.R.'s leading producer, is now of little importance. Extensive pipeline systems link the producing districts to all regions of the federation, the neighbouring former Soviet republics, and, across the western frontier, numerous European countries.

Much of the fuel produced in Russia is converted to electricity, about three-fourths of which is generated in thermal stations; some two-thirds of thermal generation is from oil and gas. The remaining power output is produced by hydroelectric and nuclear plants. Most of the hydroelectricity comes from huge stations on the Volga, Kama, Ob, Yenisey, Angara, and Zeya rivers. Nuclear power production expanded rapidly before development was checked by the Chernobyl accident. Much of Siberia's electricity output is transmitted to the European region along high-voltage lines.

Other minerals. Russia produces large quantities of iron ore, mainly from the Kursk Magnetic Anomaly (Central Black Earth region), Kola Peninsula, Urals, and Siberia. The largest steel-producing plants are located mainly in the Urals, Central Black Earth region, and Kuznetsk Basin, but there is some steel production in every economic region.

Nonferrous metals are available in great variety from many districts, but by far the most important are those of the Ural region, which is the republic's main centre of nonferrous metallurgy. Russia is a major producer of cobalt, chrome, copper, gold, lead, manganese, nickel, platinum, tungsten, vanadium, and zinc. The country produces much of its aluminum from plants powered by the Siberian hydroelectric stations, but bauxite deposits are relatively meagre.

INDUSTRY

Machine building. Russia's machine-building industry provides most of the federation's requirements for steam boilers and turbines, electric generators, grain combines, automobiles, and electric locomotives and fills much of its demand for machine tools, instruments, and automation components. Important automobile factories are located in Moscow, Nizhny Novgorod, Miass, Yaroslavl, Ulyanovsk (formerly Simbirsk), and Izhevsk; the largest plants include the Zhiguli works at Tolyatti (near Samara) and the heavy truck factory at Naberezhnye Chelny (in Tatarstan).

Chemicals. Because of the complex history of the development of the chemical industries and the great variety of raw materials involved, chemical manufacture is widely dispersed. The industry initially utilized mineral salts, coke-oven and smelter gases, timber, and foodstuffs (mainly potatoes) as their raw materials. On this basis, synthetic rubber factories were built in the Central Black Earth and Central regions, areas of large-scale potato production; sulfuric acid plants were developed in the Urals and North Caucasus, where there was nonferrous metallurgy; and potassium and phosphatic fertilizer plants were constructed at sites in several regions, near deposits of potassium salts and phosphorites.

Since the end of the 1950s the massive increase in oil and gas output has provided new raw chemical materials and lessened the dependence on traditional resources. New chemical plants have been built both in the oil- and gas-producing areas of the Volga-Ural and North Caucasus zones and in other regions at points served by pipelines. Chemical industries requiring large quantities of electric power, such as those based on cellulose, are particularly important in Siberia, where both timber and electricity are plentiful.

Textiles

Light industry. Russia's textile industries are heavily concentrated in the European sector, especially in the Central region, which produces a large share of the federation's clothing and footwear. The dominant branch is cotton textiles, with the raw cotton coming mainly from the Central Asian states. In the zone between the Volga and Oka rivers, east of Moscow, there are numerous cotton textile towns, the largest of which are Ivanovo, Kostroma, and Yaroslavl. Durable consumer goods—refrigerators, washing machines, radios and television sets, and the like—are produced primarily in areas with a tradition of skilled industry, notably in and around Moscow and St. Petersburg.

AGRICULTURE, FORESTRY, AND FISHING

Agriculture. The harshness of the Russian environment is reflected in the small proportion of land that is used for farming. Agricultural land constitutes less than one-sixth of the republic's territory. About three-fifths of Russian farmland is used to grow crops; the remainder is devoted to pasture and meadow. The main product of Russian farming has always been grain, which occupies considerably more than half of the cropland. Wheat is the chief cereal, followed by barley, rye, and oats. More than one-third of the sown area is devoted to fodder crops—sown grasses, clovers, root crops, and, in the southern districts, corn (maize). The remaining farmland is devoted to industrial crops, such as sunflowers, sugar beets, and flax, and to potatoes and other vegetables.

There are pronounced regional variations in response to the variations in relief, soil, and climate. In the European section of Russia the proportion of land devoted to crops increases southward, from virtually none in the North to about two-thirds in the Central Black Earth region. In Siberia and the Far East, crops are largely confined to the southern fringe. Even in West Siberia, where the cultivated zone is at its widest, crops occupy less than one-tenth of the region's territory, this proportion falling to negligible levels in East Siberia and the Far East. Cereals occupy more than two-thirds of the cropland in most regions but less than half in the damper Northwest and Central regions, where fodder crops and livestock are more important. The intensity of farming and the yields achieved are generally much higher in the European section than in Siberia. This is also true of livestock farming.

Forestry and wood pulp. The lumbering, pulp, paper, and woodworking industries are particularly important because Russia contains the world's largest forest reserves. Coniferous species are predominant, and the republic produces more than one-fifth of the world's softwood. In the period since World War II these activities have spread increasingly to the east of the Urals.

Fishing. The fishing industry plays a significant role in the Russian economy. With access to the substantial resources of both the Atlantic and Pacific oceans, the country has especially developed marine fishing, notably its fleet of factory ships that can process huge catches at remote locations. The chief European ocean-fishing ports are Kaliningrad and St. Petersburg on the Baltic Sea and Murmansk and Arkhangelsk in the far north. Russia's chief Pacific port is Vladivostok, but there are several others, particularly in Sakhalin and Kamchatka. Smaller-scale fishing takes place in the Sea of Azov and the Black and Caspian seas (the Caspian sturgeon is the source of the world's finest caviar), but reduced river flows and pollution from agricultural runoff, industrial waste, and

Ocean-fishing ports



The Baikal-Amur railway running through Siberia.
© Tass/Sovfoto

sewage dumping have reduced fish populations. There are important inland fisheries on lakes and rivers, including a good deal of fish farming.

TRADE

As part of the U.S.S.R., the Russian republic traded extensively with the other Soviet republics, from which it "imported" a variety of commodities that it was unable to produce in sufficient quantities itself. These included cotton (from Central Asia) and other high-value agricultural products, grain (mainly from Kazakstan), and various minerals. In return Russia "exported" oil and gas to republics with a weak energy base such as Belorussia (now Belarus) and the Baltic states and sent its skilled engineering products and consumer goods to most of its partners. By the mid-1990s, trade relations among the former union republics had not been established in any systematic manner, one problem being agreement on the prices to be charged for goods exchanged in place of the artificially low ones that prevailed during the Soviet period. It was clear that the former union republics remained heavily interdependent and that some kind of free-trade grouping was necessary if the economies of the new states were to flourish. A move in this direction occurred in late 1993, when Russia and nine other republics signed a treaty of economic union.

International trade by the U.S.S.R. remained at a rather low level until the 1960s, most of it being based on bilateral and multilateral arrangements with the Council for Mutual Economic Assistance (Comecon) states. As Soviet economic expansion slowed during the 1970s and '80s, it became apparent that further growth required large quantities of high-tech equipment from the West. To finance these imports, increasing amounts of hard currency were needed, and this could be obtained only by increasing exports to the West. In this expanding trade, oil and gas were of particular importance.

With the collapse of Comecon and the dissolution of the Soviet Union itself, individual republics began to develop their own trading relations with the outside world. Russia, with its large resources of oil, gas, and minerals, seemed well placed to continue the type of trading relations with the West already developed by the former Soviet Union. In June 1994 Russia signed an agreement that strengthened economic ties with the European Union.

TRANSPORTATION

Russia's vast size and the great distances that often separate sources of raw materials and foodstuffs from consuming areas place a heavy burden on the transport system. One result has been the continuing dominance of the railways, which account for about 90 percent of the country's freight turnover (60 percent if pipelines are included) and half of all passenger movement. Nevertheless, the rail network is a very open one, and its density varies regionally: highest in the Northwest, Central, and Central Black Earth regions; lowest in Siberia and the Far East. Indeed, east of the Urals the term "network" is a misnomer, since the system consists of only a few major trunk routes (*e.g.,* the Trans-Siberian Railroad and Baikal-Amur Mainline) with feeder branches to sites of economic importance. Russian railways are among the world's leading freight carriers, the line from the Kuznetsk to the Urals being especially prominent. Railways

Apart from highways linking the major cities of European Russia, the road system is underdeveloped and carries only a tiny fraction of the freight. A much greater volume, in fact, is carried by inland waterways. Although the greatest volume is carried on the Volga system, river transport is most vital in areas devoid of railways.

In addition to its vital role in foreign trade, maritime transport also has some importance in linking the various regions of Russia, particularly those that face the Arctic seaboard. Traffic on the Arctic Ocean route is seasonal.

Air transport plays an increasingly important role. Russian airlines carry only a minute fraction of all freight, chiefly high-value items to and from the remote parts of Siberia, where aircraft are sometimes the only means of transport. Airlines are responsible for nearly one-fifth of all passenger movement. Aeroflot, the largest airline in the world, formerly the state airline of the Soviet Union, carries more than 80 million passengers a year.

Administration and social conditions

GOVERNMENT

Government during the Soviet era. Prior to the events that brought about the dissolution of the Soviet Union, the R.S.F.S.R., like the other 14 union republics, was subject to a series of Soviet constitutions (1918, 1924, 1936, 1977). Until the late 1980s the whole structure of Soviet

government was dominated at all levels by the Communist Party, which was all-powerful and whose head was the country's de facto leader.

Under the Soviet constitutions, the R.S.F.S.R. nominally was a sovereign socialist state. A socialist economic system and socialist ownership of the means of production, distribution, and exchange were the economic basis of the republic. Even before the R.S.F.S.R. had been established, the Second All-Russian Congress of Soviets had created a Soviet republic out of the former Russian Empire. The Third All-Russian Congress of Soviets announced the creation of the R.S.F.S.R. in July 1918. During and after the Civil War (1918–20) the R.S.F.S.R. was organized to include autonomous soviet socialist republics (A.S.S.R.'s). On Dec. 30, 1922, the R.S.F.S.R. and the other republics entered the formal federation of the Union of Soviet Socialist Republics (U.S.S.R.), whose constitution was ratified in 1924. Later events were greatly influenced by that document's recognition of each republic's right to secede from the U.S.S.R.

Minority Republics, Regions, and Districts of the Russian Federation

Minority republics	Sakha (Yakutia)
Adygea	Tatarstan
Altay (Gorno-Altay)	Tuva (Tyva)
Bashkortostan	Udmurtia
Buryatia	
Chechnia	**Autonomous oblast (region)**
Chuvashia	Yevreyskaya (Jewish)
Dagestan	
Ingushetia	**Autonomous *okruga* (districts)**
Kabardino-Balkaria	Aga-Buryat
Kalmykia	Chukchi (Chukotka)
(Khalmg Tangch)	Evenk
Karachay-Cherkessia	Khanty-Mansi
Karelia	Komi-Permyak
Khakassia	Koryak
Komi	Nenets
Mari El	Taymyr (Dolgano-Nenets)
Mordvinia	Ust-Orda Buryat
North Ossetia (Alania)	Yamalo-Nenets

The 1936 constitution

The 1936 constitution, adopted in 1937, established a federal structure that lasted for about 50 years, despite changes introduced in 1978. The R.S.F.S.R. was politically divided into A.S.S.R.'s, autonomous *oblasti* (provinces), and *okruga* (districts)—which were the homelands of the more important non-Russian nationalities—as well as provinces and *kraya* (regions), the latter division including the autonomous provinces and districts (see *The people*). All these divisions were represented in the legislative branch: the Supreme Soviets of the U.S.S.R. and of the R.S.F.S.R., both of which were bicameral (a Soviet of the Federation and a Soviet of Nationalities). The autonomous republics enjoyed some political independence, each with its own constitution, executive, legislature, and judiciary; each sent deputies to the Supreme Soviets of both the U.S.S.R. and the R.S.F.S.R. Autonomous provinces and districts also sent deputies to these bodies, but their autonomy was limited. The division of Russia into these types of political units and their representation in central government were retained after independence.

Under this system the highest organ of state power in the republic was the Supreme Soviet of the R.S.F.S.R., which appointed the Council of Ministers as its highest executive and administrative organ. Each autonomous republic also had its Supreme Soviet and Council of Ministers, while local soviets were responsible for the affairs of their provinces, districts, regions, *rayony* (sectors), cities, and other localities. All these bodies were subject to decisions made by the Soviet central government.

Until 1988, elections at all levels were from a single slate of candidates, the great majority of whom were in effect chosen by the Communist Party.

The transitional period. From the late 1980s through 1991—the period of Mikhail Gorbachev's *perestroika* ("restructuring"), *glasnost* ("openness"), and *demokratizatsiya* ("democratization") reform policies—fundamental changes took place in the political system and government structures of the U.S.S.R. that altered both the nature of the Soviet federal state and the status and powers of the individual republics. The first major change consisted of legislation passed by the old Supreme Soviet in 1988 that created a new body, the Congress of People's Deputies of the U.S.S.R., together with a Congress of People's Deputies in each republic. Elections to these bodies were held in 1989. For the first time, voters were presented with a choice of candidates, including non-Communists, although the system was such that party members remained a strong element: one-third of the deputies came from territorial constituencies, another third were from national territorial constituencies, and the remaining third from organizations such as the Communist Party, trade unions, and professional bodies.

Thereafter, the pace of change accelerated, culminating in 1991. In June 1990 the Congress of the R.S.F.S.R. proclaimed that Russian laws took precedence over Soviet laws. In April 1991 the popularly elected post of president of the Russian Federation was created to head the executive branch, and in June Boris Yeltsin became the republic's first democratically elected president. An abortive coup in August by hard-liners opposed to Gorbachev's reforms led to the collapse of most U.S.S.R. government organizations, the abolition of the Communist Party's leading role in government, and the dissolution of the party itself. Republic after republic declared its "sovereignty," and in December the U.S.S.R. was formally dissolved. Concurrently, the Russian Federation and 10 other former Soviet republics established the Commonwealth of Independent States (CIS), which was to replace the Soviet Union with a more loosely structured federation. Russia played a leading role in the creation of the CIS and has maintained its status as the dominant member.

Dissolution of the U.S.S.R.

Government of the Russian Federation. The structure of the new Russian government differed significantly from that of the government of the former Soviet republic. It has been characterized by a power struggle between the executive and legislative branches, primarily over issues of constitutional authority and the pace and direction of democratic and economic reform. Conflicts came to a head in September 1993 when President Yeltsin dissolved the Russian parliament (the Congress of People's Deputies and the Supreme Soviet). Parliamentary members and their allies revolted and were suppressed only through military intervention.

In December 1993 a new constitution proposed by Yeltsin was approved in a nationwide referendum, and representatives were elected to a new legislature. The new constitution, effective Dec. 22, 1993, provides the president with significantly increased powers. The president appoints the prime minister and key judges as well as cabinet members and may override and even dissolve the legislature in some cases. The new parliament, the Federal Assembly, consists of the Federation Council (a 178-member upper house with equal representation for all 89 republics and regions) and a State Duma (a 450-member lower house elected through proportional representation on a party basis and through single-member constituencies). The new constitution provides for welfare protection, access to social security, pensions, free health care, and affordable housing. At the same time, it considerably reduced the status of the regions and made all regions subject to central authorities. In 2000 the Russian central government assigned each of the country's provinces and autonomous republics, regions, and districts to one of seven new federal districts. These districts were intended to enable the central government to control and monitor these constituent units, and they confirmed the supremacy of the central government over the regions.

Judiciary. Previously, the highest judicial body of the R.S.F.S.R. was the Supreme Court, chosen by the Supreme Soviet for a five-year period. It supervised the activities of all other judicial bodies—the Supreme Courts of the A.S.S.R.'s and the courts of other territories.

The Supreme Court is the highest judicial body of the Russian Federation as well. In 1991 a Constitutional Court was established to review Russian laws and treaties. Constitutional justice in the court is based on the equality of all citizens before the law and in exact accordance with the law; judges are independent and subordinate only to the

law; trials are to be open, and the accused are guaranteed a defense. In the past these ideals were violated on many occasions. Millions suffered from illegal actions during the Stalin period, and officially sanctioned illegality continued until the early 1980s.

Armed forces and security. The Russian armed forces consist of an army, navy, air force, air defense force, and strategic rocket force, all under the command of the president. About half the troops are conscripts: military service lasting 18 months in the army or 24 months in the navy is compulsory for men over 18 years of age, although draft evasion is widespread. There is an extensive reserve force. Russia maintains defense facilities in several former Soviet republics and contributes a small proportion of its troops to the joint forces of the CIS. Russia's military capacity has declined since the breakup of the Soviet Union. Nonetheless, it still has one of the world's largest armed-forces establishments that includes a vast nuclear arsenal.

Intelligence operations

Foreign and domestic intelligence operations are managed, respectively, by the Foreign Intelligence Service and the Federal Counterintelligence Service, agencies that emerged after the reorganization of the Soviet KGB (Committee for State Security) in 1991. High officials are protected by the Presidential Guard. Local police forces have become progressively more ineffective, especially against the organized crime that has burgeoned in Russia's emerging market economy. Well-trained private security forces have become increasingly common.

EDUCATION

Prior to the early 1990s, education in the U.S.S.R. was highly centralized. The system in the Russian Federation still resembles the old Soviet one in this respect. *Perestroika* and particularly the demise of the Communist Party have led to major revisions of the syllabi in many subjects and the removal of compulsory indoctrination in Marxist-Leninist theory. Preschool provision is well developed, and a high proportion of the preschool-age children attend crèches or kindergartens. Free compulsory schooling begins at age seven and lasts for at least eight years, with more than three-fifths now attending for up to 10 years. Admission to higher education is selective and highly competitive: first-degree courses usually take five years. Non-Russian schoolchildren are taught in their own language, but Russian is a compulsory subject at the secondary level. Higher education in Russia is almost entirely in Russian, although there are a few institutions, mainly in the minority republics, where the local language is also used. The most prestigious universities are those of Moscow and St. Petersburg.

SOCIAL CONDITIONS

The average monthly wages of workers and service personnel are considerably enhanced by benefits from public welfare funds. Wages depend on the type, chiefly the heaviness, of the work, as well as on working conditions and the importance of the branch of industry. Higher-than-average wages are paid to labourers engaged in construction work, transportation, and mining. Collective farm wages are paid only after all other commitments (taxes, compulsory investments, seed reserves) have been met.

The issue of land ownership and the sale and transfer of property is still being resolved in post-Soviet Russia. A decree issued in 1993 allowed for the private ownership of land but provided no mechanisms for sales and transfers. In 1996 a second decree was designed to provide these mechanisms.

Public welfare funds from the state budget, enterprises, and trade unions are used substantially to improve the condition of workers in Russia, both materially and socially. A major portion of public welfare funds goes for free medical service, training, pensions, and scholarships. All workers and professionals in Russia receive paid vacations of up to one month. During much of the Soviet period, advances in health care and material well-being led to a decline in mortality, the control or eradication of the more dangerous infectious diseases, and an increase in the average life span. Since the late 1980s, however, life expectancy has fallen sharply, particularly for men. In 1991 Russia began experiencing negative population growth, caused by a plummeting birth rate and a rising mortality rate.

The social, ecological, and economic factors underlying these developments include stress resulting from economic dislocation, contamination of food and drinking water by agricultural and industrial pollutants, poor nutrition, contact with toxic substances in the workplace, and high rates of alcohol and tobacco consumption. Air pollution in heavily industrialized areas has led to relatively high rates of lung cancer in these regions, and high incidences of stomach cancer occur in regions where consumption of carbohydrates is high and intake of fruits, vegetables, milk, and animal proteins is low. Rates of alcoholism—especially among young people—and drug addiction have increased.

Although Russia has a large supply of doctors, a number of factors have contributed to a decline in the quality of health care, such as poor intensive and emergency care, insufficiently trained medical personnel, shortages of medicine, and limited development of specialized services such as maternity and hospice care.

(S.A.V./R.He./J.C.D.)

Cultural life

THE DEVELOPMENT OF RUSSIAN CULTURE

Slavic and borrowed elements

Russia's unique and vibrant culture developed, as did the country itself, from a complicated interplay of native Slavic cultural material and borrowings from a wide variety of foreign cultures. In the Kievan period (*c.* 10th–13th centuries) the borrowings were primarily from Eastern Orthodox Byzantine culture. During the Muscovite period (*c.* 14th–17th centuries) the Slavic and Byzantine cultural substrates were enriched and modified by Asiatic influences carried by the Mongol hordes. Finally, in the modern period (since the 18th century) the cultural heritage of western Europe was added to the Russian melting pot.

The Kievan period. Although many traces of the Slavic culture that existed in the territories of Kievan Rus survived beyond its Christianization (which occurred, according to the Russian Primary Chronicle, in AD 988), the cultural system that organized the lives of the early Slavs is far from being understood. From the 10th century on, however, enough material has survived to give a reasonable portrait of Old Russian cultural life. High culture in Kievan Rus was primarily ecclesiastical. The level of literacy was low, and artistic composition was undertaken almost exclusively by monks. The earliest literary works to have circulated were translations from Greek into Old Church Slavonic (a South Slavic dialect that was, in this period, close enough to Old Russian to be understandable). By the 11th century, however, monks were producing original works (on Byzantine models), primarily in the genres of saints' lives, historical chronicles, and homilies. At least one great secular work was produced as well: the epic *The Song of Igor's Campaign,* which dates from the late 12th century and describes a failed military expedition against the neighbouring Polovtsy. Evidence also exists (primarily in the form of church records of suppression) of a thriving popular culture based on pre-Christian traditions centring around harvest, marriage, birth, and death rituals. The most important aspects of Kievan culture for the development of modern Russian culture, however, were not literary or folkloric but rather artistic and architectural. The early Slavic rulers expressed their religious piety and displayed their wealth through the construction of stone churches, at first in Byzantine style (like the 11th-century Cathedral of St. Sophia that still stands in Kiev, Ukraine) and later in a distinctive Russian style (best preserved today in churches in and around the city of Vladimir, to the east of Moscow). The interiors of many of these churches were ornately decorated with frescoes and icons.

The Muscovite period. The Mongol (Tatar) invasions of the early 13th century decimated Kievan Rus. By the time Russian political and cultural life began to recover in the 14th century, a new centre had arisen: Muscovy (Moscow). Continuity with Kiev was provided by the Orthodox church, which had acted as a beacon of national

life during the period of Tatar domination and which continued to play the central role in Russian cultural life into the 17th century. As a result, Russian cultural development in the Muscovite period was quite different from that of western Europe, which at this time was experiencing the secularization of society and the rediscovery of the classical cultural heritage that characterized the Renaissance. At first the literary genres employed by Muscovite writers were the same as those that had dominated in Kiev. The most remarkable literary monuments of the Muscovite period, however, are unlike anything that came before. Most noteworthy is the correspondence between Tsar Ivan IV the Terrible and Prince Andrey Kurbsky during the 1560s and '70s. Kurbsky, a former general in Ivan's army, defected to Poland, from where he sent a letter critical of the tsar's regime. Ivan's diatribes in response are both wonderful expressions of outraged pride and literary tours de force, combining the highest style of Muscovite hagiographic writing with pithy and vulgar attacks on his enemy. Similarly vigorous in style is the first full-scale autobiography in Russian literature, *The Life of the Archpriest Avvakum, by Himself* (*c.* 1672–75).

As in the Kievan period, however, the most significant cultural achievements of Muscovy were not in literature but rather in the visual arts and architecture. The Moscow school of icon painting produced great masters, among them Andrey Rublyov (whose "Old Testament Trinity," now in Moscow's State Tretyakov Gallery, is among the most beautiful icons ever painted) and Dionysius. Russian architects continued to design and build impressive churches, including the celebrated Cathedral of St. Basil the Blessed on Moscow's Red Square. Built to commemorate the Russian capture of the Tatar capital Kazan, St. Basil's is a perfect example of the confluence of Byzantine and Asiatic cultural streams that characterizes Muscovite culture.

Icon painting

The emergence of modern Russian culture. The gradual turn of Russia toward western Europe that began in the 17th century led to an almost total reorientation of Russian interests during the reign of Peter I the Great (1694–1725). Although Peter was not particularly interested in cultural questions, the influx of Western ideas (which accompanied the technology Peter found so attractive) and the weakening of the Orthodox church led to a cultural renaissance in the reigns of his successors. In the late 1730s the poets Mikhail Lomonosov and Vasily Trediakovsky carried out reforms as far-reaching as those of Peter. They adapted German syllabo-tonic versification to Russian, developing the system of "classical" metres that prevails in Russian poetry to this day. In the 1740s, in imitation of French Neoclassicism, Aleksandr Sumarokov wrote the first Russian stage tragedies. In the course of the century, Russian writers assimilated all the European genres; much of their work was derivative, but the comedies of Denis Fonvizin and the powerful solemn odes of Gavrila Derzhavin were original and have remained part of the active Russian cultural heritage. Prose fiction made its appearance at the end of the century in the works of the sentimentalist Nikolay Karamzin. By the beginning of the 19th century, after a 75-year European cultural apprenticeship, Russia had developed a flexible secular literary language, had a command of modern literary forms, and was ready to produce fully original cultural work.

THE ARTS

Literature. *The 19th century.* The first quarter of the 19th century was dominated by romantic poetry. Vasily Zhukovsky's 1802 translation of Thomas Gray's *An Elegy Written in a Country Church Yard* ushered in a vogue for the personal, elegiac mode that was soon amplified in the work of Konstantin Batyushkov, Prince Pyotr Vyazemsky, and the young Aleksandr Pushkin. Although there was a call for civic-oriented poetry in the late 1810s and early '20s, most of the strongest poets followed Zhukovsky's lyrical path. The mature Pushkin of the 1820s, however, went his own way, producing a series of masterpieces that laid the foundation for his eventual recognition as Russia's national poet (Pushkin, for Russians, is the equivalent of Shakespeare for English readers or Dante for Italians); these works include the "Byronic" long poems *The Prisoner of the Caucasus* (1820–21) and *The Gypsies* (1824), the "novel in verse" *Yevgeny Onegin* (published 1833), and the "Shakespearean" tragedy *Boris Godunov* (1825), as well as exquisite lyrical verse. Pushkin's poetry is remarkable for its classical balance, brilliant and frequently witty use of the Russian literary language, and philosophical content.

During the 1830s there was a gradual decline in poetry and a rise of prose. This shift coincided with a change in literary institutions: the aristocratic salon, which had been the seedbed for Russian literature, was gradually supplanted by the monthly "thick journals," the editors and critics of which became Russia's tastemakers. The turn to prose was signaled in the work of Pushkin, whose *Tales of the Late Ivan Petrovich Belkin* (1831), *The Queen of Spades* (1834), and *The Captain's Daughter* (1836) all appeared before his death in 1837. Also in the 1830s the first publications appeared by Nikolay Gogol, a comic writer of Ukrainian origin, whose grotesquely hilarious oeuvre includes the story "The Nose," the play *The Government Inspector* (both 1836), and the epic novel *Dead Souls* (1842). Although Gogol was known in his own day primarily as a satirist, he now is appreciated as a verbal magician, whose works seem akin to the absurdists of the 20th century. One final burst of poetic energy appeared in the late 1830s in the verse of Mikhail Lermontov, known also as the author of the first Russian psychological novel, *A Hero of Our Time* (1840).

The rise of prose

In the 1840s the axis of Russian literature shifted decisively from the personal and Romantic to the civic and realistic, a shift presided over by the great Russian literary critic Vissarion Belinsky. Belinsky called for a literature concerned with current social problems, although he never expected it to give up the aesthetic function entirely. By the end of the 1840s, Belinsky's ideas had triumphed. Early works of Russian realism include Ivan Goncharov's antiromantic novel *A Common Story* (1847) and Fyodor Dostoyevsky's *Poor Folk* (1846).

From the 1840s until the turn of the 20th century, the realist novel was the dominant genre in Russian literature. Realism was not, however, a monolithic movement. In the early period, the favoured method was the "physiological sketch," often depicting a typical member of the downtrodden classes: quintessential examples are found in Ivan Turgenev's 1852 collection *A Sportsman's Sketches.* In these beautifully crafted stories, Turgenev describes the life of Russian serfs as seen through the eyes of a Turgenev-like narrator. The power of Turgenev's artistic depiction was credited with convincing Tsar Alexander II of the need to emancipate the serfs. Turgenev followed *Sketches* with a series of novels, each of which was felt by contemporaries to have captured the essence of Russian society at the time it appeared. The most celebrated is *Fathers and Sons* (1862), in which generational and class conflict in the period of Alexander II's reforms is described through the interactions of the Kirsanov family (father, son, and uncle) with the young "nihilist" Bazarov.

The two other great realists of the 19th century were Fyodor Dostoyevsky and Leo Tolstoy. Dostoyevsky, who was arrested in 1849 for his involvement in a socialist reading group, reentered the literary scene in the late 1850s. While in prison, he experienced a religious conversion, and his novels of the 1860s and '70s are suffused with messianic Orthodox ideas. Dostoyevsky's major novels—*Crime and Punishment* (1866), *The Idiot* (1868–69), *The Possessed* (1872), and *The Brothers Karamazov* (1879–80)—are filled with riveting, often unstable characters and dramatic scenes. While Dostoyevsky delves into the psychology of men and women at the edge, Tolstoy's novels treat the everyday existence of normal people. In both *War and Peace* (1865–69) and *Anna Karenina* (1875–77) Tolstoy draws beautifully nuanced portraits filled with deep psychological and sociological insight.

Dostoyevsky and Tolstoy

By the early 1880s, the hegemony of the realist novel was waning. What was to replace the novel, however, was unclear. Russian poetry, notwithstanding the civic verse of Nikolay Nekrasov and the subtle lyrics of Afanasy Fet, had not played a central role in the literary process since

the 1830s, and drama, despite the able work of Aleksandr Ostrovsky, was a marginal literary activity for most writers. The only major prose writer to appear in the 1880s and '90s was Anton Chekhov, whose specialty was the short story. In his greatest stories—including "The Man in a Case" (1898), "The Lady with a Lapdog" (1899), "The Darling" (1899), and "In the Ravine" (1900)—Chekhov manages to attain all the power of his great predecessors in a remarkably compact form. Toward the end of his career, Chekhov also became known for his dramatic work, including such pillars of the world theatrical repertoire as *Uncle Vanya* (1897) and *The Cherry Orchard* (1903). Chekhov's heirs in the area of short fiction were the lower-class writer Maksim Gorky (who later would become the dean of Soviet letters) and the aristocrat Ivan Bunin (who emigrated after the Russian Revolution of 1917 and received the Nobel Prize for Literature in 1933).

The 20th century. The turn of the century ushered in a new renaissance in Russian poetry and drama, a "Silver Age" that rivaled, and in some respects surpassed, the Pushkinian "Golden Age." The civic orientation that had dominated Russian literature since the 1840s was, for the moment, abandoned. The avant-garde's new cry was "art for art's sake," and the new idols were the French Symbolists. The first, "decadent" generation of Russian Symbolists included the poets Valery Bryusov, Konstantin Balmont, and Zinaida Gippius. The second, more mystically and apocalyptically oriented generation included Aleksandr Blok (perhaps the most talented lyric poet Russia ever produced), the poet and theoretician Vyacheslav Ivanov, and the poet and prose writer Andrey Bely. The Symbolists remained ascendant until 1910, when internal dissension led to the collapse of the movement.

Poetry of the 1917 Revolution period

The period just before and immediately following the 1917 Revolution was marked by the work of six spectacularly talented, difficult poets. Anna Akhmatova's brief, finely chiseled lyrics brought her fame at the outset of her career, but later in life she produced such longer works as *Requiem,* her memorial to the victims of Joseph Stalin's purges. The Futurists Velimir Khlebnikov and Vladimir Mayakovsky engaged in innovative experiments to free poetic discourse from the fetters of tradition. Marina Tsvetayeva was a great poetic experimenter as well. She produced much of her major work as an émigré but returned to the Soviet Union in 1939, only to commit suicide there in 1941. Boris Pasternak, who won the Nobel Prize for Literature in 1958, produced lyrics of great depth and power in this period, and Osip Mandelshtam brought his great erudition to the creation of some of the most beautiful and haunting lyric poems in the Russian language.

Many of the writers who began to publish immediately after the Revolution turned to prose, particularly the short story and the novella. Some were inspired by the recent Revolution and Russian Civil War, including Boris Pilnyak (*The Naked Year* [1922]), Isaak Babel (*Red Cavalry* [1926]), and Mikhail Sholokhov, who won the Nobel Prize in 1965 for his writings. Others described life in the new Soviet Union with varying degrees of mordant sarcasm; the short stories of Mikhail Zoshchenko, the comic novels of Ilya Ilf and Yevgeny Petrov, and the short novel *Envy* (1927) by Yury Olesha fall into this category.

The first decade after the Revolution also was a time of significant advances in literary theory and criticism, which changed methods of literary study throughout the world. Members of the Moscow Linguistic Circle and of the Society for the Study of Poetic Language (OPOYAZ) in Petrograd (now St. Petersburg) combined to create Formalist literary criticism, a movement that concentrated on analyzing the internal structure of literary texts. At the same time, the theorist Mikhail Bakhtin began to develop a sophisticated criticism concerned with ethical problems and ways of representing them, especially in his favourite genre, the novel.

By the late 1920s the period of Soviet experimentation had ended. Censorship became much stricter, and many of the best writers were silenced. During the late 1920s and the '30s there appeared what became known as the classics of Socialist Realism, a literary method that in 1934 was declared to be the only acceptable one for Soviet writers. A few among these classics—Fyodor Gladkov's *Cement* (1925), Nikolay Ostrovsky's *How the Steel Was Tempered* (1932–34), and Valentin Katayev's *Time, Forward!* (1932)—have retained some literary interest. The masterpieces of this period, however, did not fit the canons of Socialist Realism and were not published until many years later. They include Mikhail Bulgakov's grotesquely funny *The Master and Margarita* (1966–67) and Andrey Platonov's dark pictures of rural and semi-urban Russia, *The Foundation Pit* (1973) and *Chevengur* (1972).

New writers and trends appeared during the "thaw" period of the 1950s and early '60s. The vibrant young poetic voices of Joseph Brodsky, Yevgeny Yevtushenko, and Andrey Voznesensky were heard. Aleksandr Solzhenitsyn emerged from a Soviet prison camp to shock the U.S.S.R. and the world with his story *One Day in the Life of Ivan Denisovich* (1962). "Youth" prose on the model of J.D. Salinger appeared as well, in the work of Vasily Aksyonov and Vladimir Voynovich. By the late 1960s, however, most of these writers had again been silenced. Brodsky, Aksyonov, Voynovich, and Solzhenitsyn were forced to leave the country by 1980, and the best writing was again unpublishable. Practically the only decent writing acceptable for publication from the late 1960s through the mid-1980s came from the "village prose" writers (the best include the novelist Valentin Rasputin and the short-story writer Vasily Shukshin), who treated the clash of rural traditions with modern life in a realistic idiom. Somewhat apart stands the morally complex fiction of Yury Trifonov (*The House on the Embankment* [1976]). Nevertheless, as with the 1930s and '40s, the most important literature of this period first appeared outside the Soviet Union. Notable writers include Varlam Shalamov, author of exquisitely artistic stories that chronicled the horrors of the Gulag; Andrey Sinyavsky, whose complex novel *Goodnight!* appeared in Europe in 1984, long after he had been forced to emigrate from the U.S.S.R.; and Venedikt Yerofeyev, whose grotesque latter-day picaresque *Moscow-Petushki*—published in a clandestine (samizdat) edition in 1968—is a minor classic.

"Village prose"

The literature that first appeared in the 1980s has yet to stand the test of time, but it appears that most of the best work published was in poetry—as in the work of conceptualists like Dmitry Prigov and in the meta-metaphoric poetry of Aleksey Parshchikov, Olga Sedakova, Ilya Kutik, and others.

Music. *The 19th century.* Secular music on a Western model appeared later than did Westernized literature in Russia. Although a few works of interest have survived from the 18th century, the "father" of modern Russian classical music, Mikhail Glinka, worked in the second quarter of the 19th century. Glinka created a Russian national music by grafting Russian melodies onto European harmonies. His patriotic *A Life for the Tsar* (1836) and his Pushkin-inspired *Ruslan and Lyudmila* (1842) are the oldest Russian operas that remain in the standard repertoire.

By the second half of the 19th century, an active Russian musical life was in place. Like so many other areas of Russian culture, music was split into Westernizer and Slavophile (nationalist) camps. The principal composer of the former was Peter Ilich Tchaikovsky, whose symphonies, overtures, ballets, and operas combined a careful European craftsmanship with a judicious use of native melodies. The Slavophile camp called for a more national music that would be based not merely on native melodies but also on the harmonic system of the Russian folk song. Although to the musical Westernizers' ears the work of the nationalists sounded barbaric, the major compositions of the Slavophile composers Aleksandr Borodin, Modest Mussorgsky, and Nikolay Rimsky-Korsakov have become staples of the international repertoire, just as have those of Tchaikovsky.

The 20th century. In the immediate pre-Revolutionary period, three major Russian composers emerged: Aleksandr Scriabin, Sergey Rachmaninoff, and Igor Stravinsky. Scriabin was a piano virtuoso and mystic whose compositions were close in spirit to Symbolist literature. Rachmaninoff, also a major pianist, is known primarily for his lyrical piano works. Stravinsky, who began as a student

of Rimsky-Korsakov, quickly outgrew his teacher and, in the course of the century, produced a dazzling string of groundbreaking works in a wide variety of styles, including the ballets *Petrushka* (1911) and *The Rite of Spring* (1913). Stravinsky emigrated in 1914, Rachmaninoff following after the Revolution.

Shostakovich and Prokofiev

Soviet music was dominated by two major composers: Dmitry Shostakovich and Sergey Prokofiev. Both composed in a wide variety of styles and genres, and, even though Shostakovich in particular had serious problems with the artistic authorities, both were able to remain productive even during the worst years of Stalinism. Shostakovich is known primarily for his 15 symphonies, although he also produced masterpieces of chamber music, opera, and ballet. Prokofiev's best-known works are his ballet music for *Romeo and Juliet* (1935–36) and his score for Sergey Eisenstein's film *Alexander Nevsky* (1938), although he, too, wrote a wide variety of chamber music, orchestral music, and opera. Among contemporary composers, the complex work of Alfred Schnittke is highly valued and frequently perfomed.

Popular culture also produced many renowned performers. Particularly notable is the legacy of two balladeers—composers who perform their own songs to guitar accompaniment. The raspy-voiced actor and musician Vladimir Vysotsky, whose songs circulated on thousands of bootleg cassettes throughout the 1960s and '70s, was perhaps the best-known performer in the U.S.S.R. until his death in 1980. The Georgian Bulat Okudzhava has an almost equally loyal following. The pop singer Alla Pugacheva drew large audiences in the 1970s, and the rock bands Aquarium (Akvarium) and Kino became quite popular in the 1980s.

In the area of classical musical performance, Soviet conservatories turned out generations of world-renowned soloists. Among the best-known were the violinists David Oistrakh and Gidon Kremer, the cellist Mstislav Rostropovich, the pianists Svyatoslav Richter and Emil Gilels, and the vocalist Galina Vishnevskaya.

The visual arts. *The 19th century.* Like music, the visual arts were slower to develop along European lines than was literature in Russia. The 18th and early 19th centuries did not produce any great Russian painters, with the exception of the portraitist Dmitry Levitsky. In the 1830s the Russian Academy of Arts (which had been founded in 1757) began sending Russian painters abroad for training. Among the most gifted of these were Aleksandr Ivanov and Karl Bryullov: both are known for Romantic historical canvases. A truly national tradition did not begin, however, until the 1870s with the appearance of the "Itinerants." Although their work is not well known outside Russia, the serene landscapes of Isaac Levitan, the expressive portraits of Ivan Kramskoy and Ilya Repin, and the socially oriented genre paintings of Vladimir Makovsky, Vasily Perov, and Repin deserve an international reputation.

The 20th century. As was the case with literature, there was a burst of creativity in the visual arts in the years just before the 1917 Revolution. Russian painters interacted frequently with their European counterparts and played a major role in the European art scene. This period was marked by a turning away from realism to primitivism, symbolism, and abstract painting. The careers of such major artists as Wassily Kandinsky, Marc Chagall, Natalya Goncharova, Mikhail Larionov, Kazimir Malevich, and Vladimir Tatlin all began in the immediate pre-Revolutionary years.

Constructivism

As in literature, the 1920s were a period of continued experimentation. Perhaps the most noteworthy movement was Constructivism. Led by El Lissitzky and Aleksandr Rodchenko, the Constructivists favoured strict geometric forms and crisp graphic design. Many also became actively involved in the task of creating living spaces and forms of daily life, working in such fields as furniture, ceramic, and clothing design and architecture. Non-Constructivist artists, including Pavel Filonov and Mariya Ender, also produced major works in this period.

By the end of the 1920s, however, the same pressures that confronted experimental writing were brought to bear on the visual arts. A return to the classics of realism was decreed, and the great painters of the early 1920s found themselves increasingly isolated. Eventually their works were removed from museums, and in many cases the artists themselves were almost completely forgotten. It was not until the late 1980s that the greatest works of Russian art of the 20th century were again made available to the public. Experimental art was replaced by countless pictures of Lenin, as, for example, Isaak Brodsky's "Lenin at the Smolny" (1930), and by a seemingly unending string of rose-tinted Socialist Realist depictions of everyday life bearing titles like "The Tractor Drivers' Supper" (1951).

The visual arts took longer to recover from the Stalinist years than did literature. It was not until the 1960s and '70s that a new group of artists, all of whom worked "underground," appeared. Major artists included Ernst Neizvestny, Ilya Kabakov, Mikhail Shemyakin, and Erik Bulatov. They employed techniques as varied as primitivism, hyperrealism, grotesque, and abstraction, but they shared a common distaste for the canons of Socialist Realism. By the late 1980s a large number of them had emigrated. Many became well known on the international art scene.

The performing arts. *The 19th century.* The dramatic and ballet theatres were entirely under government control until the end of the 19th century. Actors and dancers were government employees and often were treated badly. Nevertheless, theatrical life was quite active throughout the century. Famous Russian actors and dancers of the early part of the century included the ballerina Istomina and the actor Mikhail Shchepkin. From an international perspective, however, the greatest success of the Russian theatre was in the area of classical ballet. Since the 1820s, Russian dancers have reigned supreme on the ballet stage. Many great choreographers, even those of non-Russian origin, worked for the Russian Imperial Theatres, including Marius Petipa, who choreographed Tchaikovsky's ballets *Swan Lake* and *The Sleeping Beauty.*

The 20th century. Two directors and one producer dominated Russian theatrical life in the first decades of the 20th century: the producer was Sergey Diaghilev, and the directors were Konstantin Stanislavsky and Vsevolod Meyerhold. Together with Vladimir Nemirovich-Danchenko, Stanislavsky founded the Moscow Art Theatre (later called the Moscow Academic Art Theatre) in 1898. Stanislavsky's insistence on historical accuracy, exact realism, and intense psychological preparation by his actors led to a string of successful productions from the beginning of the century into the 1930s. The theatre was known particularly for its productions of the plays of Anton Chekhov, whose *The Seagull* (1896) was the hit of the theatre's inaugural season.

Moscow Art Theatre

Meyerhold was one of Stanislavsky's students, but he broke with his master's insistence on realism. He welcomed the Revolution and put his considerable talent and energy into creating a new theatre for the new state. Throughout the 1920s and into the '30s he staged brilliant, inventive productions, both of contemporary drama and of the classics. However, his iconoclastic style fell out of favour in the 1930s, and he was arrested and executed in 1940.

Diaghilev was a brilliant organizer and impresario whose innovative Ballets Russes premiered many of the most significant ballets of the first quarter of the century. Although the company was based primarily in Paris, Diaghilev employed major Russian composers (particularly Stravinsky), artists (*e.g.,* Alexandre Benois, Goncharova, and Larionov), and dancers (including Vaslav Nijinsky and Tamara Karsavina) in his legendary company.

Ballet was one of the great successes of the Soviet period as well, not because of any innovations but because the great troupes of the Bolshoi Theatre in Moscow and the Kirov Theatre in Leningrad were able to preserve the traditions of classical dance that had been perfected in pre-Revolutionary Russia. The Soviet Union's choreography schools produced one internationally famous star after another, including the incomparable Maya Plisetskaya, Rudolf Nureyev, and Mikhail Baryshnikov.

Another extremely successful area of theatrical performance was puppet theatre. The State Central Puppet The-

atre founded by Sergey Obraztsov in Moscow continues to give delightful performances for patrons of all ages. The same can be said for the spectacular presentations of the Moscow State Circus, which has performed throughout the world to great acclaim.

Motion pictures. The Soviet cinema, too, was a hotbed of invention in the immediate post-Revolutionary period. Its most celebrated director was Sergey Eisenstein (a student of Meyerhold), whose great films include *Battleship Potemkin* (1925) and *Alexander Nevsky* (1938). Film did not escape the strictures of Socialist Realism, but a few post-World War II films in this style were artistically successful, including *The Cranes Are Flying* (1957; directed by Mikhail Kalatozov) and *Ballad of a Soldier* (1959; directed by Grigory Chukhrai). A number of successful film versions of classic texts were made in the 1950s and '60s, particularly Grigory Kozintsev's spectacular versions of *Hamlet* (1964) and *King Lear* (1971). The 1960s and '70s produced a few great directors and artistically successful films. Two standouts were Andrey Tarkovsky (*Andrey Rublev* [1966] and *Mirror* [1974]) and the Georgian-born Armenian Sergey Paradzhanov (*Shadows of Forgotten Ancestors* [1964] and *The Colour of Pomegranates* [1969]).

CULTURAL AND EDUCATIONAL INSTITUTIONS

The press and other media. Russian 19th-century journalism was extremely vigorous, with newspapers and monthly "fat" journals being the most important forums. Daily newspapers and monthly journals of all political and artistic stripes continued to appear in the immediate aftermath of the Revolution. However, the state's desire to control sources of information and propaganda manifested itself quickly, and most independent newspapers were eliminated by the early 1920s. What remained were the ubiquitous daily duo of *Pravda* ("Truth") and *Izvestiya* ("News"). Journals were in a somewhat better position, especially those that published mostly works of literature. Periodicals like *Krasnaya nov* ("Red Virgin Soil") and *LEF* ("The Left Front of Art") published much significant literature in the 1920s. In the 1960s the journal *Novy mir* revived this tradition. In the 1980s it was joined by a revitalized *Ogonyok* ("Spark").

State control of news media

Radio and television from the time of their appearance in the Soviet Union were heavily dominated by the party apparatus and were seen as primary tools for propaganda. Until the mid-1980s most television programming consisted of either direct or indirect propaganda spiced with high art (*e.g.,* filmed concerts and plays) and occasional grade-B thriller motion pictures. In the period of *glasnost,* television was a leader in innovative programming, helping to create the situation in which the Soviet state was destroyed.

Museums. Some of the greatest museums in the world can be found in the cities of Moscow and St. Petersburg. In Moscow the Pushkin Fine Arts Museum houses treasures of western European art, while the Tretyakov Gallery has a strong collection of Russian art. In St. Petersburg the Hermitage is one of the great art museums of the world, and the Russian Museum has wonderful examples of Russian art. In addition, in the suburbs outside St. Petersburg, the former tsarist palaces at Pavlovsk, Pushkin, and Petrodvorets have been restored as museums. They are popular destinations for both Russian citizens and foreign tourists.

Universities. Russia's oldest university, in Moscow, was founded in 1755. Throughout the 19th century and into the 20th, Russian universities in Moscow, St. Petersburg, and Kazan produced world-class scholars, notably the mathematician Nikolay Lobachevsky and the chemist Dmitry Mendeleyev. Although universities suffered severely during the Stalinist purges, a number of universities have continued to provide high-quality education, particularly in the sciences. The most important include Moscow M.V. Lomonosov State University, St. Petersburg State University, and Novosibirsk State University. In the 1960s the school of semiotics connected with Tartu University (Estonia) under the direction of Yury Lotman produced world-renowned studies of Russian and Soviet culture and literature.

SPORTS AND RECREATION

Sports played a major role in the Soviet state in the post-World War II period. The achievements of Soviet athletes in the international arena, particularly in the Olympic Games (the Soviets first participated in the 1952 Summer and the 1956 Winter Olympics), were a source of great national pride. Although athletes were technically amateurs, they were well supported by the Sports State Committee. In team sports the U.S.S.R. was especially successful in ice hockey—winning numerous world championships and Olympic gold medals—volleyball, and later basketball. Soviet gymnasts and track-and-field athletes (male and female), weight lifters, wrestlers, and boxers were consistently among the best in the world. Even after the collapse of the Soviet empire, Russian athletes have continued to dominate international competition in these areas.

Olympic participation

On the amateur level, the lack of facilities and equipment have prevented many average Russian citizens from participating in sporting activities, but jogging, soccer, and fishing have been popular. Finally, many Russians are avid chess players, and the country has produced most of the greatest players of the 20th century. (A.B.Wa.)

For statistical data on the land and people of Russia, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

HISTORY

From the beginnings to *c.* 1700

PREHISTORY AND THE RISE OF THE RUS

Ancient trading empires

Indo-European, Ural-Altaic, and diverse other peoples have occupied what is now the territory of Russia since the 2nd millennium BC, but little is known about their ethnic identity, institutions, and activities. In ancient times, Greek and Iranian settlements appeared in the southernmost portions of what is now Ukraine. Trading empires of that era seem to have known and exploited the northern forests—particularly the vast, triangular-shaped region west of the Urals between the Kama and Volga rivers—but these contacts seem to have had little lasting impact. Between the 4th and 9th centuries AD, the Huns, Avars, Goths, and Magyars passed briefly over the same terrain, but these transitory occupations also had little influence upon the East Slavs, who during this time were spreading south and east from an area between the Elbe River and the Pripet Marshes. In the 9th century, as a result of penetration into the area from the north and south by northern European and Middle Eastern merchant adventurers, their society was exposed to new economic, cultural, and political forces.

The scanty written records tell little of the processes that ensued, but archaeological evidence—notably, the Middle Eastern coins found in eastern Europe—indicates that the development of the East Slavs passed through several stages.

From about 770 to about 830, commercial explorers began an intensive penetration of the Volga region. From early bases in the estuaries of the rivers of the eastern Baltic region, Germanic commercial-military bands, probably in search of new routes to the east, began to penetrate territory populated by Finnic and Slavic tribes, where they found amber, furs, honey, wax, and timber products. The indigenous population offered little resistance to their incursions, and there was no significant local authority to negotiate the balance among trade, tribute, and plunder. From the south, trading organizations based in northern Iran and North Africa, seeking the same products, and particularly slaves, became active in the lower Volga, the Don, and, to a lesser extent, the Dnieper region. The history of the Khazar state is intimately connected with these activities.

About 830 commerce appears to have declined in the Don and Dnieper regions. There was increased activity in the north Volga, where Scandinavian traders who had previously operated from bases on Lakes Ladoga and Onega established a new centre, near present-day Ryazan. Here, in this period, the first nominal ruler of Rus (called, like the Khazar emperor, khagan) is mentioned by Islāmic and Western sources. This Volga Rus khagan state may be considered the first direct political antecedent of the Kievan state.

Within a few decades these Rus, together with other Scandinavian groups operating farther west, extended their raiding activities down the main river routes toward Baghdad and Constantinople, reaching the latter in 860. The Scandinavians involved in these exploits are known as Varangians; they were adventurers of diverse origins, often led by princes of warring dynastic clans. One of these princes, Rurik of Jutland, is considered the progenitor of the dynasty that ruled in various portions of East Slavic territory until 1598. Evidences of the Varangian expansion are particularly clear in the coin hoards of 900–930. The number of Middle Eastern coins reaching northern regions, especially Scandinavia, indicates a flourishing trade. Written records tell of Rus raids upon Constantinople and the northern Caucasus in the early 10th century.

The Varangian expansion

In the period from about 930 to 1000, the region came under complete control by Varangians from Novgorod. This period saw the development of the trade route from the Baltic to the Black Sea, which established the basis of the economic life of the Kievan principality and determined its political and cultural development.

The degree to which the Varangians may be considered the founders of the Kievan state has been hotly debated since the 18th century. The debate has from the beginning borne nationalistic overtones. Recent works by Russians have generally minimized or ignored the role of the Varangians, while non-Russians have occasionally exaggerated it. Whatever the case, the lifeblood of the sprawling Kievan organism was the commerce organized by the princes. To be sure, these early princes were not "Swedes" or "Norwegians" or "Danes"; they thought in categories not of nation but of clan. But they certainly were not East Slavs. There is little reason to doubt the predominant role of the Varangian Rus in the creation of the state to which they gave their name.

KIEV

The rise of Kiev. The consecutive history of the first East Slavic state begins with Prince Svyatoslav (d. 972). His victorious campaigns against other Varangian centres, the Khazars, and the Volga Bulgars and his intervention in the Byzantine–Danube Bulgar conflicts of 968–971 mark the full hegemony of his clan in Rus and the emergence of a new political force in eastern Europe. But Svyatoslav was neither a lawgiver nor an organizer; the role of architect of the Kievan state fell to his son Vladimir (*c.* 980–1015), who established the dynastic seniority system of his clan as the political structure by which the scattered territories of Rus were to be ruled. He invited or permitted the patriarch of Constantinople to establish an episcopal see in Rus.

Vladimir extended the realm (to include the watersheds of the Don, Dnieper, Dniester, Neman, Western Dvina, and upper Volga), destroyed or incorporated the remnants of competing Varangian organizations, and established relations with neighbouring dynasties. The successes of



The Kievan Rus in the 11th century.

his long reign made it possible for the reign of his son Yaroslav (ruled 1019–54) to produce a flowering of cultural life. But neither Yaroslav, who gained control of Kiev only after a bitter struggle against his brother Svyatopolk (1015–19), nor his successors in Kiev were able to provide lasting political stability within the enormous realm. The political history of Rus is one of clashing separatist and centralizing trends inherent in the contradiction between local settlement and colonization, on the one hand, and the hegemony of the clan elder, ruling from Kiev, on the other. As Vladimir's 12 sons and innumerable grandsons prospered in the rapidly developing territories they inherited, they and their retainers acquired settled interests that conflicted both with one another and with the interests of unity.

Separatist and centralizing trends

The conflicts were not confined to Slavic lands: the Turkic nomads who moved into the southern steppe during the 11th century (first the Torks, later the Kipchaks—also known as the Polovtsy, or Cumans) became involved in the constant internecine rivalries, and Rurikid and Turkic princes often fought on both sides. In 1097 representatives of the leading branches of the dynasty, together with their Turkic allies, met at Liubech, north of Kiev, and agreed to divide the Kievan territory among themselves and their descendants, although, later, Vladimir Monomakh made a briefly successful attempt (1113–25) to reunite the land of Rus.

The decline of Kiev. The hegemony of the prince of Kiev depended on the cohesion of the clan of Rurik and the relative importance of the southern trade, both of which began to decline in the late 11th century. This decline seems to have been part of a general shift of trade routes that can for convenience be associated with the First Crusade (1096–99) and that made the route from the Black Sea to the Baltic less attractive to commerce. At the same time, conflicts among the Rurikid princes acquired a more pronounced regional and separatist nature, reflecting new patterns in export trade along the northern and western periphery. Novgorod, in particular, began to gravitate toward closer relations with the cities of the Hanseatic League, which controlled the Baltic trade. Smolensk, Polotsk, and Pskov became increasingly involved in trade along western land routes, while Galicia and Volhynia established closer links with Poland and Hungary. The princes of these areas still contested the crown of the "grand prince of Kiev and all of Rus," but the title became an empty one; when Andrew Bogolyubsky of Suzdal won Kiev and the title in 1169, he sacked the city and returned to the upper Volga, apparently seeing no advantage in establishing himself in the erstwhile capital. (Roman Mstislavich of Galicia and Volhynia repeated these actions in 1203.) By the middle of the 12th century, the major principalities, owing to the prosperity and colonization of the Kievan period, had developed into independent political and economic units.

Social and political institutions. The paucity of evidence about social and political institutions in Kievan Rus suggests that they were rudimentary. The East Slavs had no significant tradition of supratribal political organization before the coming of the Varangians, who themselves, until well into the 10th century, had little interest in institutions more elaborate than those necessary for the exploitation of their rich, new territory. The territory of Rus, moreover, was immense and sparsely settled. The scattered towns, some probably little more than trading posts, were separated by large primeval forests and swamps.

Thus, although the campaigns of Svyatoslav indicate the extent of the political vacuum that his clan filled, he construed his domains as a clan possession rather than as a territorial or national state. His successor Vladimir, however, seems to have been conscious of one political element—organized religion—that distinguished both the contemporary empires and the newly established principalities in Poland and Hungary from his own. The church provided the concepts of territorial and hierarchical organization that helped to make states out of tribal territories; its teachings transformed a charismatic prince into a king possessing the attributes and responsibilities of a national leader, judge, and first Christian of the realm.

Once Vladimir had adopted Christianity in 988, his rule was supported by the propagation of Byzantine notions of imperial authority. The political traditions and conditions of Rus, however, required that the actual workings of the political system and some of its style be derived from other sources. The succession system, probably a vestige of the experience of the Rus khaganate in the upper Volga, was based upon two principles: the indivisibility of the basic territory of Rus (the principalities of Kiev, Chernigov, and Pereyaslavl) and the shared sovereignty of a whole generation. Seniority passed through an ascension by stages from elder brother to younger and from the youngest eligible uncle to the eldest eligible nephew. Such a system was admirably suited to the needs of the dynasty, because, by providing a rotating advancement of members of the clan through apprenticeships in the various territories of the realm, it assured control of the key points of the far-flung trading network by princes who were subject to traditional sanctions, and it gave them experience in lands over which they could someday expect to rule from Kiev. This system served well for a century after it was given final form by Vladimir and was revived by Monomakh (Vladimir II, ruled 1113–25), but it could not survive the decline of Kiev's importance.

The adoption of Christianity

Individual Rurikid princes maintained military retinues led by boyars. The princes and boyars drew their most significant revenues from the tribute or taxes collected annually in kind from territories under their control and disposed of in the export trade. The bulk of the population, apparently free peasants living in traditional agricultural communes, had little other connection with the dynasty and its trading cities.

Little is known of law in this period; it may be assumed that juridical institutions had not developed on a broad scale. The earliest law code (1016), called the "Russian Law," was one of the "Barbarian" law codes common throughout Germanic Europe. It dealt primarily with princely law—that is, with the fines to be imposed by the prince or his representative in the case of specified offenses.

Some scholars have held that, since land was in the hands of the boyar class, who exploited the labour of slaves and peasants, Kievan society should be termed feudal. The meagre sources indicate, however, that Kiev experienced nothing like the complex and highly regulated legal and economic relationships associated with feudalism in western Europe. Kiev's political system existed primarily for and by international trade in forest products and depended on a money economy in which the bulk of the population scarcely participated. The subsistence agriculture of the forest regions was not the source of Kiev's wealth, nor was it the matrix within which law and politics and history were made.

Formal culture came to Rus, along with Christianity, from the multinational Byzantine synthesis, primarily through South Slavic intermediaries. A native culture, expressed in a now lost pagan ritual folklore and traditions in the arts and crafts, existed before the Kievan period and then persisted alongside the formal culture, but its influence on the latter is conjectural.

Formal and native cultures

No single one of the regional (or, later, national) cultures, perhaps least of all that of Muscovy, can be called the heir of Kiev, although all shared the inheritance. The strands of continuity were everywhere strained, if not broken, in the period after Kiev's decline. But "Golden Kiev" was always present, in lore and bookish tradition, as a source of emulation and renascence.

THE LANDS OF RUS

The decline of Kiev led to regional developments so striking that the subsequent period has often been called the Period of Feudal Partition. This phrase is misleading: "feudal" is hardly more applicable to the widely varying institutions of this time than to those of the Kievan period. "Partition" implies a former unity of which there is insufficient evidence. The distinctiveness of the character and historical fortunes of each of the major East Slavic regions, discussed briefly below, is clear even in the Kievan period and has persisted into the 20th century.

Novgorod. Novgorod arose in the 9th century as one of the earliest centres of the exploitation of the forest hinterland and remained the most important commercial centre of the Kievan period. The changes of the latter Kievan period did not diminish the town's importance, for it benefited both from the increased activity of the Hanseatic League and from the development of the upper Volga region, for which it was a major trade outlet. Although Novgorod was an early base for the Rurikids, the princely traditions characteristic of Kiev and other post-Kievan centres never developed there. When Kiev declined, Novgorod soon (1136) declared its independence from princely power, and, although it accepted princely protectors from various neighbouring dynasties, it remained, until conquered by Muscovy (Moscow), a sovereign city.

During the 13th century, Novgorod's burghers easily found an accommodation with the invading Mongols. In the Mongol period, its energetic river pirates pushed farther north and east toward the Urals and even down the Volga, and Novgorod's prosperity was generally unbroken until the commercial revolution of the 16th century. Its absorption by the growing principality of Muscovy in 1478 ended its political independence and changed its social structure, but Novgorod's characteristic economic and cultural life did not end with that catastrophe.

Novgorod was governed by an oligarchy of great trading boyar families who controlled the exploitation of the hinterland. They chose (from among themselves) a mayor, a military commander, and a council of aldermen, who controlled the affairs of the city and its territories. The town itself was divided into five "ends," which seem to have corresponded to the "fifths" into which the hinterland was divided. There was, in addition, a *veche* ("council"), apparently a kind of town meeting of broad but indeterminate composition, whose decisions, it would appear, were most often controlled by the oligarchy. A major role in politics was played by the archbishop, who after 1156 controlled the lands and incomes previously owned by the Kievan princes and who appears throughout Novgorod's history as a powerful, often independent figure.

The northwest. During this period, much of the territory of the principalities of Smolensk, Polotsk, Turov, and Pinsk was controlled by the grand principality of Lithuania, which was essentially an international or nonnational formation, led by a foreign dynasty (of eastern Lithuanian pagan origins) ruling over predominantly Belorussian and Ukrainian populations. By the 15th century, the dynasty had become Slavic in culture (a version of Belorussian was the official language of the realm), and at its height under Vytautas (1392–1430) it controlled all the old Kievan territory outside of Russia proper—that is, most of present-day Lithuania, Belarus (formerly Belorussia), Moldova (formerly Moldavia), and Ukraine. In 1385 the grand principality joined the kingdom of Poland, and the union was sealed shortly thereafter by the marriage of Grand Prince Jogaila (Jagiełło) to Jadwiga, the Polish queen.

The northeast. The region bounded by the Oka and Volga rivers, later to be the heartland of the grand principality of Moscow, was settled before the arrival of Slavs from Novgorod and the Baltic area by a Finnic tribe. Rostov, the earliest princely centre, was from Vladimir's time included in the princely rotation system. In the 12th century, it became the patrimony of the younger branch of Vladimir II Monomakh's family (who founded the new princely centre Vladimir in 1108). Under his son Yury Dolgoruky (1125–57) and grandson Andrew I (1157–74), the principality reached a high political and cultural development, which it retained through much of the succeeding century. Early in the 13th century, the principality of Moscow was created as an appanage (royal grant) within the grand principality of Vladimir, and this new seat grew in importance when Michael Khorobrit, brother of Alexander Nevsky, conquered Vladimir (1248) and made himself prince of both centres. Daniel, Nevsky's son and the progenitor of all the later Rurikid princes of Moscow, had a long and successful reign (1276–1303), but at his death the principality still embraced little more than the territory of the present Moscow province (an area of 140 miles [225 kilometres] in length and width). The beginning of Moscow's rise to its later preeminence came during the reign of Daniel's son Ivan (1328–41), who, by cooperating with Öz Beg, khan of the Golden Horde, and also by his shrewd purchases (probably of tax-farming rights), greatly expanded the influence of his principality.

Early Moscow

The southwest. The lands of Galicia and Volhynia were always ethnically and economically distinct from the Kievan region proper, as well as from more distant regions. Agriculture was highly developed, and trade, particularly in the valuable local salt, tended to take westward and overland routes. Galicia, already a separate principality by 1100, grew as Kiev declined. Later, Roman Mstislavich of Volhynia (ruled 1199–1205) conquered Galicia and united the two principalities. Under his son Daniel (1201–64), difficulties with the Galician landed magnates and the interference of the Hungarians weakened the principality, and it was subjugated in 1240 by the Mongol invasion. Eventually, this region came under the domination of Lithuania (Volhynia) and Poland (Galicia).

THE MONGOL PERIOD

The Mongol invasion. In 1223, when the first Mongol reconnaissance into former Kievan territory led to the disastrous defeat of a Volhynian-Galician-Polovtsian army on the Kalka River, the Rurikid principalities had for generations been intermittently at war. Kiev was in ruins, Novgorod was preoccupied with commerce and with its northern neighbours, Galicia was being torn internally and drawn increasingly into Polish and Hungarian dynastic affairs, and Vladimir-Suzdal, apparently the leading principality, was unable to resist the finely organized and skillful mounted bowmen of the steppe, the greatest military force of the age.

Pious tradition, born of the works of monkish annalists and court panegyrists, has exaggerated both the destructiveness of the first Mongol conquests and the strength of the resistance. The Mongols aimed to revive, under a unified political system, the trade that had traditionally crossed the Central Asian steppe and vitalized the economy of the pastoral nomads. As they moved westward, they gained the collaboration of groups of Turkic nomads and the predominantly Iranian and Muslim traders in the towns of the old Silk Road; they encountered the greatest resistance in sedentary political centres and among landowning elites. The lands of the Rus presented numerous similarities with the Central Asian areas that the Mongols had already conquered. Here, too, a former commercial empire had fallen apart into an aggregation of warring principalities. Here, too, ready recruits were to be found in the Polovtsians, who controlled the lower Dnieper and Volga and Don, and in the Muslim merchants, who dealt in the towns of the Crimea and the upper Volga. These merchants showed the way, first (1223) to the Crimea and up the Volga to the old centre of Bulgar, later to Ryazan, Rostov, and the Suzdalian towns, and still later (1240) to Kiev and Galicia.

Impact of Mongol victories

Many of the conquered cities made a striking recovery and adjustment to the new relationships. Some towns, such as Kiev, never fully recovered in Mongol times, but the cities of the Vladimir-Suzdal region clearly prospered. New centres, such as Moscow and Tver, hardly mentioned in any source before the Mongol period, arose and flourished in Mongol times.

Thus, the Mongol invasion was not everywhere a catastrophe. The local princely dynasties continued unchanged in their traditional seats; some princes resisted the new authority and were killed in battle, but no alien princes ever became established in Slavic territory. Few Mongols remained west of the Urals after the conquest; political and fiscal administration was entrusted to the same Turkic clan leaders and Islāmic merchants who had for generations operated in the area. The whole of the Novgorodian north remained outside the sphere of direct Tatar control, although the perspicacious burghers maintained correct relations with the khans.

Tatar rule. After a brief attempt to revive the ancient centres of Bulgar and the Crimea, the Jucids (the family of Jöchi, son of Genghis Khan, who inherited the western portion of his empire) established a new capital, Itil. (It was



The empire of the Golden Horde at its greatest extent.
From A. Herrmann, *An Historical Atlas of China*

The Mongol commercial empire

moved to New Sarai, near the site of Tsaritsyn, modern Volgograd, in about 1260.) These towns became the commercial and administrative centres of what was later to be called the "Golden Horde" (the term is probably a Western invention). Its East Slavic territories were tributaries of an extensive empire, including, at its height, the Crimea, the Polovtsian steppe from the Danube to the Ural River, the former territories of the Bulgar empire (including the fur-rich Mordvinian forests and parts of western Siberia), and in Asia the former kingdom of Khwārezm, including Urgench, the cultural capital of the Jucids. Control of the Slavic lands was exercised through the native princes, some of whom spent much of their time at the Mongol capital, and through agents charged with overseeing the activities of the princes and particularly the fiscal levies.

This multinational commercial empire was unstable. Early in the history of the Golden Horde, the khans of Sarai, who tended to reflect the interests of the Volga tribes, were challenged by the tribal princes of the west, whose control of the Danube, Bug, and Dnieper routes and of the access to the Crimea gave them considerable political and economic power. As early as 1260, Nokhai, one of these western chieftains, showed his independence of Sarai by establishing his own foreign policy, and toward the end of the 13th century he seized control of Sarai itself. At his death, the eastern tribes reestablished their control in Sarai, but in the reign of the great Öz Beg (1313–41), the high point of Golden Horde power, the west was again ascendant. Öz Beg based his power upon firm control of the Crimea and had extensive relations with the Genoese and Venetians, who controlled the main ports there. After the death of Öz Beg's son Jani Beg in 1357, however, the empire began to reveal serious internal strains. The tribes of the west paid little heed to the khans who appeared in dizzying succession in Sarai; the northern Russian princes fell to quarreling and to maneuvering for their own advantage in the internecine politics of the Golden Horde; the Volga Bulgar region was detached by a dissident Tatar prince; and the lands of the east were drawn into the orbit of the Turkic conqueror Timur (Tamerlane).

The Golden Horde's last cycle of integration and dismemberment was closely linked with events in Timur's domains. Tokhtamysh, son of a minor Tatar prince, had been unsuccessfully involved in the skirmishes around the throne of Sarai in the 1370s and had fled to the court of Timur, with whose aid he returned to Sarai and vanquished the tribal leaders who had opposed him. Having defeated and made peace with them, he now turned to defeat Mamai (1381), who had the previous year been defeated by Prince Dmitry Donskoy ("of the Don"; grand prince of Moscow, 1359–89). Mamai's western tribal allies went over to Tokhtamysh, and, for a brief time, the major components of the tribal structure of the Golden Horde were reunited. Tokhtamysh successfully attacked Moscow (just as Mamai had hoped to do) and set about consolidating his gains. As his power grew, however, Tokhtamysh was drawn into a struggle with Timur, who had conquered much of Iran, the south Caucasus, and eastern Anatolia. After a number of encounters in the northern Caucasus, Timur, who apparently was intent upon diversion of east-west trade through his own Transoxanian and North Iranian territories, set out to destroy Tokhtamysh and the latter's commercial centres. In 1395 and 1396 Timur's armies systematically annihilated Sarai, Azov, and Kaffa. The Golden Horde never recovered; its subsequent history is a record of struggles among its erstwhile subjects for supremacy and attempts to restore political and commercial stability to the steppe.

The rise of Muscovy. From the beginning of the Tatar period, the Rurikid princes displayed much disunity. During the reign of Öz Beg there was a shift of alignments. The princes of Moscow and their allies, together with Öz Beg and his Crimean supporters, generally opposed the princes of Tver, Pskov, and, intermittently, Novgorod. The major punitive measures directed by Öz Beg against Tver with Muscovite support were a part of this pattern.

The links forged in the 14th century between Moscow and the Crimea (and Sarai, while Öz Beg controlled it) were crucial to Moscow's later preeminence. They not only afforded Moscow a steady and profitable export trade for its furs but (because of contacts between Crimean merchants and Byzantium) also led quite naturally to close relations between the Muscovite hierarchy and the patriarchate of Constantinople. This special relationship was but one of the reasons for the eventual rise of Moscow as leader of the Russian lands. Admirably situated in the northeast, linked with all of the major navigable river systems and with the steppe, close to the major fur-producing regions and to the most intensely settled agricultural lands, served by a succession of shrewd and long-lived princes, Moscow came naturally to a position of preeminence during the 14th century and was best equipped to enter the struggle for the political inheritance of the Golden Horde that followed the destruction of its capitals by Timur.

Factors favouring Moscow

Cultural life and the "Tatar influence." Most traditional scholarship has accepted the notions that (1) the Mongol invasion "destroyed" Kievan culture, (2) the Tatar period was one of "stultification" and "isolation from the West", and (3) "Russian" culture was deeply influenced by Golden Horde culture, in particular by "Oriental" con-

ceptions of despotism. These views do not accord with the evidence and should probably be discarded.

In the first place, it seems incorrect to say that Kievan culture was destroyed. In the shift of the cultural centre of gravity to the numerous regional centres, Kievan traditions were in the main continued and in some cases (*i.e.,* Galician literature, Novgorodian icon painting, Suzdalian architecture) enjoyed remarkable development.

Similarly, the notions of stultification and isolation from the West cannot be supported. The enormous Novgorodian culture sphere, the upper Dnieper territories that eventually came under Lithuanian control, and the principalities of Volhynia and Galicia all had, if anything, closer contacts with western and central Europe than in the previous period.

As to "Tatar influence," in the areas of religion and intellectual life, it was practically nonexistent. Control of formal culture by the Orthodox clergy and Muslim divines and limited contact between the Slavic and Turkic populations prevented it. There is no evidence that any single Turkic or Islāmic text of religious, philosophical, literary or scholarly content was translated directly into Slavonic or any East Slavic vernacular during the period.

Concerning the secular culture of the court and counting house, the situation was radically different. These spheres were controlled by very pragmatic princes, merchants, and diplomats. Here, Slavs and Tatars elaborated together an international subculture whose language was Turkic and whose administrative techniques and chancellery culture were essentially those of the Golden Horde. Slavic merchants took full part in this culture, and the princes of Muscovy, in particular, developed their original court culture and chancellery practices within its context. These borrowings, however, were not of a theoretical or ideological nature, and to ascribe later despotism—and its theoretical basis—to "Oriental" influence is to misunderstand the development of Muscovite absolutism.

The post-Sarai period. The collapse of the Golden Horde saw a growth in the political power of the old sedentary centres—Muscovy, Lithuania, the Volga Bulgar region (which became the khanate of Kazan), and the Crimea. This growth was accompanied by dynastic struggles. This period of recovery also saw cooperation among the emerging dynasties against their internal enemies and toward the stabilization of the steppe.

Even by the end of the 14th century, Moscow's position was by no means as dominating as the cartographers' conventions or the historians' hindsight makes it seem. Other centres—Lithuania, Tver, Novgorod—were as rich and powerful as Moscow; many of the areas nominally subject to the Muscovite princes retained their own dynasties, whose members often broke away and sided with one of Moscow's rivals. Only after a series of dynastic conflicts in the early 15th century did Moscow emerge as the leader of the Russian territory.

The struggle began at the death of Vasily I, a son of Dmitry Donskoy, in 1425. The succession of his 10-year-old son Vasily II was challenged by his uncle Yury, prince of the important upper Volga commercial town of Galich. After many turns of fortune, Vasily II succeeded, with the help of Lithuanian and Tatar allies, in establishing his house permanently as the rulers of Muscovy.

RURIKID MUSCOVY

The gathering of the Russian lands

Ivan III. Ivan III (ruled 1462–1505) consolidated from a secure throne the gains his father, Vasily II, had won in the saddle. The "gathering of the Russian lands," as it has traditionally been known, became under Ivan a conscious and irresistible drive by Moscow to annex all East Slavic lands, both the Russian territories, which had traditionally had close links with Moscow, and the Belorussian and Ukrainian regions, which had developed under distinctly different historical and cultural circumstances. In 1471 Ivan mounted a simultaneous attack upon Novgorod and its upper Volga colonies, which capitulated and accepted Moscow's commercial and political demands. The trading republic, however, retained considerable de facto independence and became involved with the Lithuanian princes in an attempt to resist Moscow. Ivan, using these dealings as a pretext, attacked again, and in 1478 Novgorod was absorbed by Moscow. A Muscovite governor was installed, and 70 Novgorodian boyar families were deported and assigned estates elsewhere to hold in service tenure, being replaced by members of the Moscow military-service class.

Tver suffered a similar fate. Ivan had agreed with Prince Michael Borisovich of Tver to conduct foreign relations in concert and by consultation, but, when the Tverite complained that Ivan was not consulting him on important matters, Ivan attacked him and annexed his lands (1485). By the end of Ivan's reign, there were no Russian princes who dared conduct policies unacceptable to Moscow.

The success of Ivan's expansion was determined by his skillful dealings with the Polish-Lithuanian state, which had expanded down the Dnieper basin and into Slavic territories on the south flank of Moscow. After 1450 a competition developed for control of the numerous semi-independent principalities of the Dnieper and upper Donets regions. In the early 1490s some minor East Slavic princes defected from Lithuania to Moscow. The first phase of the conflict, confined to border skirmishes, ended in 1494 with a treaty ceding Vyazma to Moscow and with the marriage of Ivan's daughter Yelena to Alexander, grand duke of Lithuania. In 1500, on the initiative of Lithuanian defectors, Ivan's armies seized a number of important border towns; this began a war that ended somewhat inconclusively in 1503 with a truce that extended Ivan's border considerably to the west.

The third major element of Ivan's foreign policy comprised his relations with the various Tatar confederations. In the 1470s the Crimean khan Mengli Giray came into increasing conflict with Khan Ahmed of the Golden Horde and became interested in an alliance with Moscow against Ahmed and Lithuania. Ivan, eager to dissolve the connection between Lithuania and the Crimea but not wanting to alienate Ahmed, stalled for time. In 1481, when Ahmed died, Ivan was able to forge an alliance with the Nogais, Mengli Giray, and Kazan. The security provided by this system became the cornerstone of his later policies.

In addition to problems of war and diplomacy, Ivan was faced with a number of challenges from within his own family and court. In 1472 his eldest brother, Yury, died childless, and Ivan appropriated his entire estate. This action antagonized the two eldest surviving brothers, Andrey and Boris, whose grievances were further increased by Ivan's refusal to give them a share of conquered Novgorod. In 1480 they rebelled, and only with difficulty were they persuaded to remain loyal. A more serious conflict arose (1497–1502) in the form of an open and murderous struggle among Ivan's relatives for succession to the throne. Ivan had originally named as his heir his grandson Dmitry, son of his deceased son Ivan and the Moldavian princess Yelena, but a group close to Ivan's second wife, Sofia (Zoë) Palaeologus, opposed this; her son Vasily threatened and perhaps attempted an insurrection, and Ivan was forced to accept Vasily.

The entrance into Europe

Ivan made the first effort on the part of a Muscovite ruler to become involved in the diplomacy of western Europe. He and his advisers realized the need for a counterpoise to the Polish-Lithuanian power, while the diplomats of Rome and Vienna were interested in the possibility of flanking the growing empire of the Ottomans with a Muscovite-Tatar force. In the 1470s and '80s there was an unprecedented traffic between these capitals and Moscow. It was through these channels that Ivan arranged his marriage to Sofia Palaeologus, a niece of the last Byzantine emperor. Sofia has been credited with considerable influence over Ivan, in particular with urging him to adopt the Byzantine political style (*e.g.,* autocracy, state domination of the church, etc.), but in fact she probably had little influence over Ivan's policies. His reign ended on a note of failure, with his overtures to the West and his brief rapprochement with Lithuania both disrupted by the intractable territorial and religious conflicts of the Slavic East and by the opportunism of the local magnates. Moscow's situation was worsened after Ivan's death by the collapse of the alliance with the Crimean khan Mengli Giray upon the khan's death in 1515, opening a new period of chaos and readjustment in the steppe.

A similar appraisal must be made of Ivan's domestic policies. Although his reign was notable for the annexation of the rich Novgorodian provinces and for the establishment of a regular bureaucracy and a land-tenure system, these achievements created new problems for his successors. The system of land grants to military servitors, maintained for centuries (with changes) in all conquered lands, ultimately suppressed the interest of both landlords and tenants in increasing agricultural productivity.

Vasily III. Ivan's son Vasily, who came to the throne in 1505, greatly strengthened the monarchy. He completed the annexation of Russian territories with the absorption of Pskov (1510) and Ryazan (1521) and began the advance into non-Russian territories (Smolensk, 1514).

Faced with a continuing Lithuanian war and with the breakdown of his father's Tatar policy, Vasily carefully temporized in order to avoid uniting his enemies. Once he had secured peace in the west, he was able to deal directly with the khan of the Crimean Tatars. In the end, however, much of what Vasily accomplished was undone by his failure as a procreator: divorcing his first wife for her apparent barrenness, he married Yelena Glinskaya, who bore him only two children—the retarded Yury and the sickly Ivan, who was three years old at Vasily's death, in 1533.

Ivan IV the Terrible. Vasily had been able to appoint a regency council composed of his most trusted advisers and headed by his wife Yelena. The grievances created by his limitation of landholders' immunities and his antiboyar policies soon found expression in intrigue and opposition. The bureaucracy he had relied upon could not function without firm leadership. Although Yelena continued Vasily's policies with some success, on her death, in 1538, various parties of boyars sought to gain control of the state apparatus; a decade of intrigue followed, during which affairs of state, when managed at all, went forward because of the momentum developed by the bureaucracy. Foreign policy was moribund, in spite of considerable opportunities presented by the continuing decay of the khanate of Kazan, and domestic policy vacillated so wildly that scholars cannot agree upon an appraisal of its main directions.

Toward the end of the 1540s, however, there emerged a strong coalition of Muscovite boyars. Apparently inspired by a common awareness of the needs of the state, they ended the debilitating intrigues and embarked upon a thoroughgoing program of reform. The first important step was the reestablishment of the monarch (for the first time to be officially designated tsar), accomplished through the coronation of the 16-year-old Ivan in unprecedented solemnity and pomp. Shortly afterward, he was married to Anastasia Romanovna Zakharina of a leading boyar family.

Although there is a voluminous literature devoted to Ivan, almost nothing is known of his personality, his political views, or his methods of rule. There is little reliable biographical information about him aside from the facts of his six marriages, his lifelong ill health, and his mercurial temper. It is not even known when he began to rule in fact or which of the policies of his reign can be considered his.

Power of politicians over Ivan

Ivan was doubtless a puppet in the hands of the leading politicians long after his coronation. The major reforms of the middle 1550s, which produced a new law code, a new military organization, a reform of local government, and severe restrictions on the powers of hereditary landowners (including the monasteries), were probably the work of the bureaucrats and boyars, their objective being to modernize and standardize the administration of the growing state. The immediate goal was to strengthen the state and military apparatus in connection with major campaigns (the first undertaken in 1547) against the khanate of Kazan and to prepare for the major colonization of the new lands that that conquest and others were expected to secure. Toward the end of the 1550s Ivan seems to have gained the support of certain groups opposed to these policies and to have seized control of the government. The issue was evidently foreign policy. The planned conquest of the Volga and steppe region had been delayed in execution, and the Kazan campaigns had been enormously costly. By 1557, when the campaigns against the Crimea began, there was much opposition in the highest military circles. Ivan took the dissidents' part and for the first time emerged as an independent figure.

Ivan was a disastrously bad ruler, in part because no one had ever anticipated that he would rule. His lifelong ill health and the idiocy of his brother made it quite natural for the regency and the politicians to ignore him and to neglect his education. In adulthood he contracted a painful and incurable bone disease, from which he sought relief in alcohol and in potions provided by a succession of foreign doctors and quacks. Once he had acquired full power, he set about destroying those who had ruled during the interregnum, as well as the machinery of government they had built up.

Ivan established his famous *oprichnina,* an aggregate of territory separated from the rest of the realm and put under his immediate control as crown land, in 1564; this was the device through which he expressed his rejection of the established government. As it was his private domain, a state within the state, he took into it predominantly northern and commercial territories that had enjoyed a special prosperity in preceding decades. Specific towns and districts all over Russia were included in the *oprichnina,* their revenues being assigned to the maintenance of Ivan's new court and household. He established a new, much simplified officialdom and a court composed of sycophants and mercenaries, prone to rule through terror, accompanied by persecution of precisely those groups that had contributed so much to the modernization of the state. As trained statesmen and administrators were replaced by hirelings and cronies, the central government and military organization began to disintegrate. The destructiveness of the *oprichnina* was heightened by Ivan's involvement in the costly and ultimately disastrous Livonian War throughout this period (indeed, some historians have viewed the *oprichnina* as a device for the prosecution of that lengthy war's taxing campaigns). Even before the war had ended, Ivan was forced by the utter incompetence of his special *oprichnina* army to reintegrate it (1572) with the regular army and to revert, in theory at least, to the previous institutions of government. By the time he died, in 1584, the state that he had wanted to reclaim from its makers was in ruins.

Boris Godunov. Ivan had murdered his eldest son, Ivan, in a fit of rage in 1581, and his only legitimate heir, Fyodor, was an idiot. Power passed to those who were at Ivan's deathbed, among whom Boris Godunov, who had capped a rapid rise in court circles with the marriage of his sister Irina to the pathetic Fyodor, soon emerged as the leading force. Godunov's judicious combination of chicanery, vision, and force enabled him to disarm his most dangerous enemies and to have himself proclaimed tsar after Fyodor's death in 1598. His policies during Fyodor's reign had been consistently restorative and conciliatory, and he had apparently succeeded in repairing much of the damage done to the state in Ivan's time. He conducted a cautious and generally successful foreign policy: the 20 years of his reign were, except for a short, successful war against Sweden, peaceful. In domestic policy, he returned to the modernizing and standardizing policies of the mid-century. He reorganized the land-tenure system, commerce, and taxation.

The struggle for the succession

For a number of his problems Godunov could find no solution. Chief among these were the depopulation of the central Muscovite lands and the discontent among small landholders in the territories recently acquired in the south and southwest. Added to these problems was the continuing opposition of the boyars.

In spite of these difficulties and widespread famine caused by crop failures in 1601–02, Godunov remained well in control of the situation until the appearance of the so-called first False Dmitry, a defrocked monk who had appeared in Poland in 1601 claiming to be the son of Ivan IV. (The true Dmitry had died during an epileptic seizure in 1591.) The False Dmitry found some supporters in Poland—notably Jerzy Mniszech, to whose 15-year-old daughter, Maryna, he became engaged. As the impostor moved northeast toward Muscovy, he acquired growing

support among the disaffected petty gentry and Cossacks (peasants who had escaped from serfdom to a nomadic life) of the regions through which he passed, and border cities throughout the south opened their gates to him. Godunov's troops easily defeated the ragtag force, which apparently had many secret supporters among Muscovite boyars, but a few weeks later Godunov died. The boyars staged a coup against Godunov's family and declared Dmitry tsar. The pretender entered Moscow in triumph, was crowned, and married Maryna Mniszchówna.

The Time of Troubles. The events of 1606–13 cannot be captured in a few words. Chaos gripped most of central Muscovy; Muscovite boyars, Polish-Lithuanian-Ukrainian Cossacks, and assorted mobs of adventurers and desperate citizens were among the chief actors. In May 1606 a small-scale revolt supported by popular indignation at the foolishly insulting behaviour of Dmitry and his Polish garrison brought the overthrow and murder of the pretender. The boyars gave the crown to Prince Vasily Shuysky, a leader of the revolt against Dmitry, with the understanding that he would respect the special rights and privileges of the boyars. While the new tsar had the support of most boyars and of the northern merchants, he could not end the disorders in the south or the adventures of the Polish and Swedish kings, who used Muscovy as a battlefield in their continuing conflict with each other. In 1608 a number of boyars, led by the Romanovs, went over to a second False Dmitry, who had ridden a wave of discontent and freebootery from the Cossack south into the centre of Muscovy. A kind of shadow government was formed in the village of Tushino, 9 miles (14 kilometres) west of Moscow, in which the boyars and bureaucrats of the Romanov circle took leading posts. It managed to gain Cossack support and to manipulate Dmitry's pretensions while negotiating with the Polish king Sigismund III on terms by which his son Władysław IV might become tsar. Shuysky, in desperation, turned to Sweden for aid, promising territorial concessions along the Swedish-Muscovite border. At this the Polish king invaded Muscovy and besieged Smolensk (September 1609). The Tushino coalition dissolved, and Dmitry withdrew to the south. The position of the Shuysky government deteriorated, and in 1610 the tsar was deserted by his army and his allies. The boyars formed a seven-man provisional government with the aim of installing a Polish tsar. This government proved unable to settle its affairs and to restore order to the country. A new insurgent army, financed by northern merchants and staffed with Swedish troops, marched on Moscow with the intention of ousting the Polish garrison and of bringing the various Cossack bands under control. It nearly gained Moscow but fell apart because its leadership could make no arrangement with the Cossack leaders.

Cossack capture of the Kremlin

A year later, a second force, raised in the same northern cities and supported by Cossacks who had been part of the Tushino camp, was able to take possession of the Kremlin. A call was issued for the election of a new tsar.

Social and economic conditions. In the flux of social and economic life in the 15th and 16th centuries, three interconnected processes may be observed: a steady economic growth, mainly from colonization and trade; an expansion in the power of the central government; and the encroachment of the nobility upon the lands previously held by the free peasantry, accompanied by the reduction of the bulk of the peasantry to serf status.

In the middle of the 15th century, society and the economy were still organized along traditional lines. The land was sparsely settled. Life for most of the population was simple and probably close to the subsistence level. Serfdom did not yet exist. Most of the peasantry lived on state lands and paid whatever taxes could be extracted from them by their prince or his bailiff.

A number of changes occurred in this pattern in the latter part of the 15th century. About 1460 measures were taken to bring the peasantry under more regular control of the state and the landlord. Peasant registration appeared at this time, and also the requirement spread that peasants might renounce the tenancy of the land they were working only at the end of the agricultural cycle, in the week of St. Yury's Day (November 26 [December 8, New Style]). The growing controls upon the peasantry received impetus from the large-scale deportations and colonizations that accompanied the annexations of Novgorod, Tver, Pskov, and Ryazan, when the old nobility were replaced with nobility owing service to the prince of Muscovy. The nationwide promulgation of the restriction on movement to St. Yury's Day was contained in the law code of 1497, which added the stipulation that peasants leaving a former situation must pay the landlord all arrears in addition to a departure fee. All of the measures, together with the expansion of the state apparatus for tax gathering and adjudication of disputes over land and peasants, were associated with the growing complexity and power of the central government.

The law code of 1550 repeated the stipulation of 1497 limiting peasant departure, but with much more specific provisions and stronger sanctions. Other reforms put an end to local administration by rotating military governors and limited monastic landholding and the juridical rights of landlords over their peasants. The events and policies of the latter half of the reign of Ivan IV destroyed many of the beneficial results of the reforms. The Livonian War imposed unprecedented burdens upon the taxpaying population and the landowning military caste. The political disruption caused by Ivan's *oprichnina* further undermined the position of the service class and led to the looting of Novgorod and other towns. At the same time, other new trends provided the basis for economic growth: trade in local and Oriental transit goods, organized through Arkhangelsk, primarily by English and Dutch merchants, brought unprecedented wealth and luxury to the court; the opening of Siberia provided additional income; and the extension of Russian agriculture into the steppe promised, for the first time, agricultural prosperity.

Cultural trends. This period also saw the crystallization of that complex of forms and ideas that can, for the first time, be identified as Russian culture. There was a gathering and integration of the Novgorodian, Tverite, and Suzdalian cultural traditions. Moscow began to attract the artists, craftsmen, and learned monks who built the eclectic but "national" churches of Ivan III's otherwise Italianate Kremlin and who wrote the revised national, pro-Muscovite versions of the chronicles that had been kept in Rostov, Ryazan, and Novgorod. The regional traditions were not always easily reconciled. Novgorodian attitudes, in particular, clashed with those of Muscovy.

The crystallization of Russian culture

The reign of Ivan III saw a marked turning toward the West. Ivan surrounded himself with Italian and Greek diplomats and craftsmen. His palace of 1487, his Kremlin with its Latin inscription over the main gate, and his churches, the original aspect of which has been altered by successive Russifying restorations, were clearly in the Italian style, as contemporary foreign visitors noted. His marriage to Sofia Palaeologus had, in addition to its diplomatic significance, a symbolic function of bringing Ivan into the circle of Western princes. Muscovy supposedly regarded itself as the heir of Byzantium and as the spiritual leader of the Orthodox world. It may be that the church leadership, militantly anti-Roman, thought of itself in this light. Ivan and many around him viewed the Byzantine heritage as Western, by contrast with the Ottoman and Tatar world, and were at pains to associate Muscovy with Western traditions and interprincely relations. This striving to be accepted in the Western world marked most of the changes in regalia and style of Ivan's reign, although these were later to be buried in the lore of Muscovite Byzantinism.

Three significant causes can be discerned for the evolution of Muscovite culture in the 16th century. The first was the growth and prosperity of the Russian population, united under a stable and increasingly centralized monarchy, which produced the conditions for the rise of a national culture. The second was the diplomatic and cultural isolation in which Muscovy found itself, particularly in the first half of the 16th century, as a result of hostile relations with increasingly powerful Lithuania and Poland, a cause that, more than any other, brought an end to Ivan III's westward turn and to the revolutionary adjustments of the age of exploration. The third cause was the resolution of

church-state relations, in the course of which the church submitted to the power of the princes in politics but gained control over the culture, style, and ideology of the dynasty, producing the peculiar amalgam of nationalistic, autocratic, and Orthodox elements that became the official culture of high Muscovy. This new synthesis was reflected in the great undertakings associated with the name of Metropolitan Makary of Moscow: St. Basil's Cathedral in the Kremlin; the encyclopaedic Menolog, or calendar of months, which contained all the literature, translated and original, permitted to be read in the churches; and the *Illustrated Codex,* a compilation of East Slavic and Greek chronicles in an official Muscovite version.

ROMANOV MUSCOVY

Establishment of the Romanovs

Michael. The military drive that finally expelled the Poles from Moscow led to the election of Michael (Mikhail Fyodorovich), the 16-year-old son of Fyodor Romanov, as the new tsar. The composition of the coalition that elected him is not clear, but he evidently represented a compromise among the Cossacks, the boyars (especially the Tushino boyars), and the leaders of the northern army. It would be difficult to imagine circumstances less favourable for the beginning of the reign of the adolescent monarch and a new ruling coalition. The military campaigns had left much of the central and southwestern portions of the country in ruins. In many areas, populations had fled, land lay fallow, and administration was in disarray. Significant portions of the Novgorod, Smolensk, and Ryazan regions were occupied by Swedish and Polish armies and by sundry insurrectionary forces, who threatened to renew hostilities.

The Romanov government required more than a decade to establish itself politically and to restore economic and social order. Few had expected the election of a new tsar (the fourth in eight years) to bring an end to the turmoil. But the election of Michael reflected a resolution of political forces that permitted the coalition government to address itself to the problems of reconstruction. Another cause was the survival of the central bureaucracy; the civil servants in Moscow had served all successive governments without much interruption and were ready to restore administrative regularity as soon as political order was established. Fortunately, the new government refrained from involving itself in the Polish-Swedish conflicts, which reached their height at this period. This restraint was a most important element in the success of the 1613 settlement, for the international situation was, if anything, grimmer than the domestic. Polish-Swedish differences permitted Muscovite diplomats to bring the two countries to separate truces (Sweden, 1617; Poland, 1618); although these left substantial territories under the control of Poland and Sweden, they provided a needed interlude of peace. The Romanov government wisely avoided any significant participation in the Thirty Years' War, in which most European states engaged. At the death of the Polish king Sigismund III in 1632, Muscovy made an ill-advised attempt to regain Smolensk that ended in military disaster, but in 1634 it obtained Władysław's formal abjuration of the Polish king's questionable claim to the title of tsar.

After the failure of the Smolensk campaign, the government refrained from further military involvement with Poland for nearly a generation. It concentrated instead upon the extension and fortification of its southern borders, where the incursions of Crimean Tatars were an impediment to colonization. Moscow, however, was not prepared to go to war with the Ottomans, who were the protectors of the Crimean khan; when in 1637 the Don Cossacks, Muscovy's clients, captured the critical port of Azov and appealed to Moscow for aid in holding off a counterattack, a *zemsky sobor,* or national assembly (see below *Trends in the 17th century*), decided not to intervene, and the port was lost.

Domination of Morozov

Alexis. The reign of Michael's son Alexis (Aleksey Mikhaylovich), whom later generations considered the very model of a benevolent and gentle tsar, began badly. Like his father, Alexis came to the throne a mere boy. Immediately the boyar who controlled the government, Boris Ivanovich Morozov, embarked upon policies that brought the government to the brink of disaster. Morozov cut government salaries; he also introduced a tax on salt and a state monopoly of tobacco, the first of which caused widespread hardship and discontent, while the second brought the church's condemnation. At the same time, he alienated boyar groups close to the throne by his interference in his ward's marriage.

Morozov's actions exacerbated an already dangerous situation in the country. The city populations and service gentry, in particular, were heavily burdened by taxes and other obligations and were increasingly angry at the growing wealth and power of the ruling clique. During a riot in Moscow in May 1648, a mob surrounded the 19-year-old tsar, demanding the execution of Morozov and the leading officials. Some of the latter were thrown to the mob, and a brief protective exile was arranged for Morozov. Morozov's boyar enemies, who may have abetted the riot, took control of affairs and carried out a series of reforms. The salt tax and tobacco monopoly were ended, and a commission was established for the drafting of a new law code. Serious disorders continued in the cities of the north, particularly in Pskov and Novgorod, where force was required to reimpose authority.

In Novgorod the principal actor in the government's interest was the metropolitan Nikon, an energetic and authoritarian monk who had made influential friends in Moscow while archimandrite at the Romanov family church and continued assiduously to cultivate the tsar and his relations while in Novgorod. In 1652 his solicitations earned him the patriarchate. Tradition has it that, before accepting the position, Nikon demanded a declaration of full obedience in religious and moral matters from the tsar. In the first years of Nikon's tenure, his relations with Alexis and the court were good. The patriarch, with official support, carried out a number of liturgical and organizational reforms, surrounding himself with an impressive bureaucracy modeled upon the state apparatus. Relations with the tsar became strained in 1658, however, and, after he was publicly snubbed by Alexis, Nikon announced that he was abandoning the patriarchate. He later held that he had simply gone into temporary seclusion, but his effective power and influence were at an end.

The main event of Alexis' reign was the annexation of eastern Ukraine. His government had continued the previous policy of avoiding entanglements in the West while expanding eastward but could not resist the opportunity offered in 1654 when Bohdan Khmelnytsky, the leader of a Cossack revolution against Polish rule in Ukraine, appealed to Moscow for the help he had been unable to obtain from Sweden and the Turks. Moscow accepted his allegiance in return for military assistance and thus became involved in a protracted struggle with Poland and Sweden for the Ukrainian, Belorussian, and Baltic territories. At first the war went well, but the differing objectives of the Ukrainian and Muscovite allies soon revealed themselves. When Charles X of Sweden entered the fray against Poland, Alexis made peace, in 1656; he feared a strong Sweden as much as a strong Poland. Muscovite forces plunged into war with Sweden for the Estonian, Livonian, and Karelian territories along the Baltic coast. The situation in Ukraine became increasingly confused and dangerous for Moscow, and it was necessary to end the war with Sweden in 1661, even at the cost of yielding, once again, the Baltic coast.

Territorial expansion in the 17th century

In Ukraine the war took on a new aspect when in 1664 Peter Doroshenko, a new leader, put himself under the protection of the Ottomans. The Turks joined in a number of major military operations, alarming both Poland and Moscow sufficiently to bring them to a truce at Andrusovo (1667). Poland recognized Moscow's control over eastern Ukraine and Kiev, while Moscow yielded the part of Ukraine west of the Dnieper and most of Belorussia.

The peace did not greatly improve the government's position, for the same year saw the beginning of a threatening movement among the Don Cossacks and peasants of the Volga region, led by Stenka Razin, and a political battle within the inner circles at court, caused by the death of Alexis' wife. After two years, Alexis was married to Nataliya Naryshkina. In 1676, however, Alexis

himself died, and Fyodor, a sickly son of his first wife, Mariya Miloslavskaya, succeeded him. A struggle began between the rival Naryshkin and Miloslavsky families. The Naryshkins were exiled, and the Miloslavskys, with their clients and supporters, took over. In 1682, however, Fyodor died, and the Naryshkin faction sought to place his half brother Peter on the throne instead of Fyodor's full brother, the ailing Ivan. The elite corps of *streltsy* (a hereditary military caste) revolted and established Ivan's elder sister Sophia as regent. (For the accession and reign of Peter the Great, see below *The 18th century: The reign of Peter I the Great (1689–1725).*)

Trends in the 17th century. Economic reconstruction was slow, particularly in agriculture and in the old central lands, but it was accompanied by a growth of trade and manufacturing. The state revenues profited from the expansion eastward beyond the Urals and southward into the black-soil region. In the north the port of Arkhangelsk handled the export of forest products and semimanufactures (naval stores, potash) to the English and Dutch, and its merchants took a leading role in the early exploitation of Siberia. The government itself became deeply involved in the development of trade and commerce, both through its monopolistic control of certain areas and commodities and by its efforts to build up such strategic industries as metallurgy. The economy grew at unprecedented speed during the 17th century. By 1700 Russia was a leading producer of pig iron and potash, and the economic base on which Peter's military successes were to depend had been firmly established.

The political recovery of the Russian state after the so-called Time of Troubles was due largely to the survival of the central bureaucracy and ruling oligarchy. The lines of subsequent development were determined by the growth, consolidation, and almost unimpeded self-aggrandizement of these groups in the 17th century. The expansion of the bureaucratic apparatus can be measured in various ways. In 1613 there were 22 *prikazes,* or departments; by mid-century there were 80. At the beginning of the period, the jurisdiction of the bureaucracy included primarily fiscal, juridical, and military matters; by the end of the century, it also covered industrial, religious, and cultural life. At the close of the Time of Troubles, the bureaucracy's functions were exercised by leading boyars and professional administrators; by Peter's time the mercantile class, the whole of the nobility, and the clergy had become part of its ubiquitous network. This bureaucracy was the buttress—indeed, the substance—of an absolute monarchy whose prerogatives knew few internal bounds.

The expansion of central authority

The ease with which the extension of central authority overwhelmed all other political and social forces is to be explained by the frailty of local institutions and by the absence of independent ecclesiastical or social authority. The Muscovite administration was extended first into the devastated areas, where local institutions had been swept away, and then into new territories that had no significant political institutions, until it became a standardized and centralized mechanism powered by the colossal wealth generated by its own expansion.

These processes were reflected in the great law code of 1649, the first general codification since 1550, which was to remain the basis of Russian law until 1833. Its articles make clear the realities of Muscovite political practice: the rule of the bureaucrats and the extension of the powers of the state into all spheres of human activity. It was based in large measure upon the accumulated ad hoc decisions of the officials and was intended for their guidance. The code made ecclesiastical affairs a matter of state jurisdiction; it gave legal expression to the practice of serfdom; and, in an important new article, it enumerated crimes "of word and deed" against the "Sovereign"—by which were to be understood the state and all its agents.

Social development paralleled and was to a great extent determined by the developments just described. By the end of the century, only those families that had made new careers in the state apparatus through service as generals, ministers, and ambassadors remained at the apex of society; they were joined by numerous parvenu families that had risen in government service. Particularly striking was the prosperity of the *dyak* class of professional administrators, which had become a closed hereditary estate by a decree of 1640; this class had become a new and powerful "nobility of the seal" that was to survive into modern times.

During much of the 17th century, the government was run, for all practical purposes, by high officials in cooperation with relatives and cronies of the reigning tsar. Nineteenth-century historians, eager to find constitutional traditions in Russia's past, have stressed the role of the *zemsky sobor*—an assembly of dignitaries that from the time of Ivan IV had been called together when matters of crucial importance had to be decided. In the period after 1613 it was in almost continuous session for some years. After 1619, however, the services of these assemblies were no longer required. It is questionable whether they ever had, in law or in fact, any power beyond that of a crowd of military and administrative leaders. They were summoned by the government, and their composition was determined by the government.

Cultural life. No period of Russia's cultural history has been as full of change, turmoil, creativity, failure, and sheer destructiveness as the 17th century. Russian society emerged from the Time of Troubles shattered and unsure of itself, disoriented and impoverished. This shaken society was then subject to wrenching social and economic change and strong external influences.

The old culture, in its formal aspects, had been the culture of the monasteries. Art, literature, architecture, and music remained, until the end of the 16th century, traditional, canonical, and orthodox.

The emergence of a new elite

The 17th century produced, first among the officials and boyars, later among the merchants and middle classes, a new elite that was increasingly interested in European culture and that had mainly secular interests. Yet the government of these same officials and boyars worked to stifle native cultural development, and many of these merchants and nobles were drawn into movements opposed to Westernization.

The reasons for this paradoxical development were three. First, Western culture had reached Muscovy largely through Polish and Roman Catholic mediation, which rendered it unacceptable to all but those sophisticated enough to take a very broad view of the events of the Time of Troubles. In the Ukrainian and Belorussian territories, the Polish Counter-Reformation had brought a national cultural revival. The books, ideas, and people flowing from these lands into Muscovy in the 17th century, however, were hardly less suspect than those of Roman Catholic Poland, and, as these "aliens" acquired a dominant position in Muscovite cultural affairs, resentment was added to suspicion.

A second reason for the character of Muscovite cultural development in the 17th century was the preponderant role of the church and, later, of the state, which took over at last the assets, liabilities, and responsibilities of the ecclesiastical establishment. From 1620, when the patriarch Philaret pronounced an anathema upon "books of Lithuanian imprint" (in effect, the only secular books in print for the Russian reader) until the end of the century, when the government turned to imposing Greek and "Lithuanian" (*i.e.,* Ukrainian and Belorussian) views upon a resisting populace, the state and its ecclesiastical adjunct had a repressive and stultifying influence.

Finally, indigenous cultural forces were, for various reasons, unable to assert themselves. They were physically dispersed, socially diverse, and set at odds by cultural and political disaffection. The development of a vernacular literature, which can be seen in the synthetic "folk songs," pamphlets, tales, and imitations produced for and by the growing educated class, remained a marginal phenomenon; they were unpublished because of the ecclesiastical monopoly of the press, and they were anonymous. The promising experiments of a group of noble writers who worked within the formal Slavonic tradition were ended by exile and repression.

Despite these negative influences, the court itself, especially in the time of Alexis, was a centre of literary and artistic innovation, and many of the leading men of

the realm were considered cultured and cosmopolitan by Westerners who knew them.

The great schism. The contradictions of the age were reflected in the great schism within the Russian church. The doctrinal debate began over obscure and petty matters of ritual, but larger, unarticulated issues were at stake. Religion after the Time of Troubles had taken two directions, which were at first closely associated: the reformation of religious life (with stress on the pastoral functions of the clergy and the simplification of the liturgy) and the correction and standardization of the canonical books (which had come to vary widely from the Greek originals). The government had at first supported these linked objectives, but the supporters of "Old Russian piety" fell into opposition to the reforms as they were officially promulgated.

Enforcement of religious reforms

When, in the 1650s, the patriarch Nikon began to enforce the reforms in the parishes, where they had been generally ignored, the discontent developed into a massive religious and regional insurrection. Towns and parishes of the north were riven by warring "old" and "new" bishops. The Old Believers were either crushed by government force, driven to self-destruction, or reduced to silent resistance.

In the end, the Western secular culture fostered at the court and the new religious culture and education, spread by Ukrainians and Belorussians, who came to dominate church life, submerged and displaced the disparate beginnings of a modern synthesis within native matrices and cleared the way for Peter's cultural policies, which erected a Western facade over the ruins of the native traditions.

(E.L.K.)

The 18th century

THE REIGN OF PETER I THE GREAT (1689–1725)

Peter's youth and early reign. The accession of Peter I ushered in and established the social, institutional, and intellectual trends that were to dominate Russia for the next two centuries. Both Russian and Western historians, whatever their evaluation of Peter's reign, have seen it as one of the most formative periods of Russia's history. The seminal nature of the reign owes much to Peter's own personality and youth. The child of his father's second marriage, Peter was pushed into the background by his half brother Fyodor and exiled from the Kremlin during the turbulent years of the regency (1682–89) of his half sister Sophia. He grew up among children of lesser birth, unfettered by court etiquette. Playing at war and organizing his young friends into an effective military force, he could manifest his energy, vitality, and curiosity almost untrammeled. He also came into close contact with the western Europeans who lived in Moscow; the association kindled his interest in navigation and the mechanical arts—of which he became a skilled practitioner—and gave him the experience of a socially freer and intellectually more stimulating atmosphere than he might otherwise have had. He resolved to introduce this more dynamic and "open" style of life into Russia, a goal he pursued after the overthrow of Sophia in 1689 and that he erected into a policy of state after he became sole ruler following the death of his mother in 1694. (His half-witted half brother, Ivan V, remained co-tsar but played no role and died in 1696.)

Peter's first political aim was to secure Muscovy's southern borders against the threat of raids by Crimean Tatars supported by the Ottoman Empire. For lack of adequate sea power, his initial attempt, in 1695, failed to gain a foothold on the Sea of Azov. Undaunted, Peter built up a navy—he was the first Russian ruler since early Kievan times to do so—and succeeded in capturing Azov a year later. The experience convinced him of the necessity of extending his own technical knowledge and of securing tools and personnel from the West. To this end Peter traveled to western Europe, something no Muscovite tsar had ever done; he spent almost a year in Holland and England acquiring mechanical and maritime skills, hiring experts in various fields, purchasing books and scientific curiosities, and carrying on diplomatic negotiations for a crusade against the Turks. In the course of negotiations with Poland-Saxony and Denmark, an alliance was formed, not against Turkey but against Sweden. The alliance led to the Great Northern War (1700–21), which became Peter's major concern for almost the remainder of his reign.

The Great Northern War

The war started inauspiciously for Peter when King Charles XII of Sweden, disembarking suddenly on the eastern coast of the Baltic Sea, inflicted a severe defeat on the Russians before the fortress of Narva (November 1700). Thinking that he had eliminated Russia as a military factor, Charles invaded Poland to force King Augustus II to make peace and to install his own candidate, Stanisław Leszczyński, on the Polish throne (Stanisław I, ruled 1704–09, 1733). In the meantime Peter proceeded to reorganize and equip his troops systematically, while the generals B.P. Sheremetev and A.D. Menshikov gradually conquered the Swedish Baltic provinces of Ingria and Livonia. By terms of the capitulations of Riga and Revel (now Tallinn), Swedish sovereignty was ended and the provinces incorporated into the Russian Empire; the local German landed nobility and urban patriciate were confirmed in their historic corporate privileges. In 1703 Peter laid the foundations of his new capital, St. Petersburg (called Leningrad between 1924 and 1991), at the mouth of the Neva River; the site was chosen to secure a firm footing on the Gulf of Finland and to open direct sea access to western Europe.

Having forced Augustus II to withdraw from the war (Treaty of Altranstädt, September 1706), Charles again turned eastward. Invading Russia in 1708, he decided to first secure Ukraine as a source of supplies and manpower (promised by the Cossack hetman Ivan Stepanovich Mazepa, who had defected from Peter's side) and await reinforcements from the north. These reinforcements, however, were prevented from reaching Charles by Menshikov's victory at Lesnaya in September 1708. After much maneuvering, Charles laid siege to the Ukrainian town of Poltava in the spring of 1709. Peter hastened to relieve the town, and it was before its walls that the crucial battle was fought on June 27 (July 8, New Style), 1709. Russian victory was complete—Charles and Hetman Mazepa barely escaped capture, and the remainder of their troops were taken prisoner when they tried to cross the Dnieper at Perevolochnaya a few days later. Charles took refuge with the Turkish army encamped on the banks of the Prut River. Peter made the mistake of pursuing him into Turkish territory and barely escaped entrapment by the Turks, whom Charles had persuaded to renew war with Russia. With the help of bribery and diplomacy, Peter extricated himself from the trap by signing a peace treaty (July 1711) under which he gave up Azov and promised to dismantle fortresses near the Turkish border.

Charles remained interned in Turkey (he did not escape until 1714), hoping to rebuild a coalition and rejecting all peace proposals. The war dragged on: Augustus II recovered the Polish throne, and Peter consolidated his hold on the Baltic by invading southern Finland. Russia won its first significant naval victory in July 1714 off the Hangö (Gangut) peninsula and raided the Swedish mainland. The death of Charles XII (killed accidentally in Norway in 1718, soon after his return from Turkey) led to protracted negotiations (Congress of Åland) that ultimately resulted in the Peace of Nystad (Aug. 30 [Sept. 10, New Style], 1721), under the terms of which Sweden acquiesced to Russian conquests on the eastern coast of the Baltic Sea. Thereafter Russia was the dominant power in the Baltic region, while Sweden rapidly sank to second-rate status; Russia meddled in Sweden's political affairs throughout the 18th century.

The German elites

Russia's acquisition of Ingria and Livonia (and later of Kurland) brought into the empire a new national and political minority: the German elites—urban bourgeoisie and landowning nobility—with their corporate privileges, harsh exploitation of native (Estonian and Latvian) servile peasantry, and Western culture and administrative practices. Eventually these elites made significant contributions to the imperial administration (military and civil) and helped bring German education, science, and culture to Russian society. From a diplomatic point of view, Peter's triumph over Sweden secured for Russia an important

voice (enhanced by matrimonial connections) in the affairs of the German states; this culminated in Catherine II's guarantee of the constitutional integrity of the Holy Roman Empire (see below *The reign of Catherine II the Great (1762–96)*). By the same token Russia was to be drawn into all the diplomatic and military conflicts that beset western and central Europe throughout the 18th century, most particularly in connection with the rise of Prussian power, the decline of the Ottoman Empire, and the domestic turmoil in Poland. It forced Russia to maintain great military strength, which naturally put a heavy burden on the fiscal, social, and economic development of the empire.

The long war's requirements determined most domestic policy measures as well. Only when victory was well in sight could Peter devote more of his attention to a systematic overhaul of Russia's institutions. The hastiness and brutality of steps taken under the stress of war had an effect on subsequent history. Historians have debated whether Peter's legislation was informed by an overall plan based on more or less clearly formulated theoretical considerations or whether it was merely a series of ad hoc measures taken to meet emergencies as they occurred. Pragmatic elements predominated, no doubt, over theoretical principles. The prevailing intellectual climate and administrative practices of Europe, however, contributed to orient Peter's thinking.

The Petrine state. Formally, Peter changed the tsardom of Muscovy into the Empire of All Russias, and he himself received the title of emperor from the Senate at the conclusion of the peace with Sweden. Not only did the title aim at identifying the new Russia with European political tradition, but it also bespoke the new conception of rulership and of political authority that Peter wanted to implant: that the sovereign emperor was the head of the state and its first servant, not the patrimonial owner of the land and "father" of his subjects (as the tsar had been). Peter stressed the function of his office rather than that of his person and laid the groundwork of a modern system of administration. Institutions and officials were to operate on the basis of set rules, keep regular hours and records, apply laws and regulations dispassionately, and have individual and collective responsibility for their acts. Reality, of course, fell far short of this ideal, because Muscovite traditions and conditions could not be eradicated so rapidly. Furthermore, there was a great shortage of educated and reliable persons imbued with such rationality and efficiency (a problem that bedeviled the imperial government until its end). They were mainly to be found in the military establishment, where officer and noncommissioned ranks acquired the requisite outlook, experience, and values in the army and navy established by Peter.

Military organization

The changeover from the traditional militia-like military organization to a "European" professional army (as it developed in the course of the so-called military revolution of the 17th century) had been initiated during the reigns of Tsars Michael and Alexis. But it was Peter who gave it the full-fledged "modern" form it retained until the middle of the 19th century. The army—and, for the first time in Russia, the navy as well—was manned by recruits drawn from the peasantry (and other taxable groups) whose service obligation was for 25 years (*i.e.,* virtually for life). Recruitment entailed liberation from serf status both for the soldier and for all his children born after his recruitment. Eventually this provided a path, however steep and narrow, for lower-class children to follow to join the ranks of petty officialdom and nobility. Submitted to cruel and savage discipline, the soldier was isolated from direct contact with the population, and his total commitment was to the state. Drilled in modern battle order and technology, the peasant recruit was forcibly "modernized," and there are indications of some minimal influence of this on the population at large.

The officer corps was recruited in similar fashion from the landowning service class. The terms of service prevailing in Muscovite times, however, were transformed radically. The young noble serviceman was called to serve from age 15 until his death or total incapacity. In principle, service was permanent with only rare leaves granted to attend to family and estate matters. Called up individually, the service noble was assigned and transferred at the will of the state. In principle, he was remunerated by regular salary payments, though in the reign of Peter I, and for long afterward, the salary was paid neither promptly nor fully in cash; officers still had to rely on their family estates or special gifts and awards. Starting as a soldier and noncommissioned officer, the service noble was to progress through the ranks on the basis of merit and longevity; eventually the latter became the principal criterion. Minimum educational standards had to be met by officers and officials, and they came to play a crucial role in both the careers and the social status of the service nobility. The empire's large population, which grew at a rapid rate throughout the century, enabled the government to maintain the largest standing army in Europe. Good generalship and the soldiers' loyalty and resilience, as well as excellent artillery and cavalry, made for a formidable military force that achieved the notable expansion of the empire during the 18th century. The Russian bureaucracy, whose members were often drawn from the military, thus acquired a preference for uniformity and militarism that did not foster respect or concern for the individual needs of the various regions and peoples of the far-flung empire.

In the new administration, performance was to be the major criterion for appointment and promotion. Peter wanted this principle to apply to the highest offices, starting with that of the emperor himself. As a result of his bad experience with his own son, Alexis (who fled abroad, was brought back, and died in prison), Peter decreed in 1722 that every ruler would appoint his own successor. He did not have the opportunity to avail himself of this right, however, and the matter of regular succession remained a source of conflict and instability throughout the 18th century. Peter's concern for performance lay at the basis of the Table of Ranks (1722), which served as the framework for the careers of all state servants (military, civil, court) until the second half of the 19th century. In it the hierarchy was divided into 14 categories, or ranks; theoretically one had to begin at the bottom (14th rank) and proceed upward according to merit and seniority. Throughout the 18th century the 8th rank (1st commissioned officer grade) automatically conferred hereditary nobility on those who were not noble by birth. In a sense, therefore, the Table of Ranks opened all offices to merit and thus democratized the service class. But, because service was contingent on good preparation (*i.e.,* education), it was accessible only to the few—nobility and clergy—until later in the 18th century.

Policies toward the social classes

The same need for qualified personnel that had brought about the Table of Ranks also determined Peter's policies toward the several social classes of his realm. The traditional obligation of members of all estates to perform service to the state, each according to his way of life (*i.e.,* the nobleman by serving in the army and administration, the peasantry and merchants by paying taxes, the clergy by prayer), was given a modern, rational form by Peter. Paradoxically, the reform helped to transform the traditional estates into castelike groups from which—except in rare instances of clergy and rich merchants—it became impossible to escape. The nobility was most directly affected by the change, not only in Peter's lifetime but under his successors as well. The nobleman's service obligation became lifelong, regular, and permanent. The staffs of military and government institutions were no longer recruited on the basis of regional origin or family ties, but strictly according to the need of the state and the fitness of the individual for the specific task at hand. The serviceman was transferred from one assignment, branch, or locality to another as the state saw fit. The office of heraldry within the Senate kept the service rosters up to date and decided on appointments and transfers. It was not easy to break traditional family and clan ties, however. Family connection continued to be a factor in successful service careers, especially if a relative was close to the ruler or the favourite. On the level of the central government and the court, the struggle between cliques for imperial favour was the major factor in determining policy orientations and appointments to high positions.

Peter also introduced single inheritance of real estate (1714), attempting in this way to break the traditional inheritance pattern that had led to the splintering of estates. In so doing he hoped to create a professional service nobility unconnected with the land and totally devoted to the state, but the resistance the law met in its application forced its revocation in 1731. He also required the nobility to be educated as a prerequisite for service. Schooling, whether at home or in an institution, became a feature of the nobleman's way of life. Schooling was a radical innovation, at first resented and resisted; but within a generation it was accepted as a matter of course and became the decisive element in the status and self-image of the nobility.

The peasantry had been enserfed during the 17th century, but the individual peasant had retained his traditional ties to the village commune and to the land that he worked. To prevent tax evasion through the formation of artificial households, Peter introduced a new unit of taxation, the "soul"—*i.e.*, a male peasant of working age—and the lords were made responsible for the collection of the tax assessed on each of their souls. The peasant thus became a mere item on the tax roll who could be moved, sold, or exchanged according to the needs and whims of his master—whether a private landlord, the church, or the state. The serf became practically indistinguishable from a slave.

As befitted a secular-minded autocrat who saw his main task as enlightening and leading his people to "modernity," Peter had little regard for the church. He recognized its value only as an instrument of control and as an agent of modern education. When the patriarch died in 1700, Peter appointed no successor. Finally in 1721 he gave the church a bureaucratic organization: a Holy Synod composed of several appointed hierarchs and a lay representative of the emperor; the latter, called the chief procurator, came to play the dominant role. Ecclesiastic schools turned into closed institutions with a narrowly scholastic curriculum. Membership in the clerical estate became strictly hereditary; the priesthood was transformed into a closed caste of government religious servants cut off from the new secular culture being introduced in Russia and deprived of their traditional moral authority. Both on economic and religious grounds, therefore, the reign of Peter I appeared particularly oppressive to the common people. It seemed unnatural and contrary to tradition; for many it clearly was the reign of the Antichrist, from which one escaped only through self-immolation (practiced by some of the Old Believers), open rebellion, or flight to the borderlands of the empire.

Reform of the church

Resistance and flight were made possible by Peter's failure, despite all his modernizing and rationalizing, to endow the government with effective means of control on the local level. Regular officials were short in number and experience and could not be easily spared for local administration. Peter tried to have the officers of the regiments that were garrisoned in the provinces double as local officials, but the experiment failed because of the necessities of war and because regular officers proved incompetent to administer peasants. The attempts at copying Western models were also unsuccessful, for the Russian nobility lacked (and was not allowed to develop) a local corporate organization that could serve as the foundation for local self-government.

Peter concentrated his attention almost entirely on the central administration, for which his reforms provided the basic framework within which the imperial government was to operate until its fall in 1917. To prosecute the war the Petrine state had to mobilize all the resources of the country and to supervise practically every aspect of national life. This required that the central executive apparatus be extended and organized along functional lines. Peter hoped to accomplish this by replacing the numerous haphazard *prikazy* (administrative departments) with a coherent system of functional and well-ordered colleges (their number fluctuating around 12 in the course of the century). Each college was headed by a board for more effective control; it had authority in a specific area such as foreign affairs, the army, the navy, commerce, mining, finances, justice, and so on. The major problems with this form of organization proved to be the coordination, planning, and supervision of the colleges.

Peter tried to cope with these defects pragmatically through the creation of a Senate, which came to serve as a privy council as well as an institution of supervision and control. In addition, he set up a network of agents (*fiskaly*) who acted as tax inspectors, investigators, and personal representatives of the emperor.

Much reliance was put on the obligation to denounce all would-be violators of imperial orders. Those failing to do so suffered the same punishment as the actual violator, while the informer was rewarded with the property confiscated from the "criminal." Internal security was vested in 1689 in the chancery of the Preobrazhensky Guards, the tsar's own regiment, which became a much-dreaded organ of political police and repression. Under different names the police apparatus remained a permanent feature of the imperial regime. The police were also the instrument of the ruler's personal intervention, an essential function for the preservation of the autocracy as a viable political system.

The needs of war, as well as the desire to modernize Russia, led Peter to promote and expand industry, particularly mining, naval construction, foundries, and the production of glass and textiles. The emperor aimed at maximizing the use of all potential resources of the country to heighten its power and further its people's welfare; these goals were pursued in mercantilist fashion through discriminatory tariffs, state subsidies, and regulation of manufactures. Peter hoped to involve the rich merchants and the nobility in economic enterprise and expansion. As a class, however, the merchants failed to follow his lead; many were Old Believers who refused to work for the Antichrist. Nor did Peter's urban legislation provide the townspeople with the incentives and freedom necessary to change them into an entrepreneurial class; as a matter of fact, the municipal reforms were simply means to collect taxes and dues in kind. As to the nobility, only a few had the necessary capital to become entrepreneurs, and their time and energies were completely taken up by their service obligations. Nor did Peter provide for the security of property and for the landowner's right to dispose of the mineral, water, and timber resources on his estate. The shortage of capital could be, and in some specific cases was, overcome by direct government grants. But the equally serious shortage of labour was not so easily resolved. Peter permitted the use of servile labour in mines and manufactures, with the result that thousands of peasants were moved and forced to work under unfamiliar conditions, in new places, at very difficult tasks. Resentment ran high and the productivity of this forced labour was very low. Most of the enterprises established in Peter's lifetime did not survive him. But the impetus he had given to Russian industrial development was not altogether lost; it revived with new vigour—under different policies—in the middle of the 18th century.

Petrine economy

Among the important factors in Russia's economic development under Peter was the building of St. Petersburg on the then inhospitable shores of the Gulf of Finland. Its construction cost an estimated 30,000 lives (lost from disease, undernourishment, and drowning) and engulfed vast sums of public and private money. Nobles who served in the central administration and at court were required to settle in the new city and to build townhouses.

The location of the new capital symbolized the shift in the empire's political, economic, and cultural centre of gravity toward western Europe. Trade and social intercourse with western Europe became easier, and the icebound peripheral ports of what is now Murmansk and of Arkhangelsk were abandoned for the more convenient harbours of Riga, Revel, and the new St. Petersburg. After 1721 Peter also extended the borders of the empire in the south along the Caspian Sea as a result of a successful war against Persia (Treaty of St. Petersburg, 1723).

The changes that made Peter's reign the most seminal in Russian history were not the administrative reforms and the military conquests, significant as those were, but the transformation in the country's culture and style of life, at least among the service nobility. Foreign observers made much of Peter's requirement that the nobility shave

Change in Russia's culture and life-style

off their beards, wear Western clothes, go to dances and parties, and learn to drink coffee. These were only the external marks of more profound changes that in a generation or so were to make the educated Russian nobility members of European polite society. The common people, especially the peasantry, were not so immediately and positively affected, although by the end of the 18th century most peasants, and all inhabitants of towns, had moved a considerable distance from the values and habits of their 16th- and 17th-century forebears.

Most important of all, perhaps, the reign of Peter I marked the beginning of a new period in Russian educational and cultural life. Peter was the first to introduce secular education on a significant scale and to make it compulsory for all state servants. (More significant than the limited quantitative results during Peter's lifetime was the fact that education eventually became indispensable to membership in the upper class.) First, Peter tried to use the church to establish a network of primary schools for all children of the free classes; this plan failed largely because the clergy were unable to finance and staff schools for secular learning. But the specialized technical schools Peter founded, such as the Naval Academy, struck roots and provided generations of young men with the skills necessary for leadership in a modern army and navy. Although he did not live to see its formal inauguration, Peter also organized the Academy of Sciences as an institution for scholarship, research, and instruction at the higher level. The academy's beginnings were quite modest—German professors lectured in Latin to a handful of poorly prepared students—and its development was not free from difficulties, but at the end of the 18th century it was a leading European centre of science and enlightenment, preparing and guiding Russia's scientific and technological flowering in the 19th century.

Russia as a major European power

Assessment of Peter's reign. Contemporaries as well as later historians have given first place among Peter's accomplishments to his conquest of the Baltic provinces and areas on the Caspian Sea. More important was the fact that during his reign Russia became a major European power, in regular intercourse with the major trading powers and especially with Holland and Great Britain. This status of European power, however, burdened Russia with the maintenance of a large and up-to-date military establishment that became involved in many costly conflicts. The new institutional forms that Peter introduced helped to shape a less personal and more modern (*i.e.,* routinized and bureaucratized) political authority. This led to an ambiguous relationship between the autocratic ruler and his noble servants and also to a sense of alienation between the common people and the ruler.

Contemporaries and later generations alike shared the feeling that Peter's reign had been revolutionary—a radical and violent break with the centuries-old traditions of Muscovy. To some extent this was the consequence of Peter's ruthless manner, his dynamism, his harsh suppression of all opposition, and his obstinate imposition of his will. From a historical perspective, Peter's reign may appear to have been only the culmination of 17th-century trends rather than a radical break with the past. But people are more conscious of changes in manners and customs than of deeper transformations that require a long time for their working out. Thus Russia's cultural Europeanization in the early 18th century produced works of literature in a new manner, using foreign styles and techniques, such as the treatises and sermons of Feofan Prokopovich, Peter's main assistant in church matters, and the satires and translations of Prince Antiokh Kantemir, the first modern Russian poet. These writers, and many lesser ones, praised Peter's work, stressing its innovative and necessary character. The educated elite, reared on the cultural elements introduced by Peter, perceived his reign as the birth of modern Russia. This in itself became the source of critical thought and raised the question of whether the break with the past was desirable or a betrayal of the genuinely national patterns of development of Russian culture. It appeared that forcible imposition of foreign elements had led to an alienation between the elite and the Russian people. This debate as to the nature and value of the reign of Peter I served as the main stimulus to a definition of Russian national culture and to the elaboration of competing political and social philosophies in the 19th century (*e.g.,* those of the Slavophiles and the Westernizers). Peter's reign has been at the centre of all debates over Russian history, since any attempt to define its periods and to assess Russia's development in modern times requires a prior judgment of the reign and work of Peter I. (For a more detailed biography, see the *Macropædia* article PETER I THE GREAT, OF RUSSIA.)

PETER I'S SUCCESSORS (1725–62)

Peter's unexpected death in 1725 at the age of 52 left unresolved two major institutional problems. The first was the succession to the throne, which remained unsettled not only because Peter did not choose his own successor but also because during the remainder of the century almost any powerful individual or group could disregard the choice of the preceding ruler. The second problem was the lack of firm central direction, planning, and control of imperial policy; closely related to it was the question of who would have the determining role in shaping policy (*i.e.,* what would be the nature of the "ruling circle" and its relationship to the autocrat). The failure to solve these problems produced a climate of instability and led to a succession of crises in St. Petersburg and Moscow that make it difficult to give unity to the period from 1725 until the accession of Catherine the Great in 1762.

Instability and crisis

Normal and peaceful succession to the throne was thwarted by a combination of biological accidents and palace coups. At Peter's death his chief collaborators, who were headed by Prince A.D. Menshikov and were assisted by the guard regiments (the offshoots of the play regiments of Peter's youth), put on the throne Peter's widow—his second wife, Catherine I, the daughter of a Lithuanian peasant. Quite naturally, Menshikov ruled in her name. Soon, however, he was forced to share his power with other dignitaries of Peter's reign. A Supreme Privy Council was established as the central governing body, displacing the Senate in political influence and administrative significance. Catherine I's death in 1727 reopened the question of succession; Peter's grandson (the son of Alexis, who had perished in prison) was proclaimed Emperor Peter II by the council. An immature youngster, Peter II fell under the influence of his chamberlain, Prince Ivan Alekseyevich Dolgoruky, whose family obtained a dominant position in the Supreme Privy Council and brought about the disgrace and exile of Menshikov. It looked as if the Dolgorukys would rule in fact because Peter II was affianced to the chamberlain's sister, but Peter's sudden death in January 1730 crossed the plans of that ambitious family.

Anna (1730–40). Under the leadership of Prince Dmitry Golitsyn—scion of an old Muscovite boyar family and himself a prominent official under Peter I—the Supreme Privy Council elected to the throne Anna, dowager duchess of Courland and niece of Peter I (daughter of his co-tsar, Ivan V). At the same time, Golitsyn tried to circumscribe Anna's power by having her accept a set of conditions that left to the council the decisive voice in all important matters. This move toward oligarchy was foiled by top-level officials (the *generalitet—i.e.,* those with the service rank of general or its equivalent), in alliance with the rank-and-file service nobility. While the former wanted to be included in the ruling oligarchy (and Golitsyn seemed to have been ready to concede them this right), the latter opposed any limitation on the autocratic power of the sovereign. Indeed, the ordinary service nobles feared that an oligarchy, however broad its membership, would shut them off from access to the ruler and thus limit their opportunity to rise in the hierarchy of the Table of Ranks.

Anna left most of her authority to be exercised by her Baltic German favourite, Ernst Johann Biron, who acquired a reputation for corruption, cruelty, tyranny, and exploitation and who was felt to have set up a police terror that benefited the Germans in Russia at the expense of all loyal and patriotic Russians. Recent scholarship has modified this image and shown that Biron's bad reputation rested on his inflexibility in applying the law and collecting taxes, rather than on malevolence. The Supreme Privy

Role of Ernst Biron

Council was abolished upon Anna's accession in 1730, and the functions of coordination, supervision, and policy planning were vested in a cabinet of ministers composed of three experienced high officials, all Russians.

Elizabeth (1741–62). Anna was childless; she appointed as successor her infant nephew, Ivan Antonovich, under the regency of his mother, Anna Leopoldovna. Biron, who had at first retained his influence, was overthrown by Burkhard Christoph, Count von Münnich, who had made his fortune in Russia. The continuing domination of a few favourites—many of whom were Germans—much displeased the high officials, whose position was threatened by the personal caprices of ruler or favourite, and incensed even more the rank and file of the service nobility, who could not obtain rewards or favours from the sovereign without the approval and help of the favourites. The malcontents banded together around Peter I's daughter, Elizabeth, whose easygoing and open ways had gained her many friends; she was also popular because of her Russian outlook, which she emphasized, and because she shared the aura of her great father. With the help of the guard regiments and high officers and with the financial support of foreign diplomats (in particular the French envoy), Elizabeth overthrew the infant Ivan VI and the regent Anna Leopoldovna in 1741. Her 20-year reign saw the rise of certain trends and patterns in public life, society, and culture that were to reach their culmination under Catherine the Great. On the political plane, the most significant development was the restoration of the Senate to its earlier function of chief policy-making and supervising body. At the end of her reign Elizabeth also established a kind of permanent council or cabinet for planning and coordination—the Special Conference at the Imperial Court.

Stabilization under Elizabeth

During this period Peter's administrative reforms began to bear fruit. The Table of Ranks became the framework for a class of servicemen whose lives were devoted to the interests of the state. In principle, entry to this class of officials was open to anyone with the required ability and education, including the sons of priests and non-Russian landowners. In fact, however, promotion in the Table of Ranks was possible only if the individual's merit and performance were recognized by the ruler or, more likely, by high officials and dignitaries who had access to the ruler. The personal element, bolstered by family and marriage ties, came to play an important role in the formal system of promotion; most significantly, it determined the makeup of the very top echelon of the administrative and military hierarchies (which were interchangeable). This group constituted an almost permanent ruling elite, co-opting its own membership and promoting the interests of the families most directly connected with it; in order to solidify its influence and function, it aimed at bringing as many routine government operations as possible under a system of regulations that would make appeal to the ruler unnecessary. The ruler's autocratic power could not be infringed, however, because his authority was needed not only to settle special cases but also to promote, protect, and reward members of the ruling group and their clients. The greatest threat to the system was the interference or interposition of favourites—"accidental people"—and, to guard against this, the oligarchy entered into an alliance with the rank-and-file service nobles who wanted to join its ranks and could hope to do so with the help of the dignitaries' patronage. This alliance permitted successful palace coups against favourites. The system worked well enough to allow the consolidation of Peter's reforms, some success in foreign policy, and a general increase in the power and wealth of the state, despite the low calibre of the rulers and the mismanagement of favourites.

The service nobleman

The system rested on the availability to all nobles of the minimum education necessary for entrance and promotion in service. As a consequence, cultural policy became a major concern of the government and the nobility alike; the members of the service class demanded that institutions of learning be set up to prepare the nobility for better careers, permitting them to skip the lowest ranks. That demand was fulfilled in 1731 with the creation of the Corps of Cadets. In the course of the following decades, the original corps was expanded, and other special institutions for training the nobility were added. General education became accessible to a large stratum of the rank-and-file nobility with the founding of the University of Moscow in 1755, although the lack of automatic preferment for its graduates kept it from being popular among the wealthier nobles until the end of the century. The Corps of Cadets and similar public and private institutions also acted as substitutes for local and family bonds; these schools were also the seedbeds for an active intellectual life, and their students played a leading role in spreading the literature and ideas of western Europe in court circles and in the high society of the capitals.

The service noblemen were also landlords and serf owners. The majority of them, however, were quite poor for a number of reasons, chief among which were the low productivity of Russian agriculture, absentee management, and the scattered and splintered character of the landholdings. The average small or middling estate yielded only the bare necessities for the survival of the serviceman's family. As long as he remained in service, away from the estate, and without capital, he could do little to improve his property, especially since any change in the agrarian routine would have to be accepted by his peasant-serfs and the noble neighbours among whose lands his own lay scattered in an inextricable patchwork. He thus depended on the ruler for additional income, either in the form of a salary or as grants of land (and serfs) in reward for service. The salary was not very large, it was often in kind (furs), and it was paid out rather irregularly; lands and serfs could be obtained only from the ruler, and most went to favourites, courtiers, or high dignitaries. Service, it is true, provided the nobleman with some extras, such as uniforms, sometimes lodgings, and—most important—greater accessibility to court, cultural life, and education for his children. Thus, he remained in service and took little direct interest in his estates and serfs.

Shuvalov's policies

Elizabeth's chief adviser, Pyotr Shuvalov, had the government grant exclusive privileges and monopolies to some of the nobility, hoping to involve them in the development of mining and manufacturing. Shuvalov also initiated a gradual loosening of state controls over economic life in general. He began to dismantle the system of internal tariffs, so that local trade could develop; he strengthened the landlord's control over all the resources on his estate; and he gave the nobles the right to distill alcohol.

At the same time, the landlords were obtaining still greater power over their serfs. The full weight of these powers fell on the household serfs, whose number increased because their masters used them as domestics and craftsmen in their town houses to make the Western-style objects with which they surrounded themselves. When noblemen established factories or secured estates in newly conquered border areas, they transferred their serfs to them without regard for family or village ties. The operation of most estates was, in the absence of the landlord, left to the peasants. This only perpetuated the traditional patterns of agriculture and made the modernization and improvement of agricultural productivity impossible.

Elizabeth's reign also witnessed Russian victories over Turkey that expanded and consolidated the empire's control in southwestern Ukraine, between the Bug (Buh) and Dniester rivers, and promoted settlement in Ukraine. Moreover, Russia was interfering more and more in the domestic politics of Poland and in the diplomatic game of central and western Europe. Elizabeth joined Austria, France, Sweden, and Saxony in a coalition against Prussia, under Frederick II, Great Britain, and Hanover; this led to Russia's involvement in the Seven Years' War. Russian armies were successful in conquering East Prussia and occupied Berlin briefly. The empress's death saved the king of Prussia from total disaster.

THE REIGN OF CATHERINE II THE GREAT (1762–96)

Elizabeth, too, was childless, and the throne passed to the heir she had selected—her nephew the Duke von Holstein Gottorp, who became Peter III. Peter III made himself personally unpopular with St. Petersburg society; in addition, he allowed his entourage (mainly his Holstein relatives



The expansion of Russia, 1300–1796.
From W. Shepherd, *Historical Atlas;* Barnes & Noble Books, New York

and German officers) to take control of the government. The regular hierarchy of officials—particularly the Senate—was pushed into the background; power passed into the hands of the emperor's favourites, while a modernized police, under the personal control of a general who was one of the emperor's minions, spread its net over the empire. The pro-Prussian foreign and military policy pursued by Peter III (who abruptly ended Russia's victorious involvement in the Seven Years' War) and his treatment of his wife Catherine provoked much resentment. As a result, the emperor lost all support in society. It was easy for Catherine, with the help of the senators, high officials, and officers of the guard regiments (led by her lover Grigory Orlov and his brothers), to overthrow Peter III on June 28 (July 9, New Style), 1762. Thus began the long and important reign of Catherine II, whom her admiring contemporaries named "the Great."

The daughter of a poor German princeling, Catherine had come to Russia at the age of 15 to be the bride of the heir presumptive, Peter. She matured in an atmosphere of intrigue and struggle for power. She developed her mind by reading contemporary literature, especially the works of the French Encyclopaedists and of German jurists and cameralists. When she seized power at the age of 33, she was intellectually and experientially prepared, as the more than 30 years of her reign were to show.

Assessment of Catherine's reign

The historiography of Catherine's reign has been dominated by two approaches: a dramatization and romanticization of her personal life, which was indeed colourful for the number and variety of her lovers; and the viewpoint of 19th-century liberalism, which took literally her self-description as a "philosophe on the throne." Marxist and Soviet historians, to the extent that they have dealt with her reign at all, see it primarily in terms of the pressures put on the state by the serf-owning nobility faced with the demands of an expanding market economy. In recent years, scholars have seen Catherine's government as working to further the formation of a modern civil society in which social classes and groups pursue their own interests rather than serving the needs of the state exclusively.

Even before she seized power, Catherine wrote that the task of good government was to promote the general welfare of the nation by providing for the security of person and property; to that end, government should operate in a legal and orderly fashion, furthering the interests of individual subjects and giving groups and classes as much autonomy in the pursuit of their normal activities as possible. All the same, Catherine believed that the autocratic state had important functions; she had no intention of relinquishing or limiting her authority, even though she was willing to withdraw from those areas of national life that could be safely administered by an educated elite.

Expansion of the empire. Catherine's reign was notable for imperial expansion. First in importance for the empire were the securing of the northern shore of the Black Sea (Treaty of Küçük Kaynarca, 1774), the annexation of the

Crimea (1783), and the expansion into the steppes beyond the Urals and along the Caspian Sea. This permitted the adequate protection of Russian agricultural settlements in the south and southeast and the establishment of trade routes through the Black Sea and up the Danube. On the other hand, these gains involved Russia more and more in the political and military struggle over the crumbling Ottoman Empire in the Balkans.

Grigory Aleksandrovich Potemkin, Catherine's favourite in the 1770s, may be considered the chief architect of her imperial policy. He promoted large-scale foreign colonization and peasant resettlement in the south—with only mediocre success so far as agricultural settlements went but with great success in the foundation and rapid growth of such towns and ports as Odessa, Kherson, Nikolayev, Taganrog, and Mariupol (Pavlovsk). Within a generation or two these became lively cultural centres and major commercial cities for all of southern Russia, contributing to the reorientation of Russia's pattern of trade with the development of agricultural exports from Ukraine. Local society was transformed on the Russian pattern: the landlords became imperial service nobles with full control over their peasants; vast new lands were parceled out to prominent officials and made available for purchase by wealthy Russian nobles, who also received the right to resettle their own serfs from the central regions. Thus serfdom, along with elements of the plantation system, was extended to still more people and over whole new provinces. If this expansion benefited the state and a small and already wealthy part of the Russian nobility, it increased the misery and exploitation of the Ukrainian and Russian peasantries. The traditional military democracies of the Cossack hosts on the Dnieper, Don, Ural, Kuban, and Volga rivers lost their autonomy and special privileges; the wealthier officers became Russian service nobles, receiving the right to own and settle serfs on their own lands, while the rank-and-file Cossacks sank to the level of state peasants with special military obligations.

Policies toward new nationalities

Integration of the new territories required the absorption of a large number of non-Russian, non-Christian nomadic peoples. The approach that prevailed until the late 19th century was based on the idea, taken from Enlightenment writings, that there is a natural progress of society from primitive hunting and fishing groups through the stage of nomadism to settled agriculture, trade, and urbanization. Accordingly, the government sought to bring the nomadic peoples up to what it considered to be the Russian peasantry's higher way of life; this policy had the advantage also of producing uniformity in administrative and legal structures. Catherine's government was quite willing to let religious, cultural, or linguistic differences stand, although it did not feel committed to protect them actively. Inevitably, however, its effort to change the ways of the nomads affected their culture and religion and, through these, their social equilibrium and sense of national identity. While Catherine's policy led some peoples to accept (more or less under duress) changes in their way of life, thus facilitating the extension of Russian agricultural settlements onto the open steppes, it also gave rise to a growing sense of identity based on cultural, linguistic, and religious traditions. These nationalistic sentiments clashed with the outlook and practices of officials accustomed to thinking in universal categories. The policy thus defeated its own aims: it handicapped the economic development of the empire's border regions (*e.g.*, in Siberia) and worked against the social and cultural integration of the natives into the fold of the dominant Russian culture (although Russification did take place on a significant scale in the case of some native elites, as in the Caucasus and the Crimea).

In the course of the Russo-Turkish war of 1768–74, considerations of balance of power led Frederick II of Prussia to suggest that Russia, Austria, and Prussia find territorial compensation at the expense of Poland rather than squabble over the spoils of the Ottoman Empire. The internal situation of the Polish Commonwealth—in particular the treatment of non-Catholics, which allegedly was grossly discriminatory—had led the three neighbours to meddle in Poland's domestic affairs. After much diplomatic and political maneuvering, Russia, Prussia, and Austria compelled Poland to cede large chunks of its territory in the so-called First Partition (1772–73), the major beneficiaries of which were Russia (which obtained the Belorussian lands) and Austria (Prussia obtained less actual territory, but what it acquired was of great economic value). Polish patriots attempted to bring political stability to their country by drafting the "Constitution of 3 May 1791," which provided for stronger royal authority, established four-year sessions of the elected Sejm (the Polish diet), abolished the liberum veto in its proceedings (under the liberum veto, any single member of the Sejm could kill a measure), and introduced significant liberal reforms in education and law. The prospect of social and political progress within the framework of a stable government did not suit the partitioning powers, so that a Second Partition was forced on the Poles in 1792. The revolt led by Tadeusz Kościuszko to save Poland's independence was crushed, and in 1795 the three neighbours seized the remainder of the country and ended its political sovereignty and national independence.

In the short term the partitions seemed a significant success for the Russian Empire, completing the "gathering of 'Russian' lands" (begun in the 15th century) with the acquisition of Belorussia and Volyn, but in the long run they proved more of a liability than an asset. Russia became politically tied to Prussia and had to shoulder an increased military burden to defend its new boundaries as well as to maintain law and order among a people restive under foreign occupation. It also proved difficult to co-opt the Polish elites into the imperial establishment, as had been the case with the Ukrainians, the Baltic Germans, and non-Slavic natives. In addition, the empire acquired for the first time a large Jewish population, which created numerous unforeseen problems. It may also be argued that controlling the obstreperous nation resulted in a regime of harsh police supervision and oppressive censorship throughout the empire.

Reforms of local government

Government administration under Catherine. The reforms of local government carried out by Catherine also contained contradictions. The successors of Peter I had not solved the problem of local administration. St. Petersburg relied on appointed officials, too few in number and much given to abuse and corruption, and on the informal control exercised by individual landowners and village communes. However, a great peasant rebellion led by Yemelyan Ivanovich Pugachov in 1773–74 demonstrated the inadequacy of this system. Taking up suggestions of various officials and mindful of the information and complaints offered by the deputies to the Legislative Commission (1767–68), Catherine shaped the local administration into a structure that remained in force until the middle of the 19th century and also served as a foundation for the zemstvos (local elected councils), established in 1864. The basic pattern was established by the statute on the provinces of 1775 and complemented by the organization of corporate self-administration contained in the Charters to the Nobility and the Towns (1785). Essentially, the reforms divided the empire's territory into provinces of roughly equal population; the division paid heed to military considerations. Each of these units (*guberniya*) was put under the supervision and responsibility of a governor or governor-general acting in the name of the ruler, with the right of direct communication with him. A governor's chancery was set up along functional lines (paralleling the system of colleges) and subordinated to and supervised by the Senate. The regular provincial administration was assisted by officials who were elected from among the nobility for the countryside and from the higher ranks of townspeople for the cities; these elected officials took care of routine police matters in their jurisdictions, helped to enforce orders received from the central authorities, and assisted in the maintenance of law and the collection of taxes. Other elected personalities (marshals of the nobility and heads of city councils) protected the interests of their respective classes and helped to settle minor conflicts without recourse to regular tribunals. This delegation of some administrative functions to the local level multiplied the number of state agents on the local level but also fostered

a sense of responsibility among the active and cultured members of the local upper classes. On the other hand, the serfs and the lower classes in the towns found themselves without anyone to protect their interests.

The central government under Catherine

Catherine made no fundamental changes in the administration of the central government. The system of colleges was retained, but the authority of the presidents increased at the expense of the boards, initiating an evolution that culminated in the establishment of monocratic ministries in 1802. The Senate supervised all branches of administration, regulating the orderly flow of business. The Senate was also involved—albeit indirectly—in coordination, mainly because its procurator general, Prince Aleksandr A. Vyazemsky, held the office for a quarter of a century with the full trust of the empress. At the same time, the judicial functions of the Senate as a high court of appeal and administrative review were widened.

The major institutional weakness of the Petrine system remained—namely, the lack of a body to coordinate the jurisdictions and resolve the conflicts of the colleges and to plan policies and control their implementation. A ruler as energetic, hardworking, and intelligent as Catherine could perform these tasks almost single-handedly, as had Peter I; but with the growing complexity of administration even Catherine felt the need for such a body, if only to reduce her involvement in every small detail or contested matter and to provide a wider scope for government by permanent laws and uniform regulations.

A major need of the empire was an up-to-date code of laws. The last code, issued in 1649, had become largely inoperative as a result of Peter's reforms and the transformation of society. Peter and his successors had recognized this need by appointing commissions to prepare a new code; none of the several efforts having reached a successful conclusion, Catherine tried to tackle the job again, but in a different manner. In 1767 she convoked a commission of representatives elected by all classes except private serfs. For their guidance she drafted an instruction largely inspired by Western political thinkers; but, far from providing a blueprint for a liberal code, it emphasized the need for autocracy. In its civil part the instruction owed much to German political philosophy and natural-law jurisprudence, putting the individual's duties before his rights, emphasizing the state's responsibility for the welfare of the nation, and encouraging the pursuit of material self-interest within the established order. Although not implemented by the commission (which was adjourned indefinitely in 1768), the instruction stimulated the modernization of Russian political and legal thought in the early 19th century.

Catherine's social policy

In her social policy Catherine aimed at steering the nobility toward cultural interests and economic activity so as to reduce their dependence on state service. (They had already been freed from compulsory service by Peter III in 1762.) To this end she ordered a general land survey that fixed clearly and permanently the boundaries of individual estates, and she granted the nobility the exclusive right to exploit both the subsoil and surface resources of their land and to market the products of their estates and of their serfs' labour. The nobles also obtained a monopoly of ownership of inhabited estates, which, in fact, restricted ownership of agricultural serfs to the noble class. Catherine hoped to stimulate agricultural expansion and modernization by providing easy credit and by disseminating the latest techniques and achievements of Western agriculture through the Free Economic Society, founded in 1765. She also fostered the nobility's corporate organization. The Charter to the Nobility, 1785, gave the corps of nobility in every province the status of a legal entity. The corporation's members gathered periodically in the provincial and district capitals to elect a marshal, who represented their interests before the governor and the ruler himself; they also elected a number of officials to administer welfare institutions for the nobility (schools, orphanages, and so on), to help settle disputes, and to provide guardianships for orphans. The corporate life of the nobility did not develop as well as expected, however, and the nobility never became the class it was in Prussia or England, but the charter did foster a sense of class consciousness and afforded legal security to the members and their property. The periodic electoral meetings stimulated social intercourse, led to a livelier cultural life in the provinces, and helped to involve the nobility in local concerns. The charter provided both a framework and the stage for the gradual formation of a "civil society" whose members cultivated interests, activities, and values independent of the state's—a trend that would come to full bloom and manifest itself in the first half of the 19th century.

Turning the nobility's interests toward economic activity brought the return home of many landowners to supervise the operation of their estates. Interested in obtaining greater income, they not only intensified the exploitation of serf labour but also interfered in the traditional routine of the village by attempting to introduce new agricultural techniques. In most cases, this meant increased regimentation of the serfs. The secularization of the lands (estates) of monasteries and episcopal sees in 1764 had brought a considerable amount of land into the possession of the state. To reward her favourites and to encourage the nobility to economic activity, Catherine gave away large tracts with many peasants, who now had to work for ambitious and capricious masters.

Serfdom had never been acceptable to the Russian peasant; it now became particularly burdensome and unjust, the more so since the lord's extensive police powers removed his serfs from the state's protection, and the new local officials enforced strictly the prohibition against appealing to the sovereign for relief. There were also the specific grievances of the Cossacks, whose traditional liberties had been sharply curtailed and their social organization undermined, as well as the discontent of the nomadic peoples forced to accept a new way of life. Peasant misery erupted in rebellion, led by the Cossack Yemelyan Pugachov, that engulfed all of eastern European Russia in 1773–74. The peasant forces captured a number of towns and cities before they were finally defeated by government armies. The revolt demonstrated the inadequacy of local controls and was thus partly responsible for the reform of provincial administration mentioned above. It also brought the educated elite to a new awareness of the profound alienation of the peasantry from the culture of St. Petersburg.

Town planning and building

The reign of Catherine II was a period of active town planning and building. The number and size of the urban centres grew slowly but steadily. Along with new cities in the south, many old towns were rebuilt and developed. The renaissance of the old provincial centres was in part due to the administrative reforms of 1775 and 1785, which brought an influx of officials and nobles. Along with them came craftsmen, artisans, and merchants. An act of Peter III that permitted peasants to trade in neighbouring towns without passports or controls at the gates gave impetus to the emergence of a class of small merchants from among the peasantry. This trend received support from the administrative reorganization of the towns and the limited degree of corporate self-administration granted by the Charter to the Towns of 1785.

EDUCATION AND SOCIAL CHANGE IN THE 18TH CENTURY

Secular education had been actively propagated by Peter I. At first it focused on technical subjects—those directly related to the prosecution of war, the building of a navy, and the running of the government. This was also the original emphasis of the Academy of Sciences and the school connected with it. But, as education became the prerequisite for advancement in service and as Western ways of life spread among the upper classes, the focus of education gradually broadened. There developed a class of nobles who were interested in culture for the sake of their own development, as well as for cutting a good figure in society. Beginning in the 1760s, the demand for western European artistic and cultural works grew increasingly in the salons of St. Petersburg. By the 1780s the major classics of European literature had become easily available in translation to any educated person. Private boarding and day schools proliferated, as did the tutors hired by wealthy nobles for their children (and for less fortunate neighbours and relatives). The Academy of Sciences took its place

among the major academies of Europe. The University of Moscow and the chief schools of the military, naval, and civil services had become regular institutions.

There were also ecclesiastical schools. The seminaries and theological academies not only trained future members of the episcopate and officials of the Holy Synod but also staffed government bureaus on the middle and higher levels and produced the first native Russian academics, scholars, and scientists. Russia's lack of professional experts in such fields as jurisprudence, civil and military engineering, astronomy, and geophysics brought a great influx of foreigners. They brought with them French and German philosophy: the metaphysics and epistemology of René Descartes and the natural-law doctrines of the German school of Gottfried Leibniz, Samuel von Pufendorf, and Christian Wolff. These emphasized social obligation and the individual's dependence on the community and thereby laid the foundation for a critique of society. The critique was at first directed against the moral inadequacies of individuals, but it soon broadened into the view that the educated man had an obligation to help his fellowmen improve themselves. In the Russian context the class most obviously in need of improvement was the peasantry. Moral progress, it was quickly realized, was not possible without material progress, and this led quite naturally to an advocacy of practical philanthropy and social action.

Western intellectual influences

Imported German professionals furthered the dissemination of German Pietism, with its emphasis on spiritual progress and on the need to serve man and the community. Similar tendencies underlay the most influential branch of Freemasonry; the Freemasons devoted themselves to disseminating knowledge, relieving hunger, and caring for orphans and other destitutes. The publisher Nikolay Novikov carried the Pietist and Masonic messages to the public in his satiric journals and periodicals for women and children. The major writers of Catherine II's reign (including the empress herself, who dabbled in journalism and drama) produced satires, fables, and comedies of manners attuned to the belief that moral and spiritual progress would lead to social improvements. A similar approach was noticeable in education, which stressed the development of moral feeling in the conviction that a good heart would guide the well-filled head in the proper direction.

All these intellectual currents combined to awaken among educated Russians a sense of national pride and a feeling that, thanks to the impetus given by Peter I, Russia had managed to lift itself to the cultural and political level of a great European state. The educated Russian was no longer a servile and mute slave of the tsars; he had made himself into a gentleman, a man of heart and honour, a "true son of the fatherland," concerned about his fellowmen and his country's condition and future.

Growth of national pride

The response of the empress and her entourage to these intellectual developments was ambivalent. The new sense of national pride and personal dignity enhanced the government's prestige and was in line with Catherine's own aspirations for the nobility. But moral criticism of abuses could easily turn into criticism of Russia's social and political system. The outbreak of the French Revolution made Catherine II particularly anxious. She felt that large-scale private philanthropic and educational activities without government guidance and control were trespassing on her own prerogatives as an enlightened autocrat. By the end of the 18th century, the ideal of service to the state, which had underlain the Russian nobility's value system, had been transformed into one of service to the people; this meant the elite's separation from the state, which Catherine II could not accept. A dramatic illustration of Catherine's concern occurred on the appearance in 1790 of Aleksandr Radishchev's *Journey from St. Petersburg to Moscow.* In it Radishchev depicted social conditions as he saw them, particularly the dehumanization of the serfs and the corruption of their masters, warning that these threatened the stability of the existing order. Incensed by the book, Catherine had Radishchev arrested and banished to Siberia. He became the first political martyr of the Russian elite; his book and his fate foreshadowed the antagonism between the intelligentsia and the government that was to dominate Russia's history in the 19th century.

THE REIGN OF PAUL I (1796–1801)

Catherine died in 1796 and was succeeded by her son Paul. A capricious, somewhat unstable individual, Paul had a passion for military order that conflicted with the basic values of the developing civil society; he felt that the nobility should again become a service class (or withdraw completely into agriculture) and help the ruler in implementing his reform program, even at the expense of its private interests. In trying to reestablish compulsory state service, he made it more rigid, harsh, and militaristic. He sought to promote the welfare of the serfs, but the manner of his approach—a decree permitting a maximum of three days of labour service per week—was clumsy and high-handed; it did nothing to help the serfs and angered their lords. Paul also wanted to govern with his own minions, disregarding both tradition and the administrative patterns that had developed during his mother's 30-year reign. Paul's hatred of the French Revolution and of everything connected with it led him to impose tight censorship on travel abroad and to prohibit foreign books, fashions, music, and so forth. He thereby earned the enmity of upper society in St. Petersburg. On March 11 (March 23, New Style), 1801, he was murdered by conspirators drawn from high officials, favourites of Catherine, his own military entourage, and officers of the guard regiments. The accession of his son, Alexander I, inaugurated a new century and a new period in the history of imperial Russia. (M.Ra.)

Russia from 1801 to 1917

THE REIGNS OF ALEXANDER I AND NICHOLAS I

General survey. When Alexander I came to the throne in March 1801, Russia was in a state of hostility with most of Europe, though its armies were not actually fighting; its only ally was its traditional enemy, Turkey. The new emperor quickly made peace with both France and Britain and restored normal relations with Austria. His hope that he would then be able to concentrate on internal reform was frustrated by the reopening of war with Napoleon in 1805. Defeated at Austerlitz, the Russian armies fought Napoleon in Poland in 1806 and 1807, with Prussia as an ineffective ally. After the Treaty of Tilsit (1807), there were five years of peace, ended by Napoleon's invasion of Russia in 1812. From the westward advance of its arms in the next two years of heavy fighting, Russia emerged as Europe's greatest land power and the first among the continental victors over Napoleon. The immense prestige achieved in these campaigns was maintained until mid-century. During this period, Russian armies fought only against weaker enemies: Persia in 1826, Turkey in 1828–29, Poland in 1830–31, and the mountaineers of the Caucasus during the 1830s and '40s. When Europe was convulsed by revolution in 1848, Russia and Great Britain alone among the Great Powers were unaffected, and in the summer of 1849 the tsar sent troops to crush the Hungarians in Transylvania. Russia was not loved, but it was admired and feared. To the upper classes in central Europe, Nicholas I was the stern defender of monarchical legitimacy; to democrats all over the world, he was "the gendarme of Europe" and the chief enemy of liberty. But the Crimean War (1853–56) showed that this giant had feet of clay. The vast empire was unable to mobilize, equip, and transport enough troops to defeat the medium-size French and English forces under very mediocre command. Nicholas died in the bitter knowledge of general failure.

Russian victories against weak enemies

Alexander I as a young man had longed to reform his empire and benefit his subjects. His hopes were disappointed, partly by the sheer inertia, backwardness, and vastness of his domains, partly perhaps because of defects of his own character, but also because Napoleon's aggressive enterprises diverted Alexander's attention to diplomacy and defense. Russia's abundant manpower and scanty financial resources were both consumed in war. The early years of his reign saw two short periods of attempted reform. During the first, from 1801 to 1803, the tsar took counsel with four intimate friends, who formed his so-called Unofficial Committee, with the intention of drafting ambitious reforms. In the period from 1807 to 1812, he had as his chief adviser the liberal Mikhail Speransky. Both periods

produced some valuable administrative innovations, but neither initiated any basic reform. After 1815 Alexander was mainly concerned with grandiose plans for international peace; his motivation was not merely political but also religious—not to say mystical—for the years of war and national danger had aroused in him an interest in matters of faith to which, as a pupil of the 18th-century Enlightenment, he had previously been indifferent. While he was thus preoccupied with diplomacy and religion, Russia was ruled by conservatives and reactionaries, among whom the brutal but honest General Aleksey Arakcheyev was outstanding. Victory in war had strengthened those who upheld the established order, serfdom and all. The mood was one of intense national pride: Orthodox Russia had defeated Napoleon, the Corsican Antichrist, and therefore it was not only foolish but also impious to copy foreign models. Educated young Russians, who had served in the army and seen Europe, who read and spoke French and German and knew contemporary European literature, felt otherwise. Masonic lodges and secret societies flourished in the early 1820s. From their deliberations emerged a conspiracy to overthrow the government, inspired by a variety of ideas: some men looked to the United States for a model, others to Jacobin France. The conspirators, known as the Decembrists because they tried to act in December 1825 when the news of Alexander I's death became known and there was uncertainty about his successor, were defeated and arrested; five were executed, and many more sentenced to various terms of imprisonment in Siberia. Nicholas I, who succeeded after his elder brother Constantine had finally refused the throne, was deeply affected by these events and set himself against any major political change, though he did not reject the idea of administrative reform. After the revolutions of 1848 in Europe, his opposition to all change, his suspicion of even mildly liberal ideas, and his insistence on an obscurantist censorship reached their climax.

The sections that follow cover the development under Alexander I and Nicholas I of the machinery of government, of social classes and economic forces, of education and political ideas, of the relations between Russians and other peoples within the empire, and of Russian foreign policy.

Government. The discussions of Alexander I's Unofficial Committee were part of an ongoing debate that was to remain important until the end of the imperial regime. This may be called the debate between enlightened oligarchy and enlightened autocracy. The proponents of oligarchy looked back to a somewhat idealized model of the reign of Catherine II. They wished greater power to be placed in the hands of the aristocracy for the purpose of achieving a certain balance between the monarch and the social elite, believing that both together were capable of pursuing policies that would benefit the people as a whole. Their opponents, of whom the most talented was the young Count Pavel Stroganov, were against any limitation on the power of the tsar. Whereas the oligarchs wished to make the Senate an important centre of power and to have it elected by senior officials and country nobility, Stroganov maintained that if this were done the sovereign would have "his arms tied, so that he would no longer be able to carry out the plans which he had in favour of the nation." In any event, neither enlightened oligarchs nor enlightened absolutists had their way: Russia's government remained autocratic but reactionary. Alexander, however, never quite abandoned the idea of representative institutions. He encouraged Speransky to prepare in 1809 a draft constitution that included a pyramid of consultative elected bodies and a national assembly with some slight powers of legislation. In 1819 he asked Nikolay Novosiltsev, a former member of the Unofficial Committee who had made a brilliant career as a bureaucrat, to prepare another constitution, which turned out to be rather similar to the first, although somewhat more conservative and less centralist. Neither was ever implemented, though Alexander took some features of the first, notably the institution of the State Council, and used them out of their intended context.

The debate between oligarchy and autocracy

In 1802 Alexander instituted eight government departments, or ministries, of which five were essentially new. The organization of the departments was substantially improved in 1811 by Speransky. In the 1820s the ministry of the interior became responsible for public order, public health, stocks of food, and the development of industry and agriculture. Inadequate funds and personnel and the dominant position of the serf-owning nobility in the countryside greatly limited the effective power of this ministry. There was no question of a formal council of ministers, or of anything corresponding to a cabinet, and there was no prime minister. A committee of ministers coordinated to some extent the affairs of the different departments, but its importance depended on circumstances and on individuals. When the tsar was abroad, the committee was in charge of internal affairs. Aleksey Arakcheyev was for a time secretary of the committee, but he did not cease to be the strongest man in Russia under the tsar when he ceased to hold this formal office. The committee had a president, but this office did not confer any significant power or prestige.

Under Nicholas I the committee of ministers continued to operate, but the individual ministers were responsible only to the emperor. The centre of power to some extent shifted into the emperor's personal chancery, which was built up into a formidable apparatus. The Third Department of the chancery, under Count Aleksandr Benckendorff, was responsible for the security police. Its head was also chief of gendarmes, and the two offices were later formally united. The task of the security force was to obtain information on the state of political opinion and to track down and repress all political activity that might be considered dangerous to the regime. The Third Department was also considered by the tsar as an instrument of justice in a broad sense, the defender of all those unjustly treated by the powerful and rich. Some of the department's reports show that there were officials who took these duties seriously, but as a whole it showed more talent for wasting time and effort and for repressing opposition and stifling opinion than for redressing the grievances of the powerless. In addition, the department was often on the worst of terms with other branches of the public service.

Rule by bureaucracy

Russia under Alexander I and Nicholas I was ruled by its bureaucracy. The efforts of successive sovereigns after Peter the Great to establish a government service of the European type had had partial success. The Russian bureaucracy of 1850 combined some features of a central European bureaucracy of 1750 with some features of pre-Petrine Russia. One may speak of a "service ethos" and trace this back to 16th-century Muscovy. But the foundation of this ethos was, for the great majority of Russian officials, servile obedience to the tsar and not service to the state as that phrase was understood in a country such as Prussia. The notion of the state as something distinct from and superior to both ruler and ruled was incomprehensible to most government servants. Russian bureaucrats were obsessed with rank and status. Indeed, this was the only incentive that the government could give, as salaries were very meagre. Rank was not so much a reward for efficient service as a privilege to be grasped and jealously guarded. In order to prevent able persons, especially of humble origin, from rising too quickly, great emphasis was laid on seniority. There were exceptions, and outstandingly able, cultured, and humane men did reach the top under Nicholas I, but they were few.

The rank and file of the bureaucracy was mediocre, but its numbers steadily increased, perhaps trebling in the first half of the century. It remained poorly paid. The government's poverty was caused by the backward state of the economy, by the fact that no taxes could be asked of the nobility, and by the cost of waging wars—not only the great wars but also the long colonial campaigns in the Caucasus. Government officials were badly educated. They lacked not only precise knowledge but also the sort of basic ethical training that competent officials need. They were reluctant to make decisions: responsibility was pushed higher and higher up the hierarchy, until thousands of minor matters ended on the emperor's desk. Centralization of responsibility meant slowness of decision, and delays

of many years were not unusual; death often provided the answer. There were also many antiquated, discriminatory, and contradictory laws. Large categories of the population, such as Jews and members of heretical Christian sects, suffered from various legal disabilities. Since not all those discriminated against were poor and since many small officials were unable to support their families, bending or evasion of the law had its market price, and the needy official had a supplementary source of income. Corruption of this sort existed on a mass scale. To a certain extent it was a redeeming feature of the regime: if there had been less corruption the government would have been even slower, less efficient, and more oppressive.

Social classes. No significant changes were made in the condition of the serfs in the first half of the century.

Problems of the countryside

Alexander I, perhaps from fear of the nobility and with the memory of his father's fate in mind, approached the problem with caution, though with a desire for reform, but first war and then diplomacy diverted him. His successor, Nicholas, disliked serfdom, but there were political hazards in eliminating it. The power of the central government extended down to the provincial governors and, more tenuously, down to the *ispravnik,* or chief official of the district, of which each province had several. The *ispravnik* was elected by the local nobility. Below the level of the district, the administration virtually ceased to operate: the sole authority was the serf owner. If serfdom were to be abolished, some other authority would have to be put in its place, and the existing bureaucratic apparatus was plainly inadequate. The Decembrist conspiracy in 1825 had greatly increased the tsar's distrust of the nobility. He was determined to avoid public discussion of reform, even within the upper class.

The one important exception to the general picture of bureaucratic stagnation was the creation of the Ministry of State Domains, under General Pavel Kiselev. This became an embryonic ministry of agriculture, with authority over peasants who lived on state lands. These were a little less than half the rural population: in 1858 there were 19 million state peasants and 22.5 million private serfs. Kiselev set up a system of government administration down to the village level and provided for a measure of self-government under which the mayor of the *volost* (a district grouping several villages or peasant communes) was elected by male householders. There was also to be a *volost* court for judging disputes between peasants. Kiselev planned to improve medical services, build schools, establish warehouses for stocks of food in case of crop failure, and give instruction in methods of farming. Something was done in all these fields, even if less than intended and often in a manner that provoked hostility or even violent riots; the personnel of the new ministry was no more competent than the bureaucracy as a whole.

Opposition to serfdom

Only minor measures were taken to benefit the serfs on private estates. Opposition to serfdom grew steadily, however, not only among persons of European outlook and independent thought but also among high officials. It seemed not only unjust but intolerable that in a great nation men and women could be owned. Serfdom was also obviously an obstacle to economic development.

Whether serfdom was contrary to the interests of serf owners is a more complex question. Those who wished to abolish it argued that it was, since their best hope of getting the nobility to accept abolition lay in convincing them that their self-interest required it. Certainly in parts of southern Russia where the soil was fertile, labour was plentiful, and potential profits in the grain trade with Europe were high, a landowner would do better if he could replace his serfs with paid agricultural labour and be rid of obligations to those peasants whom he did not need to employ. In other regions, where the population was scanty, serfdom provided the landowner with an assured labour supply; if it were abolished, he would have to pay more for his labour force or see it melt away. In large parts of northern Russia where the land was poor, many serfs made a living from various crafts—in cottage industry or even in factories—and from their wages had to pay dues to their masters. The abolition of serfdom would deprive the serf owner of this large income and leave him with only what he could make from farming and from tenants with rather poor economic prospects. On balance, it seems likely that the short-term interests of the great majority of serf owners favoured the maintenance of serfdom, and, in any case, there is no doubt that this is what most serf owners believed.

Industry and trade made slow progress during these years. In the latter part of the 18th century, Russia had been, thanks to its Urals mines, one of the main producers of pig iron. In the next 50 years, it was left far behind by Great Britain, Germany, and the United States. In cotton textiles and sugar refining Russia was more successful. Count Yegor Frantsevich Kankrin, minister of finance from 1823 to 1844, tried to encourage Russian industry by high protective tariffs. He also set up schools and specialized institutes for the advancement of commerce, engineering, and forestry. Russia's exports of grain increased substantially, though its share of total world trade remained about the same in 1850 as in 1800. The first railways also appeared; rail traffic between St. Petersburg and Moscow was opened in 1851. The road system remained extremely inadequate, as was demonstrated in the Crimean War.

The urban population grew significantly. There were a few prosperous merchants, well protected by the government. Some centres, such as Ivanovo in central Russia, with its textile industry, had the beginnings of an industrial working class. The rest of the inhabitants of the cities consisted of small tradesmen and artisans, together with serfs living in town with their owners' permission as household servants or casual labourers.

Education and intellectual life. Alexander I's School Statute (1804) provided for a four-tier system of schools from the primary to the university level, intended to be open to persons of all classes. Under its provisions several new universities were founded, and gymnasiums (pre-university schools) were established in most provincial capitals. Less was done at the lower levels, for the usual reason of inadequate funds. In the latter part of Alexander's reign, education was supervised by Prince Aleksandr Nikolayevich Golitsyn, head of the ministry of education and spiritual affairs. In an effort to combat what he believed to be dangerous irreligious doctrines emanating from western Europe, Golitsyn encouraged university students to spy on their professors and on each other; those who taught unacceptable ideas were frequently dismissed or threatened with prison. Under Nicholas I there was some improvement. Count Sergey Uvarov, minister of education from 1833 to 1849, permitted a much freer intellectual atmosphere, but he also began the practice of deliberately excluding children of the lower classes from the gymnasiums and universities, which was continued under his successors.

Nevertheless, in increasing numbers the children of minor officials, small tradesmen, and especially priests were acquiring education. Together with the already Europeanized nobility, they began to form a new cultural elite. Direct political criticism was prevented by the censorship of books and periodicals. Petty police interference made life disagreeable even for writers who were not much concerned with politics. Aleksandr Pushkin, Russia's greatest poet, got into trouble with the police for his opinions in 1824; he was also a friend of some leading Decembrists. After 1826 he lived an unhappy life in St. Petersburg, tolerated but distrusted by the authorities, producing magnificent poetry until he met his death in a duel in 1837. The writers Mikhail Lermontov and Nikolay Gogol were also objects of suspicion to the bureaucrats.

The censorship was not always efficient, and some of the censors were liberal. It became possible to express political ideas in the form of philosophical arguments and literary criticism. Thus, it was partly in intellectual periodicals and partly in discussions in the private houses of Moscow noblemen that the controversy between "Westernizers" and "Slavophiles" developed. It began with the publication of a "philosophical letter" by Pyotr Chaadayev in the periodical *Teleskop* in 1836. One of the most brilliant essays ever written about Russia's historical heritage, it argued that Russia belonged neither to West nor to East, neither to Europe nor to Asia:

Westernizers and Slavophiles

> Standing alone in the world, we have given nothing to the world, we have learnt nothing from the world, we have not added a single idea to the mass of human ideas; we have made no contribution to the progress of the human spirit, and everything that has come to us from that spirit, we have disfigured. . . . Today we form a gap in the intellectual order.

Nicholas declared that Chaadayev must be mad and gave orders that he should be confined to his house and regularly visited by a doctor.

It is misleading to represent the Westernizers as wishing to slavishly copy all things Western or the Slavophiles as repudiating everything European and rejecting reform. The chief Slavophiles—Aleksey S. Khomyakov, the brothers Ivan and Pyotr Kireyevsky, the brothers Konstantin and Ivan Aksakov, and Yury Samarin—were men of deep European culture and, with one exception, bitter opponents of serfdom. Indeed, as landowners they knew more about the problems and sufferings of the serfs than did many Westernizers. The leading Westernizers—Aleksandr Herzen, Vissarion Belinsky, and Mikhail Bakunin—were for their part profoundly Russian. Belinsky was ill at ease with foreigners, and Herzen and Bakunin, despite many years' residence in France, Germany, England, and Italy, remained not only hostile to the world of European bourgeois liberalism and democracy but also strangely ignorant of it.

The difference between Westernizers and Slavophiles was essentially that between radicals and conservatives, a familiar theme in the history of most European nations. It was the difference between those who wished to pull the whole political structure down and replace it with a new building, according to their own admirable blueprints, and those who preferred to knock down some parts and repair and refurnish others, bit by bit. Another basic difference was that the Slavophiles were Orthodox Christians and the Westernizers either atheists or, like the historian T.N. Granovsky, deists with their own personal faith. Belinsky described the Orthodox church in his famous "Letter to Gogol" (1847) as "the bulwark of the whip and the handmaid of despotism." He maintained that the Russian populace was "by its nature a profoundly atheistic people" and that it viewed the priesthood with contempt. These were but half-truths: the church was indeed subject to the government and upheld autocracy, and priests were often unpopular, but this did not mean that the peasants and a large part of the upper and middle classes were not devoted to the Orthodox faith.

The Slavophiles idealized early Russian history. They believed that there had once been a happy partnership between tsar and people: the tsar had consulted the people through their elected spokesmen in the *zemsky sobor.* This had been changed by Peter the Great when he sought to copy foreign models and interposed an alien bureaucracy, staffed largely by Germans, between himself and his people. The Slavophiles held that Russia should return to the way from which it had strayed under Peter. They asked not for a legislative body of the Western type, still less for parliamentary government, but for a consultative assembly to advise the emperor. This was quite unacceptable to Nicholas, who was proud of Peter the Great and believed himself his political heir. To the Westernizers, on the other hand, Peter the Great was a symbol of radical change, not of autocracy.

The multilingual and multireligious empire

The Russian Empire. Russia in the 19th century was both a multilingual and a multireligious empire. Only about half the population was at the same time Russian by language and Orthodox by religion. The Orthodox were to some extent privileged in comparison with the other Christians; all Christians enjoyed a higher status than Muslims; and the latter were not so disadvantaged as the Jews. The basis of legitimacy was obedience to the tsar: Nicholas expected all his subjects to obey him, but he did not expect non-Russians to become Russians. Admittedly, he detested the Poles, but that was because they had been disloyal subjects and revolted against him.

The idea that Russians, as such, should have a status superior to that of other peoples of the empire was distasteful to Nicholas. Russian nationalism nevertheless received some support from Count Uvarov, who, in his famous report to the tsar in 1832, proclaimed three principles as "truly Russian": Orthodoxy, autocracy, and the national principle (*narodnost*). In 1833 Uvarov set up a new university in Kiev to be the centre for a policy of spreading Russian language and culture through the schools in the western provinces, at the expense of the Polish. Nicholas approved of this, for the Poles had been guilty of rebellion, but when the attempt was made to Russify the Germans of the Baltic provinces, he objected. The Baltic Germans were loyal subjects and provided admirable officers and officials: they were therefore allowed to preserve their German culture and to maintain their cultural and social domination over the Estonians and Latvians. The young Slavophile and landowning nobleman Yury Samarin, a junior official in Riga, was severely reprimanded by the emperor for his anti-German activities.

The most revolutionary of the Decembrist leaders, Pavel Pestel, had insisted that all non-Russian peoples of the empire except the Poles should "completely fuse their nationality with the nationality of the dominant people." Another group of Decembrists, however, the Society of United Slavs, believed in a federation of free Slav peoples, including some of those living under Austrian and Turkish rule. In 1845 this idea was put forward in a different form in the Brotherhood of SS. Cyril and Methodius, in Kiev. This group, among whose members was the Ukrainian poet Taras Shevchenko, believed that a federation of Slav peoples should include the Ukrainians, whom they claimed were not a part of the Russian nation but a distinct nationality. The society was crushed by the police, and Shevchenko was sent as a private soldier to the Urals; Nicholas himself gave orders that the great poet should be forbidden to write or draw. But Ukrainian national consciousness, though still confined to an educated minority, was growing, and nothing did more to crystallize Ukrainian as a literary language than Shevchenko's poetry.

Russian victories in Asia

During the first half of the century, Russia made substantial conquests in Asia. In the Caucasus, the kingdom of Georgia united voluntarily with Russia in 1801, and other small Georgian principalities were conquered in the next years. Persia ceded northern Azerbaijan, including the peninsula of Baku, in 1813 and the Armenian province of Erivan (Yerevan) in 1828. The mountain peoples of the northern Caucasus, however, proved more redoubtable. The Chechens, led by Shāmil, resisted Russian expeditions from 1834 until 1859, and the Circassians were not finally crushed until 1864. In the 1840s Russian rule was established over the pastoral peoples of Kazakstan. In East Asia, Russian ships explored the lower course of the Amur River and discovered the straits between Sakhalin and the mainland of Asia in 1849. The Russian-American Company, founded in 1799, controlled part of the coast and islands of Alaska.

Foreign policy. At the beginning of the 19th century, Russian foreign policy was essentially concentrated on the three western neighbour countries with which it had been preoccupied since the 16th century: Sweden, Poland, and Turkey. The policy toward these countries also determined Russian relations with France, Austria, and Great Britain.

Russo-Swedish relations were settled during the Napoleonic era. When Napoleon met with Alexander at Tilsit, he gave the latter a free hand to proceed against Sweden. After two years of war, in which the Russians did not always fare well, the Swedish government ceded Finland to the tsar in 1809. Alexander became grand duke of Finland, but Finland was not incorporated into the Russian Empire, and its institutions were fully respected. In 1810, when Napoleon's former marshal, Jean-Baptiste Bernadotte, was elected heir to the Swedish throne, he showed no hostility toward Russia. In 1812 he made an agreement recognizing the tsar's position in Finland in return for the promise of Russian support in his aim to annex Norway from Denmark. Bernadotte achieved this in the Treaty of Kiel of 1814, and thereafter the relations between Russia and Sweden, now a small and peaceful state, were not seriously troubled.

Alexander I, influenced by his Polish friend Prince Adam Czartoryski, had plans for the liberation and unity of Poland, which had ceased to exist as a state in the 18th



Russian expansion in Asia.

century, when it was partitioned among Russia, Prussia, and Austria. After his defeat by Napoleon in 1805, Alexander abandoned those plans in favour of an alliance with Prussia. In 1807 Napoleon established a dependency called the Grand Duchy of Warsaw and in 1809 increased its territory at the expense of Austria. Alexander's attempts to win the Poles to his side in 1811 and to persuade Austria to make concessions to them failed; when Napoleon invaded Russia in 1812, he had 100,000 first-class Polish troops fighting for him. After Napoleon's defeat, Alexander was not vindictive. He protected the Poles against the demands of Russian nationalists who wanted revenge and sought once more to create a large Polish kingdom comprising the territories annexed by Russia and Prussia in the partitions of the 18th century. He was opposed at the Congress of Vienna in 1814–15 by Austria and Britain; the ensuing kingdom of Poland, which, though nominally autonomous, was to be in permanent union with the Russian Empire, consisted of only part of the Prussian and Russian conquests.

Alexander was popular in Poland for a time after 1815. But real reconciliation between Poles and Russians was made impossible by their competing claims for the borderlands, which had belonged to the former grand duchy of Lithuania. The majority of the population of this region was Belorussian, Ukrainian, or Lithuanian; its commercial class was Jewish; and its upper classes and culture were Polish. Neither Russians nor Poles considered Belorussians, Ukrainians, or Lithuanians to be nations, entitled to decide their own fates: the question was whether Lithuania was to be Polish or Russian. Russians could argue that most of Lithuania had been part of "the Russian land" until the 14th century, and the Poles that it had been Polish since the 16th. Alexander had some sympathy for the Polish point of view and allowed the Poles to hope that he would reunite these lands with Poland, but the effective political forces in Russia were strongly opposed to any change. The disappointment of Polish hopes for Lithuania was probably the most important single cause of the growing tension between Warsaw and St. Petersburg in the late 1820s, which culminated in the revolt of the Poles in November 1830 and the war of 1831 between Polish and Russian armies. It ended in the defeat of the Poles and the exile of thousands of political leaders and soldiers to western Europe. Poland's constitution and thus its autonomy were abrogated, and there began a policy of Russification of Poland.

International reactions to the Russo-Polish war were of some importance. Although the governments of France and Britain had failed to come to the aid of Poland during the war, there was much sympathy for the Poles in these countries; nonetheless, sympathy alone was not sufficient to influence Russian actions. On the other hand, the governments of Prussia and Austria strongly supported Russia. It is arguable that the cooperation among the three monarchies, which continued over the next two decades and was revived from time to time later in the century, had less to do with their eloquently proclaimed loyalty to monarchical government than with their common interest in suppressing the Poles.

Balkan problems

Turkey had long been the main object of Russian territorial expansion; through a certain inertia of tradition, the Turkish policy had become almost automatic. It was to some extent reinforced by religious motives—by the romantic desire to liberate Constantinople (Istanbul), the holy city of Orthodoxy—but more important in the second half of the 19th century was the desire to assure the exit of Russian grain exports through the Black Sea. During certain periods, Russia sought to dominate Turkey as a powerful ally: this was its policy from 1798 to 1806 and again from 1832 to 1853. When this policy was successful, Russia supported the integrity of the Ottoman Empire and made no territorial demands. When it was not successful, Russia sought to undermine Turkey by supporting rebellious Balkan peoples or, more directly, by war: this was the case in 1806–12, 1828–29, and 1853–56.

The periods of cooperation were more profitable for Russia than those of conflict. During the first period, a promising foothold was established in the Ionian Islands, which had to be abandoned after the Treaty of Tilsit. During the second period of cooperation, Russia achieved a great success with the 1833 Treaty of Hünkâr İskelesi, which in effect opened the Black Sea Straits to Russian warships. Russia achieved a more limited but more durable gain by the Straits Convention of 1841, signed by all the Great Powers and by Turkey, which forbade the passage of foreign warships through either the Dardanelles or the Bosporus as long as Turkey was at peace, thus protecting Russia's position in the Black Sea unless it was itself at war with Turkey.

In the periods of hostility between Russia and Turkey, the main object of Russian expansion was the area later known as Romania—the Danubian principalities of Moldavia and Walachia. In 1812 Moldavia was partitioned between Russia and Turkey: the eastern half, under the name of Bessarabia, was annexed to Russia. In the war of 1828–29, Russian armies marched through the principalities and afterward remained in occupation until 1834. In 1848 the Russians returned, with Turkish approval, to suppress the revolution that had broken out in Bucharest. It appeared to be only a matter of time before the two Romanian principalities were wholly annexed to Russia. This did not occur, however, because of Russia's defeat in the Crimean War.

The Crimean War (1853–56) pitted Russia against Great Britain, France, and Turkey. It arose from a series of misunderstandings and diplomatic errors among the powers in their conflict of interests in the Middle East, particularly over Turkish affairs. It has been called "the unnecessary war." The fact that it was fought in the Crimea was due to Austrian diplomacy. In June 1854 the Russian government accepted the Austrian demand that Russian troops be withdrawn from the Danubian principalities, and in August Austrian troops entered. It is arguable whether, on balance, the presence of Austrian troops benefited Russia by preventing French and British forces from marching on Ukraine or whether it damaged Russia by preventing its troops from marching on Istanbul. The tsar resented the Austrian action as showing ingratitude toward the power that had saved Austria from the Hungarian rebels in 1849. When the British and French were unable to attack in the principalities, they decided to send an expedition to the Crimea to destroy the Russian naval base at Sevastopol. It was here that the war dragged out its course. The war showed the inefficiency of Russia's top military command and of its system of transport and supply. The Russian armies, nevertheless, won victories over the Turks in the Caucasus, and the defense of Sevastopol for nearly a year was a brilliant achievement.

FROM ALEXANDER II TO NICHOLAS II

The emancipation of the serfs

Emancipation and reform. Defeat in the Crimea made Russia's backwardness clear, and the first step toward modernization was the abolition of serfdom. It seemed to the new tsar, Alexander II (reigned 1855–81), that the dangers to public order of dismantling the existing system, which had deterred Nicholas I from action, were less than the dangers of leaving things as they were. As the tsar said to the nobility of Moscow in March 1856, "it is better to abolish serfdom from above than to wait until the serfs begin to liberate themselves from below." The main work of reform was carried out in the ministry of the interior, where the most able officials, headed by the deputy minister Nikolay Milyutin, were resolved to get the best possible terms for the peasants. In this they were assisted by a few progressive landowners, chief among whom was the Slavophile Yury Samarin. But the bulk of the landowning class was determined, if it could not prevent abolition of serfdom, to give the freed peasants as little as possible. The settlement, proclaimed on Feb. 19 (March 3, New Style), 1861, was a compromise. Peasants were freed from servile status, and a procedure was laid down by which they could become owners of land. The government paid the landowners compensation and recovered the cost in annual "redemption payments" from the peasants. The terms were unfavourable to the peasants in many, probably most, cases. In the north, where land was poor, the price of land on which the compensation was based was unduly high: in effect, this served to compensate the landowners for the loss of their serfs and also for the loss of the share that they had previously enjoyed of the peasants' earnings from nonagricultural labour. In the south, where land was more valuable, the plots given to the peasants were very small, often less than they had had for their own use when they were serfs.

It is arguable that the main beneficiary of the reform was neither peasant nor landowner but the state. A new apparatus of government was established to replace the authority of the serf owner. From the *ispravnik,* the chief official of the district, who in 1862 ceased to be elected by the nobility and became an appointed official of the ministry of the interior, the official hierarchy now stretched down to the village notary, the most powerful person at this level, who was assisted by an elder elected by an assembly of householders. The lowest effective centre of power was the village commune (*obshchina*), an institution of uncertain origin but great antiquity, which had long had the power to redistribute land for the use of its members and to determine the crop cycle, but which now also became responsible for collecting taxes on behalf of the government.

Introduction of the zemstvos

Further important reforms followed the emancipation. A new system of elected assemblies at the provincial and county levels was introduced in 1864. These assemblies, known as zemstvos, were elected by all classes including the peasants, although the landowning nobility had a disproportionately large share of both the votes and the seats. The zemstvos were empowered to levy taxes and to spend their funds on schools, public health, roads, and other social services, but their scope was limited by the fact that they also had to spend money on some of the tasks of the central government. In 1864 a major judicial reform was completed. Russia received a system of law courts based on European models, with irremovable judges and a proper system of courts of appeal. Justices of the peace were instituted for minor offenses; they were elected by the county zemstvos. A properly organized, modern legal profession now arose, and it soon achieved very high standards. The old system of endless delays and judicial corruption rapidly disappeared. There were, however, two important gaps in the system: one was that the ministry of the interior had power, regardless of the courts, to banish persons whom it regarded as politically dangerous; the other was that the courts for settling disputes between peasants were maintained and operated on the basis of peasant custom. Their institution by Kiselev in the 1840s had been a well-intentioned reform, but their continuation after emancipation meant that the peasants were still regarded as something less than full citizens.

During the first years of Alexander II's reign there was some demand from a liberal section of the nobility for representative government at the national level—not for full parliamentary rule, still less for a democratic suffrage, but for some sort of consultative assembly in which public issues could be debated and which could put before the emperor the views of at least the educated section of the Russian people. The tsar and his bureaucrats refused to consider this. The principle of autocracy must remain sacred; such was the view not only of bureaucrats but also of such men as Nikolay Milyutin and Yury Samarin, both of whom rested their hopes for the progressive reforms they so ardently desired on the unfettered power of the emperor. Their attitude was essentially that of Pavel Stroganov at the beginning of the century, that the sovereign must not have "his arms tied" and so be prevented from realizing "the plans which he had in favour of the nation." The decision against a national assembly in the early 1860s was a negative event of the greatest importance: it deprived Russia of the possibility of public political education such as that which existed, for example, in contemporary Prussia, and it deprived the government of the services of hundreds of talented men.

From reform to terrorism

Revolutionary activities. The emancipation was received with bitter disappointment by many peasants as well as by the radical intellectuals. The serfs' view of their relationship to the landowners had been traditionally summed up in the phrase, "We are yours, but the land is ours." Now they were being asked to pay for land that they felt was theirs by right. During the 1860s small revolutionary groups began to appear. The outstanding figure was the socialist writer N.G. Chernyshevsky; the extent of his involvement in revolutionary action remains a subject of controversy, but of his influence on generations of young Russians there can be no doubt. In 1861 and 1862 revolutionary leaflets were distributed in St. Petersburg, ranging from the demand for a constituent assembly to a passionate appeal for insurrection. The Polish uprising of 1863 strengthened the forces of repression. An unsuc-

cessful attempt on the tsar's life in 1866 led to a certain predominance of extreme conservatives among Alexander's advisers. Nevertheless, there were still some valuable reforms to come. In 1870 the main cities of Russia were given elected municipal government (on a very narrow franchise), and in 1874 a series of military reforms was completed by the establishment of universal military service. This was the work of Dmitry Milyutin, the brother of Nikolay and like him a liberal, who was minister of war from 1861 to 1881.

In the 1870s revolutionary activity revived. Its centre was the university youth, who were increasingly influenced by a variety of socialist ideas derived from Europe but adapted to Russian conditions. These young people saw in the peasantry the main potential for revolutionary action. In 1873 and 1874 hundreds of the youth, including women, "went to the people," invading the countryside and seeking to rouse the peasants with their speeches. The peasants did not understand, and the police arrested the young revolutionaries. Some were sentenced to prison, and hundreds were deported to remote provinces or to Siberia. It became clear that no progress could be expected from overt action: conspiratorial action was the only hope. In 1876 a new party was founded that took the title of Land and Freedom. Some of its members favoured assassination of prominent officials in reprisal for the maltreatment of their comrades and also as a means to pressure the government in order to extract Western-type political liberties. Experience also had shown them that, while the peasants were physically too scattered to be an effective force and were in any case too apathetic, the workers in the new industrial cities offered a more promising audience. This faction was opposed by others in the party who deprecated assassination, continued to pay more attention to peasants than to workers, and were indifferent to the attainment of political liberties. In 1879 the party split. The politically minded and terrorist wing took the name People's Will and made its aim the assassination of Alexander II. After several unsuccessful attempts, it achieved its aim on March 1 (March 13, New Style), 1881, when the tsar was fatally wounded by a bomb while driving through the capital. All the main leaders of the group were caught by the police, and five of them were hanged.

Shortly before his death the tsar had been considering reforms that would have introduced a few elected representatives into the apparatus of government. His successor, Alexander III (reigned 1881–94), considered these plans. Under the influence of his former tutor, Konstantin Pobedonostsev, the procurator of the Holy Synod, he decided to reject them and to reaffirm the principle of autocracy without change. In 1882 he appointed Dmitry Tolstoy minister of interior. Tolstoy and Pobedonostsev were the moving spirits of the deliberately reactionary policies that followed. Education was further restricted, the work of the zemstvos was hampered, and the village communes were brought under closer control in 1889 by the institution of the "land commandant" (*zemsky nachalnik*)—an official appointed by the ministry of interior, usually a former officer or a local landowner, who interfered in all aspects of peasant affairs. The office of elected justice of the peace was abolished, and the government was authorized to assume emergency powers when public order was said to be in danger. By this time Russian public officials were better paid and educated, and less addicted to crude corruption, than they had been in the reign of Nicholas I, but they retained their arrogant contempt for the public and especially for the poorer classes. The discriminatory laws against Jews and members of dissenting Christian sects remained a source of widespread injustice, hardship, and resentment.

Attempts at repression

The repressive policies of Dmitry Tolstoy worked for a time. But the economic development of the following decades created new social tensions and brought into existence new social groups, from whom active opposition once more developed. The zemstvos were in growing conflict with the central authorities. Even their efforts at social improvement of a quite unpolitical type met with obstruction. The ministry of the interior, once the centre of Russia's best reformers, now became a stronghold of resistance. In the obscurantist view of its leading officials, only the central government had the right to care for the public welfare, and zemstvo initiatives were undesirable usurpations of power. Better that nothing should be done at all than that it should be done through the wrong channels. This attitude was manifested in 1891, when crop failures led to widespread famine and relief work by zemstvo officials met with government obstruction. The revival of political activity may be dated from this year. It was accelerated by the death of Alexander III in 1894 and the succession of his son Nicholas II (reigned 1894–1917), who commanded less fear or respect but nevertheless at once antagonized the zemstvo liberals by publicly describing their aspirations for reforms as "senseless dreams." In the late 1890s moderate liberalism, aiming at the establishment of a consultative national assembly, was strong among elected zemstvo members, who were largely members of the landowning class. A more radical attitude, combining elements of liberalism and socialism, was to be found in the professional classes of the cities, including many persons employed by the zemstvos as teachers, doctors, engineers, or statisticians. The growth of an industrial working class provided a mass basis for socialist movements, and by the end of the century some interest in politics was beginning to penetrate even to the peasantry, especially in parts of the middle Volga valley.

Economic and social development. Liberation from serfdom was a benefit for the peasants that should not be underrated. The decades that followed brought a growth of prosperity and self-reliance to at least a substantial minority. In 1877, when about four-fifths of the land due to be transferred to the former serfs was actually in their possession, this "allotment land" constituted about half of the arable land in 50 provinces of European Russia. A further one-third of the arable land was still owned by the nobility, and the rest belonged to a variety of individual or collective owners. In 1905 substantially more than half the arable land was in allotment land and another 10 percent belonged to individual peasants or to peasant associations; the nobility's share of arable land had fallen to a little more than 20 percent. Peasant land had increased by more than 99 million acres (40 million hectares) between 1877 and 1905, of which more than half had been obtained by purchase from landowners and the remainder by the completion of the transfer of allotment land. Peasant purchases had been assisted by loans from the Peasants' Land Bank, set up by the government in 1882. The Nobles' Land Bank, set up in 1885, made loans to landowners at more favourable rates of interest; it may have retarded, but did not prevent, the passage of land from landowners to peasants. In 1894 the rate of interest charged by the two banks was equalized.

Difficulties in agriculture

Though many peasants improved their position, agriculture remained backward, and poverty increased more rapidly than prosperity. One of the main reasons for this was the indifference of the government to agriculture. The government's economic policy was motivated by the desire for national and military power. This required the growth of industry, and great efforts were made to encourage it. Agriculture was regarded mainly as a source of revenue to pay for industry and the armed forces. Exports of grain made possible imports of raw materials, and taxes paid by peasants filled the state's coffers. The redemption payments were a heavy charge on the peasants' resources, though a gradual fall in the value of money appreciably reduced that burden with the passage of years. Consumption taxes, especially on sugar, tobacco, matches, and oil, affected the peasants, and so did import duties. In 1894 the government introduced a liquor monopoly that drew enormous revenues from the peasants, to whom vodka was a principal solace in a hard life. The techniques and tools of agriculture remained extremely primitive, and farm output low; virtually nothing was done to instruct peasants in modern methods.

The second main cause of peasant poverty was overpopulation. The vast landmass of Russia was, of course, sparsely populated, but the number of persons employed in agriculture per unit of arable land, and relative to output, was extremely high compared to western Europe.

There was a vast and increasing surplus of labour in the Russian villages. Outlets were available in seasonal migration to the southern provinces, where labour was needed on the great estates that produced the grain that Russia exported. Peasants could also move permanently to new land in Siberia, which at the end of the century was absorbing a yearly influx of 200,000, or they could find seasonal work in the cities or seek permanent employment in the growing industrial sector. These alternatives were not enough to absorb the growing labour surplus, which was most acute in the southern part of central Russia and in northern Ukraine, in the provinces of Kursk and Poltava. Peasants competed with each other to lease land from the landlords' estates, and this drove rents up. The existence of the large estates came to be resented more and more, and class feeling began to take the form of political demands for further redistribution of land.

The difficulties of agriculture were also increased by the inefficiency of the peasant commune, which had the power to redistribute holdings according to the needs of families and to dictate the rotation of crops to all members. In doing so, it tended to hamper enterprising farmers and protect the incompetent. In defense of the commune it was argued that it ensured a living for everyone and stood for values of solidarity and cooperation that were more important than mere profit and loss. Russian officials also found it useful as a means of collecting taxes and keeping the peasants in order. The 1861 settlement did provide a procedure by which peasants could leave the commune, but it was very complicated and was little used. In practice, the communal system predominated in northern and central Russia, and individual peasant ownership was widespread in Ukraine and in the Polish borderlands. In 1898 in 50 provinces of European Russia, about 198 million acres of land were under communal tenure, and about 54 million were under individual tenure.

The dispute over the peasant commune divided the ranks both of officialdom and of the government's revolutionary enemies. The ministry of the interior, which stood for paternalism and public security at all costs, favoured the commune in the belief that it was a bulwark of conservatism, of traditional Russian social values, and of loyalty to the tsar. The Socialist Revolutionaries favoured it because they took the view that the commune was, at least potentially, the natural unit of a future socialist republic. The ministry of finance, concerned with developing capitalism in town and country, objected to the commune as an obstacle to economic progress; it hoped to see a prosperous minority of individual farmers as a basis of a new and more modern type of Russian conservatism. The Social Democrats agreed that the commune must and should be replaced by capitalist ownership, but they saw this only as the next stage in the progress toward a socialist revolution led by urban workers.

Industrial growth

The emancipation of the serfs undoubtedly helped capitalist development, though this began rather slowly. A rapid growth of railways came in the 1870s, and in the same decade the exploitation of petroleum began at Baku in Azerbaijan. There was also progress in the textile and sugar industries. Only in the 1890s did the demand for iron and steel begin to be satisfied on a large scale within Russia. By the end of the century there was a massive metallurgical industry in Ukraine, based on the iron ore of Krivoy Rog and the coal of the Donets Basin. The iron industry of the Urals, which lost a large part of its labour force when the serfs became free to leave, lagged far behind. Poland was also an important metallurgical centre. Textiles were concentrated in the central provinces of Moscow and Vladimir; by the end of the century they were drawing much of their raw cotton from the newly conquered lands of Central Asia. Baku was also booming, especially as a supplier of petroleum to the Moscow region. St. Petersburg had begun to develop important engineering and electrical industries. Count Sergey Witte, minister of finance from 1892 to 1903, was able to put Russia on the gold standard in 1897 and to encourage foreign investors. French and Belgian capital was invested mainly in the southern metallurgical industry, British in petroleum, and German in electricity.

Industrial growth began to produce an urban working class, which seemed fated to repeat the history of workers in the early stages of industrial capitalism in Western countries. The workers were unskilled, badly paid, overworked, and miserably housed. Uprooted from the village communities in which they had at least had a recognized place, the peasants' children who flocked into the new industrial agglomerations suffered both physical and moral privation. This was especially true of central Russia, where the surplus of labour kept wages down to the minimum. It was in St. Petersburg, where employers found it less easy to recruit workers, that the transformation of the amorphous mass of urban poor into a modern working class made the most progress. St. Petersburg employers were also less hostile to government legislation on behalf of the workers.

Labour reforms

In 1882 Finance Minister Nikolay Khristyanovich Bunge introduced an inspectorate of labour conditions and limited hours of work for children. In 1897 Witte introduced a maximum working day of 11½ hours for all workers, male or female, and of 10 hours for those engaged in night work. Trade unions were not permitted, though several attempts were made to organize them illegally. The ministry of the interior, being more interested in public order than in businessmen's profits, occasionally showed some concern for the workers. In 1901 the head of the Moscow branch of the security police, Colonel Sergey Vasilyevich Zubatov, encouraged the formation of a workers' society intended to rally the workers behind the autocracy, but it was largely infiltrated by Social Democrats. Strikes were strictly forbidden but occurred anyway, especially in 1885, 1896, 1902, and 1903.

A Russian business class also developed rapidly under the umbrella of government policy, benefiting especially from the high protective tariffs and the very high prices paid for government purchases from the metallurgical industry. Russia's industrial progress took place under private capitalism, but it differed from classical Western capitalism in that the motivation of Russian industrial growth was political and military, and the driving force was government policy. Russian and foreign capitalists provided the resources and the organizing skill, and they were richly rewarded. The richness of their rewards accounted for a second difference from classical capitalism: Russian capitalists were completely satisfied with the political system as it was. English and French capitalists had material and ideological reasons to fight against absolute monarchs and aristocratic upper classes; Russian businessmen accepted the principle and the practice of autocracy.

Education and ideas. In 1897, at the time of the first modern census in Russia, there were 104,000 persons who had attended or were attending a university—less than 0.1 percent of the population—and 73 percent of these were children of nobles or officials. The number who had studied or were studying in any sort of secondary school was 1,072,977, or less than 1 percent of the population, and 40 percent of these were children of nobles and officials. In 1904 primary schools managed by the ministry of education had rather more than 3,000,000 pupils, and those managed by the Orthodox church not quite 2,000,000. The combined figure represented only 27 percent of the children of school age in the empire at that time. Persistent neglect of education could no longer be explained by sheer backwardness and lack of funds: the Russian Empire of 1900 could have afforded a modern school system if its rulers had thought it important.

The Russian intelligentsia

In the last half of the 19th century, the word *intelligentsia* came into use in Russia. This word is not precisely definable, for it described both a social group and a state of mind. Essentially, the intelligentsia consisted of persons with a good modern education and a passionate preoccupation with general political and social ideas. Its nucleus was to be found in the liberal professions of law, medicine, teaching, and engineering, which grew in numbers and social prestige as the economy became more complex; yet it also included individuals from outside those professions—private landowners, bureaucrats, and even army officers. The intelligentsia was by its very nature opposed to the existing political and social system, and this opposition coloured its attitude toward culture in general. In

particular, the value of works of literature was judged by the intelligentsia according to whether they furthered the cause of social progress. This tradition of social utilitarianism was initiated by the critic Vissarion Belinsky and carried further by N.A. Dobrolyubov in the late 1850s. Its most extreme exponent was Dmitry I. Pisarev, who held that all art was useless and that the only aim of thinking people should be "to solve forever the unavoidable question of hungry and naked people." In the last decades of the century the chief spokesman of social utilitarianism was the sociological writer Nikolay K. Mikhaylovsky, a former supporter of the terrorist group People's Will. It is hardly an exaggeration to say that Russian literature was faced with two censorships—that of the official servants of the autocracy and that of the social utilitarian radicals. Yet the great writers of this period—Leo Tolstoy, Fyodor Dostoyevsky, and others—though profoundly concerned with social issues, did not conform to these criteria.

The intelligentsia did not consist of active revolutionaries, although it preferred the revolutionaries to the government, but it was from the intelligentsia that the professional revolutionaries were largely recruited. The lack of civil liberties and the prohibition of political parties made it necessary for socialists to use conspiratorial methods. Illegal parties had to have rigid centralized discipline. Yet the emergence of the professional revolutionary, imagined in romantically diabolical terms in the "Revolutionary Catechism" of Mikhail Bakunin and Sergey Nechayev in 1869 and sketched more realistically in *What Is to Be Done?* by Vladimir Ilyich Ulyanov, better known as Lenin, in 1902, was not entirely due to the circumstances of the underground political struggle. The revolutionaries were formed also by their sense of mission, by their absolute conviction that they knew best the interests of the masses. For these men and women, revolution was not just a political aim; it was also a substitute for religion. It is worth noting that a proportion of the young revolutionaries of the late 19th century were children of Orthodox priests or persons associated with religious sects. It is also worth noting that the traditional Russian belief in autocracy, the desire for an all-powerful political saviour, and the contempt for legal formalities and processes had left its mark on them. The autocracy of Nicholas II was, of course, odious to them, but this did not mean that autocratic government should be abolished; rather, it should be replaced by the autocracy of the virtuous.

The socialists

Russian revolutionary socialism at the end of the century was divided into two main streams, each of these being subdivided into a section that favoured conspiratorial tactics and one that aimed at a mass movement to be controlled by its members. The Socialist Revolutionaries (founded in 1901 from a number of groups more or less derived from People's Will) first hoped that Russia could bypass capitalism, and when it became clear that this could not be done they aimed to limit its operation and build a socialist order based on village communes. The land was to be socialized but worked by peasants on the principle of "labour ownership." The Social Democrats (founded in 1898 from a number of illegal working-class groups) believed that the future lay with industrialization and a socialist order based on the working class. The Socialist Revolutionaries were divided between their extreme terrorist wing, the "Fighting Organization," and a broader and looser membership that at one end merged imperceptibly with radical middle-class liberalism. The Social Democrats were divided between Lenin's group, which took the name Bolshevik (derived from the Russian word for "majority," after a majority won by his group at one particular vote during the second congress of the party, held in Brussels and London in 1903), and a number of other groups that were by no means united but that came to be collectively known as Menshevik (derived from the word for "minority"). The personal, ideological, and programmatic issues involved in their quarrels were extremely complex, but it is a permissible oversimplification to say that Lenin favoured rigid discipline while the Mensheviks aimed at creating a mass labour movement of the western European type, that the Mensheviks were much more willing to cooperate with nonsocialist liberals than were the Bolsheviks, and that Lenin paid much more attention to the peasants as a potential revolutionary force than did the Mensheviks.

Russification policies. After the Crimean War the Russian government made some attempt to introduce in Poland a new system acceptable to the Polish population. The leading figure on the Polish side was the nobleman Aleksander Wielopolski. His pro-Russian program proved unacceptable to the Poles. Tension increased, and in January 1863 armed rebellion broke out. This rebellion was put down, being suppressed with special severity in the Lithuanian and Ukrainian borderlands. In order to punish the Polish country gentry for their part in the insurrection, the Russian authorities carried out a land reform on terms exceptionally favourable to the Polish peasants. Its authors were Nikolay Milyutin and Yury Samarin, who genuinely desired to benefit the peasants. The reform was followed, however, by an anti-Polish policy in education and other areas. In the 1880s this went so far that the language of instruction even in primary schools in areas of purely Polish population was Russian. At first, all classes of Poles passively acquiesced in their defeat, while clinging to their language and national consciousness, but in the 1890s two strong, though of course illegal, political parties appeared—the National Democrats and the Polish Socialist Party, both fundamentally anti-Russian.

The repression of nationalities

After 1863 the authorities also severely repressed all signs of Ukrainian nationalist activity. In 1876 all publications in Ukrainian, other than historical documents, were prohibited. In Eastern Galicia, however, which lay just across the Austrian border and had a population of several million Ukrainians, not only the language but also political activity flourished. Here the great Ukrainian historian Mikhail Hrushevsky and the socialist writer Mikhail Drahomanov published their works; Ukrainian political literature was smuggled across the border. In the 1890s small illegal groups of Ukrainian democrats and socialists existed on Russian soil.

From the 1860s the government embarked on a policy designed to strengthen the position of the Russian language and nationality in the borderlands of the empire. This policy is often described as "Russification," though the government never explicitly demanded that its non-Russian subjects abandon their own nationalities or religions. On the other hand, conversions to Orthodoxy were welcomed, and converts were not allowed to revert to their former religions. The government policy of Russification found its parallel in the overtly Russian nationalist tone of several influential newspapers and journals. Nor was Russian society immune to the attraction of national messianism, as the popularity of Nikolay Yakovlevich Danilevsky's *Russia and Europe* in the decades after its first appearance in 1869 attested.

Ironically, some of the most prominent objects of Russification were peoples who had shown consistent loyalty to the empire and now found themselves confronted by government policies that aimed to curtail the rights and privileges of their culture and nationality. The Germans of the Baltic provinces were deprived of their university, and their ancient secondary schools were Russified. The Latvians and Estonians did not object to action by the government against the Germans, whom they had reason to dislike as landowners and rich burghers, but the prospect of the German language being replaced by the Russian had no attraction for them, and they strongly resented the pressure to abandon their Lutheran faith for Orthodoxy. The attempt to abolish many aspects of Finnish autonomy united the Finns in opposition to St. Petersburg in the 1890s. In 1904 the son of a Finnish senator assassinated the Russian governor-general, and passive resistance to Russian policies was almost universal. Effective and widespread passive resistance also occurred among the traditionally Russophile Armenians of the Caucasus when the Russian authorities began to interfere with the organization of the Armenian church and to close the schools maintained from its funds.

Resistance to Russification

Of the Muslim peoples of the empire, those who suffered most from Russification were the most economically and culturally advanced, the Tatars of the Volga valley. At-

tempts by the Orthodox church to convert Muslims and the rivalry between Muslims and Orthodox to convert small national groups of Finno-Ugrian speech who were still pagans caused growing mutual hostility. By the end of the century the Tatars had developed a substantial merchant class and the beginnings of a national intelligentsia. Modern schools, maintained by merchants' funds, were creating a new Tatar educated elite that was increasingly receptive to modern democratic ideas. In Central Asia, on the other hand, modern influences had barely made themselves felt. In those newly conquered lands, Russian colonial administration was paternalistic and limited: like the methods of "indirect rule" in the British and French empires, it made no systematic attempt to change old ways. Here there was no Russification.

The position of the Jews was hardest of all. As a result of their history and religious traditions, as well as of centuries of social and economic discrimination, the Jews were overwhelmingly concentrated in commercial and intellectual professions. They were thus prominent both as businessmen and as political radicals, hateful to the bureaucrats as socialists and to the lower classes as capitalists. The pogroms, or anti-Jewish riots, which broke out in various localities in the months after the assassination of Alexander II, effectively ended any dreams for assimilation and "enlightenment" on the western European pattern for Russia's Jewish community. At this time there also arose the oft-repeated accusation that anti-Semitic excesses were planned and staged by the authorities, not only in Ukraine in 1881 but also in Kishinev in 1903 and throughout the Jewish Pale of Settlement in 1905. The view of government-sponsored pogroms has not, however, been corroborated by documental evidence, though some local officials were certainly at least remiss in their duties in protecting Jewish lives and properties and at worse in cahoots with the anti-Semitic rioters. The most important result of the 1881 pogrom wave was the promulgation in May 1882 of the notorious "temporary rules," which further restricted Jewish rights and remained in effect to the very end of the Russian Empire. By the turn of the century the terms "Jews" and "revolutionaries" had come to be synonymous for some officials.

Foreign policy. During the second half of the 19th century, Russian foreign policy gave about equal emphasis to the Balkans and East Asia. The friendship with Germany and Austria weakened, and in the 1890s the Triple Alliance of Germany, Austria-Hungary, and Italy stood face to face with a dual alliance of France and Russia.

The Balkans

The demilitarization of the Black Sea coast that had resulted from the Crimean War was ended by the London Conference of 1871, which allowed Russia to rebuild its naval forces. In 1876 the Serbo-Turkish War produced an outburst of Pan-Slav feeling in Russia. Partly under its influence, but mainly in pursuit of traditional strategic aims, Russia declared war on Turkey in April 1877. After overpowering heavy Turkish resistance at the fortress of Pleven in Bulgaria, the Russian forces advanced almost to Istanbul. By the Treaty of San Stefano of March 1878 the Turks accepted the creation of a large independent Bulgarian state. Fearing that this would be a Russian vassal, giving Russia mastery over all the Balkans and the straits, Britain and Austria-Hungary opposed the treaty. At the international Congress of Berlin, held in June 1878, Russia had to accept a much smaller Bulgaria. This was regarded by Russian public opinion as a bitter humiliation, for which the German chancellor Otto von Bismarck was blamed. In 1885–87 a new international crisis was caused by Russian interference in Bulgarian affairs, with Britain and Austria-Hungary again opposing Russia. Once more, Russia suffered a political reverse. In the 1890s, despite the pro-Russian sentiment of many Serbs and Bulgarians, neither country's government was much subject to Russian influence. In the crises that arose in connection with the Turkish Armenians and over Crete and Macedonia, Russian policy was extremely cautious and on the whole tended to support the Turkish government. In 1897 an Austro-Russian agreement was made on spheres of influence in the Balkans.

The attempt of Bismarck to restore Russo-German friendship through the Reinsurance Treaty of 1887, with a view to an ultimate restoration of the alliance of Russia, Germany, and Austria, did not survive Bismarck's fall from power in 1890. The Russian government, alarmed by indications of a closer cooperation between the Triple Alliance and Britain and by some signs of a pro-Polish attitude in Berlin, reluctantly turned toward France. The French needed an ally against both Germany and Britain; the Russians needed French capital, in the form both of loans to the Russian government and of investment in Russian industry. The Franco-Russian alliance was signed in August 1891 and was supplemented by a military convention. Essentially the alliance was directed against Germany, for it was only in a war with Germany that each could help the other.

Russia established diplomatic and commercial relations with Japan by three treaties between 1855 and 1858. In 1860, by the Treaty of Peking, Russia acquired from China a long strip of Pacific coastline south of the mouth of the Amur and began to build the naval base of Vladivostok. In 1867 the Russian government sold Alaska to the United States for $7.2 million. The Treaty of St. Petersburg between Russia and Japan in 1875 gave Russia sole control over all of Sakhalin and gave Japan the Kuril Islands.

The systematic Russian conquest of Turkistan, the region of settled population and ancient culture lying to the south of the Kazak steppes, began in the 1860s. This was watched with distrust by the British authorities in India, and fear of Russian interference in Afghanistan led to the Anglo-Afghan War of 1878–80. In the 1880s Russian expansion extended to the Turkmen lands on the east coast of the Caspian, whose people offered much stiffer military resistance. The Russian conquest of Merv in 1884 caused alarm in Calcutta, and in March 1885 a clash between Russian and Afghan troops produced a major diplomatic crisis between Britain and Russia. An agreement on frontier delimitation was reached in September 1885, and for the next decades Central Asian affairs did not have a major effect on Anglo-Russian relations.

Much more serious was the situation in East Asia. In 1894–95 the long-standing rivalry between the Japanese and Chinese in Korea led to a war between the two Asian empires, which the Japanese won decisively. Russia faced the choice of collaborating with Japan (with which relations had been fairly good for some years) at the expense of China or assuming the role of protector of China against Japan. The tsar chose the second policy, largely under the influence of Count Witte. Together with the French and German governments, the Russians demanded that the Japanese return to China the Liaotung Peninsula, which they had taken in the treaty of peace. Russia then concluded an alliance with China in 1896, which included the establishment of the Russian-owned Chinese Eastern Railway, which was to cross northern Manchuria from west to east, linking Siberia with Vladivostok, and was to be administered by Russian personnel and a Russian police force with extraterritorial rights. In 1898 the Russian government went still further and acquired from China the same Liaotung Peninsula of which it had deprived the Japanese three years earlier. Here the Russians built a naval base in ice-free waters at Port Arthur. They also obtained extraterritorial rights of ownership and management of a southern Manchurian railroad that was to stretch from north to south, linking Port Arthur with the Chinese Eastern Railway at the junction of Harbin. When in 1900 the European powers sent armed forces to relieve their diplomatic missions in Peking, besieged by the Boxer Rebellion, the Russian government used this as an opportunity to bring substantial military units into Manchuria. All of this bitterly antagonized the Japanese. They might have been willing, nonetheless, to write off Manchuria as a Russian sphere of influence provided that Russia recognize Japanese priority in Korea, but the Russian government would not do this. It was not so much that the tsar himself wished to dominate all of East Asia; it was rather that he was beset by advisers with several rival schemes and could not bring himself to reject any of them. The British government, fearing that Russia would be able to establish domination over the Chinese govern-

War with Japan

ment and so interfere with the interests of Britain in other parts of China, made an alliance with Japan in January 1902. Negotiations between Russia and Japan continued, but they were insincere on both sides. On the night of Jan. 26/27 (Feb. 8/9, New Style), 1904, Japanese forces made a surprise attack on Russian warships in Port Arthur, and the Russo-Japanese War began. (H.S.-W./N.V.R.)

THE LAST YEARS OF TSARDOM

The revolution of 1905–06. The Russo-Japanese War brought a series of Russian defeats on land and sea, culminating in the destruction of the Baltic fleet in the Tsushima Strait. The defeat finally brought to a head a variety of political discontents simmering back at home. First the professional strata, especially in the zemstvos and municipalities, organized a banquet campaign in favour of a popularly elected legislative assembly. Then, on Jan. 9 (Jan. 22, New Style), 1905, the St. Petersburg workers, led by the priest Georgy Gapon (leader of the Assembly of Russian Factory Workers), marched on the Winter Palace to present Emperor Nicholas with a loyal petition containing similar but wider-ranging demands. They were met by troops who opened fire on them, and about 130 were killed.

Bloody Sunday

News of this massacre, known as Bloody Sunday, spread quickly, and very soon most of the other social classes and ethnic groups in the empire were in uproar. There were student demonstrations, workers' strikes, peasant insurrections, and mutinies in both the army and navy. The peasants organized themselves through their traditional village assembly, the mir, to decide when and how to seize the land or property of the landlords. The workers, on the other hand, created new institutions, the "Soviets of Workers' Deputies": these, consisting of elected delegates from the factories and workshops of a whole town, organized the strike movement there, negotiated with the employers and police, and sometimes kept up basic municipal services during the crisis.

The revolutionary movement reached its climax in October 1905, with the declaration of a general strike and the formation of a soviet in St. Petersburg itself. Most cities, including the capital, were paralyzed, and Witte, who had just concluded peace negotiations with the Japanese, recommended that the government yield to the demands of the liberals and create an elected legislative assembly. This the tsar reluctantly consented to do, in the Manifesto of Oct. 17 (Oct. 30, New Style), 1905. It did not end the unrest, however. In a number of towns, armed bands of monarchists, known as Black Hundreds, organized pogroms against Jewish quarters and also attacked students and known left-wing activists. In Moscow the soviet unleashed an armed insurrection in December, which had to be put down with artillery, resulting in considerable loss of life. Peasant unrest and mutinies in the armed services continued well into 1906 and even 1907.

Throughout the period from 1905 to 1907, disorders were especially violent in non-Russian regions of the empire, where the revolutionary movement took on an added ethnic dimension, as in Poland, the Baltic provinces, Georgia, and parts of Ukraine. There was also persistent fighting between Armenians and Azerbaijanis in the towns of Transcaucasia.

A campaign of terrorism, waged by the Maximalists of the Socialist Revolutionary Party against policemen and officials, claimed hundreds of lives in 1905–07. The police felt able to combat it only by infiltrating their agents into the revolutionary parties and particularly into the terrorist detachments of these parties. This use of double agents (or agents provocateurs, as they were often known) did much to demoralize both the revolutionaries and the police and to undermine the reputation of both with the public at large. The nadir was reached in 1908, when it was disclosed that Yevno Azef, longtime head of the terrorist wing of the Socialist Revolutionary Party, was also an employee of the department of police and had for years been both betraying his revolutionary colleagues and organizing the murders of his official superiors.

The split in the Social Democratic Party was deepened by the failure of the 1905 revolution. Both Mensheviks and Bolsheviks agreed that a further revolution would be needed but disagreed fundamentally on the way to bring it about. The Mensheviks favoured cooperation with the "bourgeois" parties in the Duma, the new legislative assembly, in order to legislate civil rights and then use them to organize the workers for the next stage of the class struggle. The Bolsheviks regarded the Duma purely as a propaganda forum, and Lenin drew from 1905 the lesson that in Russia, where the bourgeoisie was weak, the revolutionaries could combine the "bourgeois" and "proletarian" stages of the revolution by organizing the peasantry as allies of the workers. He was also moving closer to Leon Trotsky's theory that the forthcoming Russian revolution, taking place in the country that was the "weak link" of international imperialism, would spark a world revolution. Lenin did not reveal the full extent of the changes in his ideas until 1917, but in 1912 the split with the Mensheviks was finalized when the Bolsheviks called their own congress in Prague that year, claiming to speak in the name of the entire Social Democratic Party.

Deepening of the split between Bolsheviks and Mensheviks

The State Duma. The October Manifesto had, however, split the opposition. The professional strata, now reorganizing themselves in liberal parties, basically accepted it and set about trying to make the new legislature, the State Duma, work in the interest of reform. The two principal socialist parties, the Socialist Revolutionaries and the Social Democrats, saw the manifesto as just a first step, and the Duma (which at first they boycotted) as merely a tribune to be exploited to project their revolutionary ideas.

The empire's Fundamental Laws were amended in 1906 to take account of the Duma. Russia was still described as an "autocracy," though the adjective "unlimited" was no longer attached to the term, and an article confirming that no law could take effect without the consent of the Duma effectively annulled its meaning. Alongside the Duma there was to be an upper chamber, the State Council, half of its members appointed by the emperor and half elected by established institutions such as the zemstvos and municipalities, business organizations, the Academy of Sciences, and so on. Both chambers had budgetary rights, the right to veto any law, and the ability to initiate legislation. On the other hand, the government was to be appointed, as before, by the emperor, who in practice seldom chose members of the Duma or State Council to be ministers. In addition, the emperor had the right to dissolve the legislative chambers at any time and, under Article 87, to pass emergency decrees when they were not in session.

The Duma electoral law, though complicated, did give the franchise to most adult males. The first elections, held in spring 1906, produced a relative majority for the Constitutional Democratic Party (Kadets), a radical liberal group drawn largely from the professional strata: it wished to go beyond the October Manifesto to a full constitutional monarchy on the British model and to grant autonomy to the non-Russian nationalities. The next largest caucus, the Labour Group (Trudoviki), included a large number of peasants and some socialists who had ignored their comrades' boycott. The two parties demanded amnesty for political prisoners, equal rights for Jews, autonomy for Poland and—most important of all—expropriation of landed estates for the peasants. These demands were totally unacceptable to the government, which used its powers to dissolve the Duma. The new premier, Pyotr Arkadyevich Stolypin, then used Article 87 to pass his own agrarian reform (see below) and to institute special summary courts-martial against terrorists; under the jurisdiction of these courts some 600–1,000 suspects were executed.

Pyotr Arkadyevich Stolypin

In early 1907 new elections were held, and to the government's disappointment the Social Democrats, having abandoned their boycott, did very well, coming in as the third largest party, behind the Kadets and the Trudoviki. The monarchists also performed better than before, so that the house was sharply polarized, but with a preponderance on the left. Unable to pass his agrarian law through it or to cooperate with its majority in any other way, Stolypin advised the tsar to dissolve the Second Duma on June 3 (June 16, New Style), 1907.

Nicholas did not, however, abolish the Duma altogether,

as some of his advisers wished. Instead, he and Stolypin altered the electoral law in favour of landowners, wealthier townsfolk, and Russians to the detriment of peasants, workers, and non-Russians. The Third Duma, elected in autumn 1907, and the Fourth, elected in Autumn 1912, were therefore more congenial to the government. The leading caucus in both Dumas was the Union of October 17 (known as the Octobrists), whose strength was among the landowners of the Russian heartland. The Octobrists acknowledged the October Manifesto as a sufficient basis for cooperation with the government and accepted Stolypin's agrarian program as well as his desire to strengthen the position of the Russian nation throughout the empire.

In practice, however, their cooperation did not bear much legislative fruit beyond the agrarian reform. Many nobles were worried by Stolypin's proposed reform of local-government and justice, which would have weakened their dominant position in the localities. They were also alarmed that more and more land was passing from their control to other social classes. Their opposition was articulated by a pressure group known as the United Nobility, which had numerous members in the State Council and close personal links with the imperial court. Stolypin increasingly found that his reform measures, passed by the Duma, were being blocked in the State Council.

Frustrated, but not wanting to lose all momentum, Stolypin fell back on nationalist measures, for which he could rely on support from his right-wing opponents both in the Duma and the State Council. Such was the bill restricting Finland's special liberties, passed in 1910. He proposed introducing zemstvos into the western provinces: since most landowners there were Polish, he added a special provision to bolster the vote of Russian peasants. The right wing of the State Council objected to this weakening of the landowners, and, receiving the tacit support of the emperor, they defeated the vital clause in the bill in March 1911. Stolypin, dismayed and angry, suspended both houses for three days and introduced the western zemstvos under Article 87. This egregious violation of the spirit of the Fundamental Laws lost him the support of the Octobrists, who went into opposition. Stolypin was, then, already fatally weakened politically when he was assassinated in September 1911. His murderer was both a Socialist Revolutionary and a police agent whose motives have remained obscure.

Achievements and failures of the Duma

Although the legislative achievements of the Duma were meagre, it should not be written off as an ineffective body. It voted credits for a planned expansion of education that was on target to introduce compulsory primary schooling by 1922. Although it could not create or bring down governments, it could exert real pressure on ministers, especially during the budget debates in which even foreign and military affairs (constitutionally the preserve of the emperor alone) came under the deputies' scrutiny. These debates were extensively reported in the newspapers, where they could not be censored, and enormously intensified public awareness of political issues. Partly as a result, the period 1905–14 saw a huge growth in the publication of newspapers, periodicals, and books, both in the capital cities and in the provinces.

Not all the results of this heightened political awareness were happy for the government, of course. In 1910–11, following the death of Leo Tolstoy, who had been excommunicated by the Orthodox church and was refused an ecclesiastical burial, there was serious student unrest, and several Moscow State University professors resigned in protest at government arbitrariness. Furthermore, in 1912, after a disorder at the Lena gold mines, where some 200 workers were killed by troops, the workers' movement revived. Strikes and demonstrations broke out in many of the largest cities, culminating in the erection of barricades in St. Petersburg in July 1914. This time, however, the workers were on their own: there was no sign that peasants, students, or professional people were prepared to join their struggle.

One area where the failure to reform had very serious effects was in the church. Most prelates and clergymen wanted to see the Orthodox church given greater independence in relation to the state, perhaps by restoring the patriarchate and assigning authority within the church to a synod elected by clergy and laity. Many also favoured internal reform by strengthening the parish, ending the split between white (parish) and black (monastic) clergy, and bringing liturgy and scriptures closer to the people. An elected church council was to have taken place in 1906 to debate these reforms, but in the end Stolypin and Nicholas decided not to convene it, as they feared its deliberations would intensify political discontent in the country. Thus the church remained under secular domination until 1917 and fell increasingly under the influence of Grigory Yefimovich Rasputin, a starets (holy man) of dubious reputation, who became a favourite of the imperial couple because he was able to stanch the bleeding of their son Alexis, who suffered from hemophilia.

Agrarian reforms. The 1905 revolution showed that the village commune (mir) was not a guarantor of stability, as its protagonists had claimed, but rather an active promoter of unrest. Stolypin's attempt to undermine it was therefore part of his program for restoring order. But he had economic aims in mind as well. He aimed to give peasant households the chance to leave the commune and also to consolidate their strip holdings, enclosing them in one place as privately owned smallholdings in order to lay the basis for a prosperous peasant commercial agriculture.

The reforms, promoted energetically by the minister of agriculture, Aleksandr Vasilevich Krivoshein, enjoyed a tangible if not sensational measure of success. By 1915 some 20 percent of communal households had left the communes, and about 10 percent had taken the further step of consolidating their strips into one holding. All over the country, land settlement commissions were at work surveying, redrawing boundaries, and negotiating with the village assemblies on behalf of the new smallholders. Not unnaturally, individual withdrawals often aroused resentment, and the reform worked more effectively when whole villages agreed to consolidate and enclose their strips. Many households, both within and outside the commune, were joining cooperatives to purchase seeds and equipment or to market their produce. A good many peasants from the more densely settled regions of Russia were migrating to the open spaces of Siberia and northern Turkestan, whither Krivoshein attracted them by offering free land, subsidies for travel, and specialist advice. In nearly all categories, agricultural output rose sharply between 1906 and 1914, though in international grain markets Russia was beginning to lose ground to the United States, Canada, and Argentina.

The non-Russian peoples

While the non-Russian peoples had made considerable political and cultural gains in 1905–06, these were largely reversed after 1907. Ukrainian nationalism gained ground despite the efforts to suppress it and spread from its nucleus among the professional strata to embrace a growing number of both peasants and workers. In Poland, Russian was restored (after a brief interval in 1905–07) as the language of tuition in all schools, while local government assemblies were introduced with artificially inbuilt Russian majorities. The Finnish Diet, resisting a reduction in its powers, was reduced to the status of a provincial zemstvo, and Finland was submitted to direct rule from St. Petersburg.

Among Muslims the reform movement known as Jādid temporarily found an outlet for its political aspirations in the Muslim Group in the Duma. With the new electoral law of 1907, however, nearly all Muslims lost their representation in the house. Many of their leaders subsequently emigrated to Turkey, encouraged by the Young Turk Revolution of 1908. In Central Asia, industrialization and the increasing colonization of the grazing lands of the Turkic nomadic peoples by immigrants from European Russia caused bitter resentment and led to a widespread and violent rebellion that broke out in 1916.

War and the fall of the monarchy. After 1906 Russia for some time had to pursue a cautious foreign policy in order to gain time to carry out reforms at home, to refit its army, and to rebuild its shattered navy. It set about these goals with the help of huge French loans that were contingent on the strengthening of the Franco-Russian alliance in both the diplomatic and military sense.

Excluded as a serious player in East Asia, Russia paid much more attention to the affairs of the Balkans, where the vulnerability of the Habsburg monarchy and that of the Ottoman Empire were generating an increasingly volatile situation. Besides, the Octobrists and many of the Rights who supported the government in the Duma took a great interest in the fate of the Slav nations of the region and favoured more active Russian support for them.

Operating from a position of weakness and under pressure from home, the Russian foreign minister, Aleksandr Petrovich Izvolsky, attempted to conclude a deal with his Austrian counterpart, Aloys, Count Lexa von Aerenthal, whereby Austria would occupy Bosnia and Hercegovina (over which it had exercised nominal suzerainty since 1878) in return for permitting a revision of the Straits Convention that would allow Russia to bring its warships out of the Black Sea if it were at war but Turkey were not. There was subsequent disagreement about what had been agreed and, in the event, Austria occupied Bosnia and Hercegovina unilaterally, without making Russia any reciprocal concessions. Russia protested but was unable to achieve anything, as Germany threw its support unequivocally behind Austria.

Izvolsky had to resign after this public humiliation, and his successor, Sergey Dmitriyevich Sazonov, set about building an anti-Austrian bloc of Balkan states, including Turkey. This failed, but instead Russia was able to sponsor a Serbian-Greek-Bulgarian-Montenegrin alliance, which was successful in the First Balkan War against Turkey (1912–13). This seemed to herald a period of greater influence for Russia in the Balkans. Austria, however, reacted by demanding that the recently enlarged Serbia be denied an outlet to the Adriatic Sea by the creation of a new state of Albania. Russia supported the Serbian desire for an Adriatic port, but the European powers decided in favour of Austria. The Balkan alliance then fell apart, with Serbia and Greece fighting on the side of Turkey in the Second Balkan War (1913).

The assassination of Archduke Francis Ferdinand in June 1914 and the subsequent Austrian ultimatum to Serbia thus placed Russia in a very ticklish situation. If it let Serbia down and yielded yet again to Austrian pressure, it would cease to be taken seriously as a participant in Balkan affairs and its prestige as a European Great Power would be seriously compromised. The alternative was to escalate the Balkan conflict to the point where Germany would come in behind Austria and a general European war would ensue. Understandably by the standards of the time, Russia chose the second alternative, but the result was World War I and the destruction of the monarchy.

Outbreak of World War I

The immediate effect of the outbreak of war was to strengthen social support for the monarchy. The Duma allowed its sessions to be suspended for some months, and a number of a voluntary organizations came into existence to lend support to the war effort. Zemstvo and Municipal unions were set up to coordinate medical relief, supplies, and transport. Unofficial War Industry Committees were established in major cities and some provinces to bring together representatives of local authorities, cooperatives, merchants, industrialists, and workers for mutual consultation on economic priorities. These were supplemented in the summer of 1915 by government-sponsored Special Councils in the fields of defense, transport, fuel, and food supplies. Civil society seemed to be maturing and diversifying as a result of the national emergency.

In 1914 the Franco-Russian alliance proved its value. The German army could have crushed either France or Russia alone, but not both together. The Russian invasion of East Prussia in August 1914 was a failure: in two unsuccessful battles nearly 150,000 Russians were taken prisoner. The invasion did, however, cause the Germans to withdraw troops from their western front and thus enable the French to win the First Battle of the Marne. The entry of Turkey into the war on the side of Germany was a major setback, since it not only created a new front in the Caucasus (where the Russian armies performed rather well) but, by closing the straits, enormously reduced the supplies that the allies could deliver to Russia. The failure of the British and French campaign in the Dardanelles and the entry of Bulgaria into the war on the German side meant that no relief could come from the south.

When the Central Powers launched a spring offensive in 1915, therefore, the Russian army was already short of munitions. The Germans and Austrians were able to occupy the whole of Poland and begin advancing into the western provinces and the Baltic region, unleashing a flood of refugees, who aggravated the already serious transport situation.

The military reverses of 1915, and especially the shortage of munitions, generated a strong swell of opinion in the Duma and State Council in favour of trying to compel the government to become more responsive to public opinion. The centre and left of the State Council combined with all the centre parties in the Duma, from the Moderate Rights to the Kadets, to form a Progressive Bloc. Its aim was to bring about the formation of a "government enjoying public confidence," whose ministers would be drawn if possible partly from the legislative chambers. The bloc called for a broad program of political reform, including the freeing of political prisoners, the repeal of discrimination against religious minorities, emancipation of the Jews, autonomy for Poland, elimination of the remaining legal disabilities suffered by peasants, repeal of anti-trade-union legislation, and democratization of local government. This program had the support of eight ministers, at least as a basis for negotiation, but not of the premier, Ivan Logginovich Goremykin, who regarded it as an attempt to undermine the autocracy.

Progressive Bloc

The emperor did not approve of the Progressive Bloc either. He entertained quite different notions of how to deal with the crisis. In August 1915 he announced that he was taking personal command of the army and moved with his suite to Mogilyov, in Belorussia, where he remained until the revolution. During the next few months he dismissed all eight ministers who had supported the Progressive Bloc. Cut off from the capital city, Nicholas found it increasingly difficult to play the coordinating role that was so vital, while he still insisted that he was autocrat. From afar he ordained frequent pointless ministerial changes (dubbed by malicious gossip "ministerial leapfrog"), partly under the influence of his wife and Rasputin. Even loyal monarchists despaired of the situation, and in December 1916 Rasputin was murdered in a conspiracy involving some of them.

Ironically, the military situation improved under the emperor's direct command. The Polish and Baltic fronts were stabilized, and in 1916 General Aleksey Alekseyevich Brusilov launched a successful offensive in Galicia, took nearly 400,000 Austrian prisoners, and captured Chernovtsy (Czernowitz).

In the end it was the economic effect of the war that proved too much for the government. The shock of the munitions shortage prompted a partly successful reorganization of industry to concentrate on military production, and by late 1916 the army was better supplied than ever before. But life on the home front was grim. The German and Turkish blockade choked off most imports. The food supply was affected by the call-up of numerous peasants and by the diversion of transport to other needs. The strain of financing the war generated accelerating inflation, with which the pay of ordinary workers failed to keep pace. Strikes began in the summer of 1915 and increased during the following year, taking on an increasingly political tinge and culminating in a huge strike centred on the Putilov armament and locomotive works in Petrograd (the name given to St. Petersburg in August 1914) in January 1917. The government made matters worse by arresting all the members of the worker group of the Central War Industries Committee.

The February (March, New Style) revolution began among the food queues of the capital, which started calling for an end to autocracy. Soon workers from most of the major factories joined the demonstrations. The vital turning point came when Cossacks summoned to disperse the crowds refused to obey orders and troops in the city garrison mutinied and went over to the insurgents. The workers and soldiers rushed to re-create the institution they remembered from 1905, the Soviet of Workers' and

Soldiers' Deputies. Soon their example was followed in many other towns and army units throughout the empire.

By agreement between the Petrograd soviet and the Duma, the Provisional Government was formed, headed by Prince Georgy Yevgenyevich Lvov (chairman of the Zemstvo Union) and consisting mainly of Kadets and Octobrists, though Aleksandr Fyodorovich Kerensky joined it from the Trudoviki. On March 2 (March 15, New Style), this government's emissaries reached Pskov, where the emperor had become stranded in his train, attempting to reach Petrograd. He dictated to them his abdication and thus brought to an end the 300-year-old Romanov dynasty. (H.S.-W./G.A.H.)

Soviet and post-Soviet Russia

RUSSIA AFTER THE MONARCHY

The following is a general overview of the history of Russia during the period of Soviet domination. For full coverage of the history of the Soviet Union, see the article UNION OF SOVIET SOCIALIST REPUBLICS.

The February Revolution

The February Revolution of 1917 was spontaneous, leaderless, and fueled by deep resentment over the economic and social conditions that had prevailed in imperial Russia under Tsar Nicholas. The country, having been sucked into World War I, found the strains of fighting a modern war with a premodern political and economic system intolerable. The tsar was well-meaning but fell short as a war leader and was unable to cope with the burdens of being head of state. His wife Alexandra meddled in government and, while encouraging her husband to be a strong tsar, sought the advice of Rasputin on matters of state. The strain of the war, complicated by the intrigues and machinations within the royal house, caused a great gulf to develop between the monarchy and educated society and between the tsar and the rest of the population.

Hardly a hand was raised in support when the imperial order collapsed in February (March, New Style) 1917. The key factor had been the defection of the military. Without this instrument of coercion, the tsar could not survive. Most Russians rejoiced, but a political vacuum had been created that needed immediate attention. The political government that had been formed was to remain in office until a democratic parliament, the Constituent Assembly, was convened. The new government was bourgeois, or middle-class, representing a tiny segment of the population. However, the soviets, which were proliferating rapidly, did not contest the right of the bourgeoisie to rule.

As Bolshevik domination grew in Petrograd, Moscow, and other major cities, the soviets accepted the idea that the revolution that would give them power would take place in two stages: the bourgeois and the socialist. How long this transition period would last was a debatable point. The Mensheviks, the moderate socialists, held that Russia had to pass through its capitalist phase before the socialist one could appear. The Bolsheviks, the radical socialists, wanted the transition period to be short. Their firebrand leader, Lenin, sensed that power could be seized rather easily. The government was weak, and it could not rely on the army. With its large complement of peasants and workers in uniform, it was this group that formed the natural constituency of the socialists. Like the Mensheviks, the Socialist Revolutionaries, the main agrarian party, did not advocate a rush to power. More than 80 percent of the population lived in the countryside, a fact that made the Socialist Revolutionaries certain to be the leading party when the Constituent Assembly was elected.

The Provisional Government was undone by war, economic collapse, and its own incompetence. Being a temporary administration, it postponed all hard decisions—what should be done about land seizures by the peasants, for example—for the Constituent Assembly. A fatal mistake by the government was its continued prosecution of the war. Middle-class politicians believed wrongly that one of the reasons for the February Revolution was popular anger at the incompetence of the conduct of the war. Disgruntled peasant-soldiers wanted to quit the army. They did not perceive Germany to be a threat to Russian sovereignty and deserted in droves to claim their piece of the landlord's estate. Industrial decline and rising inflation radicalized workers. The Bolshevik slogan of "All Power to the Soviets" was very attractive. Dual power prevailed. The government seemingly spoke for the country, but in reality represented only the middle class; the soviets represented the workers and peasants. Moderate socialists—Mensheviks and Socialist Revolutionaries—dominated the Petrograd and Moscow soviets after February, but the radical Bolsheviks began to win local elections and by September had a majority in the Petrograd Soviet.

Growth in influence of the Bolsheviks

THE OCTOBER (NOVEMBER) REVOLUTION

One of the turning points in the struggle for power was the attempt by General Lavr Kornilov, who had been appointed commander in chief, to take control of Petrograd in August 1917 and wipe out the soviet. Alexandr Kerensky, the prime minister, had been negotiating with Kornilov but then turned away and labeled Kornilov a traitor, perceiving his attack as a possible attempt to overthrow the government. Kerensky agreed to the arming of the Petrograd soviet, but after the failed coup the weapons were retained. The Bolsheviks could now consider staging an armed uprising. Had the Constituent Assembly been called during the summer, it could have undercut Lenin and his close colleague, Leon Trotsky. Probably a majority of the population favoured state power passing to the soviets in October. They envisaged a broadly based socialist coalition government taking over. The October Revolution was precipitated by Kerensky himself, when, angered by claims that the Bolsheviks controlled the Petrograd garrison, he sent troops to close down two Bolshevik newspapers. The Bolsheviks, led by Trotsky, feared that Kerensky would attempt to disrupt the Second All-Russian Congress, scheduled to open on October 25 (November 7, New Style); they reacted by sending troops to take over key communications and transportation points of the city. Lenin, who had been in hiding, appeared on the scene to urge the Bolsheviks to press forward and overthrow the Provisional Government; this they did on the morning of October 26. After the almost bloodless siege, Lenin proclaimed that power had passed to the soviets.

Lenin, at the Second All-Russian Congress of Soviets in October (November, New Style) 1917, managed to secure and head a solely Bolshevik government—the Council of People's Commissars, or Sovnarkom. The Bolsheviks also had a majority in the Soviet Central Executive Committee, which was accepted as the supreme law-giving body. It was, however, the Central Committee of the Communist Party, the Bolsheviks' party, in which true power came to reside. If the October Revolution was accepted as democratic—supported by a majority of the population—then it ceased to be so soon after its institutionalization. This was the result of deliberate manipulation by Lenin and the Bolsheviks. However, in the immediate post-October days, a majority of the Bolshevik Central Committee favoured a coalition government, and Lenin eventually had to give in. Some Socialist Revolutionaries were added in December 1917, but the first and last coalition government remained in office only until March 1918, when, making great land concessions, the Bolsheviks accepted the defeatist Treaty of Brest-Litovsk, ending Russian participation in World War I. The Socialist Revolutionaries, disagreeing with the terms of the treaty, resigned. The Bolsheviks, through the refined skills of the party leader Yakov Sverdlov, had the Congress of Soviets under control by the summer of 1918. Local soviets continued to defy the Bolsheviks, but to no avail. Democracy received little nurturing and was never institutionalized; politics remained personalized. The cult of the strong leader gradually emerged, with local "Lenins" cropping up throughout the land.

Treaty of Brest-Litovsk

THE CIVIL WAR AND WAR COMMUNISM (1918–21)

The Civil War. One side can start a war, but it takes two to end a war. The Bolsheviks found that this principle applied to themselves after October, when they expected to disengage quickly from World War I. Of the three points of their effective slogan—"Peace, Land, and Bread"—the first proved to be the most difficult to realize. Trotsky, the silver-tongued Bolshevik negotiator, had lectured the

Germans and Austrians on Georg Hegel's philosophy and other abstruse subjects at Brest-Litovsk. He thought that he had time on his side. He was waiting for news of revolution in Berlin and Vienna. It never came, and the Bolsheviks found themselves at the Germans' mercy. The issue of peace or war tore the Bolsheviks apart. Lenin favoured peace at any price, believing that it was purely an interim settlement before inevitable revolution. Nikolay Bukharin, a left-wing Bolshevik in the early Soviet period, wanted revolutionary war, while Trotsky wanted neither war nor peace. Trotsky believed the Germans did not have the military muscle to advance, but they did, and eventually the very harsh peace of the Brest-Litovsk treaty was imposed on Russia. The Socialist Revolutionaries left the coalition, and some resorted to terrorism, the target being the Bolshevik leadership. Ukraine slipped under German influence, and the Mensheviks held sway in the Caucasus. Only part of Russia—Moscow, Petrograd, and much of the industrial heartland—was under Bolshevik control. The countryside belonged to the Socialist Revolutionaries. Given the Bolshevik desire to dominate the whole of Russia and the rest of the former tsarist empire, civil war was inevitable.

The Red Army was formed in February 1918, and Trotsky became its leader. He was to reveal great leadership and military skill, fashioning a rabble into a formidable fighting force. The Reds were opposed by the "Whites," anticommunists led by former imperial officers. There were also the "Greens" and anarchists, who fought the Reds and were strongest in Ukraine; the anarchists' most talented leader was Nestor Makhno. The Allies (Britain, the United States, Italy, and a host of other states) intervened on the White side and provided much matériel and finance. The Bolsheviks controlled the industrial heartland of Russia, and their lines of communication were short. Those of the Whites, who were dispersed all the way to the Pacific, were long. The Reds recruited many ex-tsarist officers but also produced many of their own. By mid-1920 the Reds had consolidated their hold on the country.

Joseph Stalin

The feat of winning the Civil War and the organizational methods adopted to do so made a deep impact on Bolshevik thinking. Joseph Stalin, a party leader, talked about the party in terms of an army. There were political fronts, economic struggles, campaigns, and so on. The Bolsheviks were ruthless in their pursuit of victory. The Cheka, or political police, was formed in December 1917 to protect communist power; it was the forerunner of the notorious KGB. By the end of the Civil War the Cheka had become a powerful force. Among the targets of the Cheka were Russian nationalists who objected strongly to the bolshevization of Russia. They regarded bolshevism as alien and based on western European and not Russian norms. Lenin was always mindful of Great Russian chauvinism, which was one reason he never permitted the formation of a separate Russian Communist Party apart from that of the Soviet Union. Russia, alone of the U.S.S.R.'s 15 republics, did not have its own communist party. It was belatedly founded in 1990.

War Communism. Lenin did not favour moving toward a socialist economy after October, because the Bolsheviks lacked the necessary economic skills. He preferred state capitalism, with capitalist managers staying in place but supervised by the work force. Others, like Bukharin, wanted a rapid transition to a socialist economy. The Civil War caused the Bolsheviks to adopt a more severe economic policy known as War Communism, characterized chiefly by the expropriation of private business and industry and the forced requisition of grain and other food products from the peasants. The Bolsheviks subsequently clashed with the labour force, which understood socialism as industrial self-management. Ever-present hunger exacerbated the poor labour relations, and strikes became endemic, especially in Petrograd. The Bolsheviks, however, pressed ahead, using coercion as necessary. The story was the same in the countryside. Food had to be requisitioned in order to feed the cities and the Red Army. The Reds informed the peasants that it was in their best interests to supply food, because if the landlords came back the peasants would lose everything.

Soviet Russia adopted its first constitution in July 1918 and fashioned treaties with other republics such as Ukraine. The latter was vital for the economic viability of Russia, and Bolshevik will was imposed. It was also imposed in the Caucasus, where Georgia, Armenia, and Azerbaijan were tied to Bolshevik Russia by 1921. Many communists regarded Russia as acquiring imperialist ambitions. Indeed Moscow, under the Georgian Stalin, the Commissar for Nationalities, regarded imperial Russia's territory as its natural patrimony. Russia lost control of the Baltic states and Finland, however. Lenin's nationality policy was based on the assumption that nations would choose to stay in a close relationship with Russia, but this proved not to be the case. Many republics wanted to be independent in order to develop their own brand of national communism. The comrade who imposed Russian dominance was, ironically, Stalin. As commissar for nationalities, he sought to ensure that Moscow rule prevailed.

NEW ECONOMIC POLICY (1921–28)

Forced requisitioning led to peasant revolts, and the Tambov province revolt of 1920 in particular forced Lenin to change his War Communism policy. He and the Bolshevik leadership were willing to slaughter the mutinous sailors of the Kronstadt naval base in March 1921, but they could not survive if the countryside turned against them. They would simply starve to death. A tactical retreat from enforced socialism was deemed necessary, a move that was deeply unpopular with the Bolshevik rank and file. The New Economic Policy (NEP) was inaugurated at the 10th Party Congress in March 1921. A ban on factionalism in the party was also imposed. This ban was needed to prevent local party groups from overturning the decisions of the congress. The key sectors of the economy—heavy industry, communications, and transport—remained in state hands, but light and consumer-goods industries were open to the entrepreneur. The monetary reform of 1923 provided a money tax that brought an end to forced requisitioning. The economy was back to its 1913 level by the mid-1920s, and this permitted a vigorous debate on the future. All Communist Party members agreed that the goal was socialism, and this meant the dominance of the industrial economy. The working class, the natural constituency of the Communist Party, had to grow rapidly. There was also the question of the country's security. Moscow lived in fear of an attack during the 1920s and concluded a number of peace treaties and nonaggression pacts with neighbouring and other countries.

Creation of the U.S.S.R.

Soviet Russia gave way to the Union of Soviet Socialist Republics (U.S.S.R.) in 1922, but this did not mean that Russia gave up its hegemony within the new state. As before, Moscow was the capital, and it dominated the union. Lenin's death in January 1924 set off a succession struggle that lasted until the end of the decade. Stalin eventually outwitted Trotsky, Lenin's natural successor, and various other contenders. Stalin, who had become general secretary of the party in 1922, used the party as a power base. The economic debate was won by those who favoured rapid industrialization and forced collectivization. The NEP engendered not only a flowering of Russian culture but also that of non-Russian and non-Slavic cultures. Russia itself had been an empire with many non-Russian citizens, and the emergence of numerous national elites was a trend of considerable concern to Stalin and his leadership.

RUSSIA UNDER STALIN (1928–53)

Stalin, a Georgian, surprisingly turned out to be a great Russian nationalist. During the 1930s and '40s he promoted Russian history, Russian language, and Russian national and cultural heroes, and he held the Russians up as the elder brother for the non-Slavs to emulate. Industrialization benefited first and foremost Russia. Collectivization, though, met with considerable resistance in rural areas. Ukraine, in particular, suffered harshly at Stalin's hands because of forced collectivization. He encountered strenuous resistance there, for which he never forgave the Ukrainians. His policies thereafter brought widespread starvation to that republic, especially in 1932–33, when possibly millions may have died. Nevertheless

many party officials from Ukraine came to Moscow to make their careers, among them Nikita S. Khrushchev, who succeeded Stalin. The armed forces were dominated by Russians and Ukrainians, but the upper echelons of the Communist Party did not contain as many Ukrainians as might have been expected, given the size of that republic. The political police, on the other hand, had many non-Russians at the top, especially Georgians and Armenians.

Russian industry expanded rapidly under Stalin, with Ukraine in second place. The industrialization of the Caucasus and Central Asia began during the 1930s, and it was the Russians, aided by the Ukrainians, who ran the factories. The labour force was also predominantly Russian, as was the emerging technical intelligentsia. Stalin's nationality policy promoted native cadres and cultures, but this changed in the late 1920s. Stalin appears to have perceived that the non-Russians were becoming dangerously self-confident and self-assertive, and he reversed his nationality policy. He came to the conclusion that a Sovietized Russian elite would be more effective as an instrument of modernization. In the non-Russian republics, Russians and Ukrainians were normally second secretaries of the Communist Party and occupied key posts in the government and political police. Diplomats were predominantly Russian. The Soviet constitution of 1936 was democratic—but only on paper. It rearranged the political and nationality map. The boundaries of many autonomous republics and provinces (*oblasts*) were fashioned in such a way as to prevent non-Russians from forming a critical mass. Moscow's fear was that they would circumvent central authority. For example, Tatars found themselves in the Tatar (Tatarstan) and Bashkir (Bashkiriya) autonomous republics, although Tatars and Bashkirs spoke essentially the same language. Tatars also inhabited the region south of Bashkiriya and northern Kazakstan, but this was not acknowledged, and no autonomous republic was established. Moscow played off the various nationalities to its own advantage. This policy was to have disastrous long-term consequences for Russians, because they were seen as imperialists bent on Russifying the locals. New industry usually attracted Russian and Ukrainian labour rather than the locals, and this changed the demographic pattern of the U.S.S.R. Russians spread throughout the union, and by 1991 there were 25 million living outside the Russian republic, 11 million of these in Ukraine. Russians and Ukrainians made up more than half the population of Kazakstan in 1991. Almost half the population of the capital of Kirgiziya (now Kyrgyzstan) and more than a third of the population of Tashkent, the capital of Uzbekistan, were Russian at the time the union ended in 1991.

Stalin's nationality policy

The German invasion in June 1941 resulted in much of Ukraine being overrun. Many Ukrainians welcomed the Wehrmacht (German armed forces). Stalin was already displeased with the Ukrainians, and this reinforced his feelings. In his victory toast he drank to the Russian triumph over the Germans. This was in line with Stalin's wartime policies, through which he rehabilitated the Russian Orthodox church while identifying himself personally with previous Russian leaders such as the medieval prince Dmitri Donskoy and the tsars Ivan IV the Terrible and Peter I the Great.

The Russians, however, suffered as much as anyone else during the purges and repression that characterized Stalin's reign. Stalin also vandalized Russian cultural monuments and destroyed many fine examples of Russian architecture. He was personally responsible for the destruction of some of Moscow's finest cathedrals. It was as if Stalin were trying to expunge Russia's past and to build a new Russia in his own image. This was ironic given that Stalin spoke Russian with a Georgian accent.

Victory over Germany precipitated an upsurge of Russian national pride. Russia, in the guise of the U.S.S.R., had become a great power and by the 1970s was one of two world superpowers. The advent of the Cold War led to Stalin tightening his grip on his sphere of influence in eastern and southeastern Europe. Russian was imposed as the main foreign language, and Russian economic experience was copied. This was effected by having Russian and other communist officials in ministries. A dense network of treaties enmeshed the region in the Russian web. War reparations went first and foremost to Russian factories. Paradoxically, when the United Nations (UN) was set up, Stalin did not insist that Russia have a separate seat like Ukraine and Belorussia, suggesting that he regarded the U.S.S.R. seat as Russia's.

The Cold War

The Bolsheviks had always been mindful of minorities on their frontiers, and the first deportation of non-Russian minorities to Siberia and Central Asia began in the 1920s. Russian Cossacks also were removed forcibly from their home areas in the North Caucasus and elsewhere because of their opposition to collectivization and communist rule. On security grounds Stalin deported some entire small nationality groups, many with their own territorial base, such as the Chechen and Ingush, from 1944 onward. They were accused of collaborating with the Germans. The Volga Germans were deported in the autumn of 1941 lest they side with the advancing Wehrmacht. Altogether, more than 50 nationalities, embracing about 3.5 million people, were deported to various parts of the U.S.S.R. The vast majority of these were removed from European Russia to Asiatic Russia. Nearly 50 years later, President Boris Yeltsin apologized for these deportations, identifying them as a major source of interethnic conflict in Russia.

The late Stalin period witnessed campaigns against Jews and non-Russians. Writers and artists who dared to claim that Russian writers and cultural figures of the past had learned from the West were pilloried. Russian chauvinism took over, and anything that was worth inventing was claimed to have been invented by a Russian.

RUSSIA UNDER KHRUSHCHEV (1953–64)

After Stalin's death in 1953, a power struggle for leadership ensued, which was won by Nikita Khrushchev. His landmark decisions in foreign policy and domestic programs markedly changed the direction of the Soviet Union, bringing détente with the West and a relaxation of rigid controls within the country. Khrushchev, who rose under Stalin as an agricultural specialist, was a Russian who had grown up in Ukraine. During his reign Ukrainians prospered in Moscow. He took it for granted that Russians had a natural right to instruct less fortunate nationals. This was especially evident in the non-Slavic republics of the U.S.S.R. and in eastern and southeastern Europe. His nationality policies reversed the repressive policies of Stalin. He grasped the nettle of the deported nationalities and rehabilitated almost all of them; the accusations of disloyalty made against them by Stalin were declared to be false. This allowed many nationalities to return to their homelands within Russia, the Volga Germans being a notable exception. (Their lands had been occupied by Russians who, fearing competition from the Germans, opposed their return.) The Crimean Tatars were similarly not allowed to return to their home territory. Their lot was complicated by the fact that Russians and Ukrainians had replaced them in the Crimea, and in 1954 Khrushchev made Ukraine a present of the Crimea. Khrushchev abided by the nationality theory that suggested that all Soviet national groups would come closer together and eventually coalesce; the Russians, of course, would be the dominant group. The theory was profoundly wrong. There was in fact a flowering of national cultures during Khrushchev's administration, as well as an expansion of technical and cultural elites.

Changes in nationality policy

Khrushchev sought to promote himself through his agricultural policy. As head of the party Secretariat after Stalin's death, he could use that vehicle to promote his campaigns. *Pravda,* the party newspaper, served as his mouthpiece. His main opponent in the quest for power, Georgy M. Malenkov, was skilled in administration and headed the government. *Izvestiya,* the government's newspaper, was Malenkov's main media outlet. Khrushchev's agricultural policy involved a bold plan to expand rapidly the sown area of grain. He chose to implement this policy on virgin land in the North Caucasus and West Siberia, lying in both Russia and northern Kazakstan. The Kazak party leadership was not enamoured of the idea, since they did not want more Russians in their republic. The Kazak

leadership was dismissed, and the new first secretary was a Malenkov appointee; he was soon replaced by Leonid I. Brezhnev, a Khrushchev protégé, who eventually replaced Khrushchev as the Soviet leader. Thousands of young communists descended on Kazakstan to grow crops where none had been grown before.

Khrushchev's so-called "secret speech" at the 20th Party Congress in 1956 had far-reaching effects on both foreign and domestic policies. Through its denunciation of Stalin, it substantially destroyed the infallibility of the party. The congress also formulated ideological reformations, which softened the party's hard-line foreign policy. De-Stalinization had unexpected consequences, especially in eastern and southeastern Europe in 1956, where unrest became widespread. The Hungarian uprising in that year was brutally suppressed, with Yury V. Andropov, Moscow's chief representative in Budapest, revealing considerable talent for double-dealing. (He had given a promise of safe conduct to Imre Nagy, the Hungarian leader, but permitted, or arranged for, Nagy's arrest.) The events in Hungary and elsewhere stoked up anti-Russian fires.

Khrushchev had similar failures and triumphs in foreign policy outside the eastern European sphere. Successes in space exploration under his regime brought great applause for Russia. Khrushchev improved relations with the West, establishing a policy of peaceful coexistence that eventually led to the signing of the Nuclear Test-Ban Treaty of 1963. But he was at times eccentric and blunt, traits that sometimes negated his own diplomacy. On one occasion he appeared at the UN and, in his speech, emphasized his point by banging a shoe on his desk. Such behaviour tended to reinforce certain Western prejudices about oafish, peasant behaviour by Soviet leaders and harmed the Russian image abroad. Khrushchev's offhanded remarks occasionally caused massive unrest in the world. He told the United States, "We will bury you," and boasted that his rockets could hit a fly over the United States, statements that added to the alarm of Americans, who subsequently increased their defense budget. Hence he turned out to be his own worst enemy, accelerating the arms race with the United States rather than decelerating it, which was his underlying objective. His alarmingly risky policy of installing nuclear weapons in Cuba and permitting local Soviet commanders to use them should they perceive that the Americans were attacking brought the world seemingly close to the brink of nuclear war.

Relations with the West

Khrushchev was a patriot who genuinely wanted to improve the lot of all Soviet citizens. Under his leadership there was a cultural thaw, and Russian writers who had been suppressed began to publish again. Western ideas about democracy began to penetrate universities and academies. These were to leave their mark on a whole generation of Russians, most notably Mikhail S. Gorbachev, who was the last leader of the Soviet Union. Khrushchev had led the Soviet Union away from the harsh Stalin period. Under his rule Russia continued to dominate the union but with considerably more concern for minorities. Economic problems, however, continued to plague the union. Khrushchev attempted to reform the industrial ministries and their subordinate enterprises but failed. He discovered that industrial and local political networks had developed, which made it very difficult for the central authority to impose its will. Under him there was a gradual dissipation of power from Moscow to the provinces. This strengthened the Russian regions. The agricultural policy, which was successful for a few years, eventually fell victim to lean drought years, causing widespread discontent.

RUSSIA UNDER BREZHNEV (1964–82)

After Khrushchev came the triumvirate of Brezhnev, Aleksey N. Kosygin, and N.V. Podgorny. The first was the party leader, the second headed the government, and the third became chairman of the Presidium of the Supreme Soviet, a ceremonial position. By the late 1960s Brezhnev was clearly the dominant leader. His strengths were in manipulating party and government cadres, but he was weak on policy ideas. Kosygin dominated the government, but the 1968 Prague Spring, an unprecedented period of free political expression in Czechoslovakia, led to the Warsaw Pact invasion of that country in August, ruining any lasting hopes of radical economic reform. Kosygin knew such reform was necessary to prevent the slow decline of the country. The U.S.S.R. had reached its apogee in the mid-1970s: it had acquired nuclear parity with the United States and was recognized as a world superpower. Détente flourished in the 1970s but was disrupted by the Soviet invasion of Afghanistan in December 1979. The economic decline of the U.S.S.R. had become such a severe problem that at the time of Brezhnev's death in 1982 there was negative growth. The Brezhnev period was labeled by Gorbachev as one of stagnation, but this may not be an accurate assessment. Stagnation only really set in toward the end of his regime.

Under Brezhnev, Russia dominated the U.S.S.R. as never before. Three-fourths of the defense industries, the priority sector, were in Russia, and the republic accounted for about three-fourths of the Soviet gross national product. The rapid expansion of the chemical, oil, and gas industries boosted exports so that Russia earned most of the union's hard-currency income. There was unprecedented stability of cadres under Brezhnev, and the result of this was rampant corruption. Under Brezhnev the state gradually lost its monopoly on information control. A counterculture, influenced by Western pop music, especially rock, spread rapidly. Russian youth had become enamoured of Western pop stars, and the advent of the audiocassette made it easier to experience their music. The widespread teaching of foreign languages further facilitated access to outside ideas. By the end of the Brezhnev era the Russian intelligentsia had rejected Communist Party values. The party's way of dealing with uncomfortable critics, such as the dissenting novelist Aleksandr Solzhenitsyn, was to deport them. These exiles then became the voice of Russian culture abroad. The academician Andrey Sakharov could not be imprisoned for fear of Western scientists cutting off contact with the Soviet Union, but he was exiled to the closed city of Gorky (now Nizhny Novgorod). Sakharov was released in 1986 and returned to Moscow. In 1989 he was elected to the Congress of People's Deputies, and many of the causes for which he originally suffered became official policy under Mikhail Gorbachev's reforms.

Russian domination

GORBACHEV: PERESTROIKA AND GLASNOST

When Brezhnev died in 1982 the plight of the U.S.S.R. had reached crisis proportions. Brezhnev had not been in effective control of the country during his last few years owing to senility, and Kosygin had died in 1980. The Politburo was dominated by old men, and they were overwhelmingly Russian. Non-Russian representation at the top of the party and the government had declined over time. The Soviet crisis, thus, was largely a Russian crisis. Yury V. Andropov and then Konstantin Chernenko led the country from 1982 until 1985, but their administrations failed to address critical problems. When Gorbachev took over as general secretary of the party in March 1985 he was unaware of the depth of the crisis and launched *perestroika,* a program of economic and political restructuring. He believed that, although it was necessary to accelerate economic growth, the system itself was basically sound. The Gorbachev team was even more heavily Russian than that of his predecessors.

Gorbachev perceived that the defense burden, perhaps equivalent to 25 percent of the gross national product, was crippling the country. This had led to cuts in expenditures in education, social services, and medical care. The corruption of the administrative apparatus fueled the popular belief that officials were enjoying illegal privileges and not responding to the needs of the citizens. The legitimacy of the party—that is, the willingness of the population to accept its right to rule—was very low. The party did not study in depth the malaise of society. This was particularly true of ethnic problems, especially within Russia. However, the party did have some knowledge of the looming threat of Islām. (In 1989 Muslims made up more than a sixth of the U.S.S.R.'s population. It was feared that with the decline of communist ideology Islām would reassert itself.) Economic decline led many minority Soviet citizens to blame their misery on the Russians.

Social and economic problems

Gorbachev had also instituted a policy of *glasnost* ("openness"), which some Russian officials and others opposed on the grounds that it was punishing to the Russians, who could now be openly blamed for all the ills of Soviet society. *Glasnost* also allowed the media more freedom of expression, and editorials complaining of depressed conditions and the government's inability to correct them began to appear. Reactionary elements, including anti-Semitic groups, also began to emerge, especially in Leningrad (formerly Petrograd, now St. Petersburg).

Gorbachev concentrated his energies on increasing his authority (that is to say, his ability to make decisions). He did not, however, develop the power to implement these decisions. He became a constitutional dictator—but only on paper. His policies were simply not put into practice. When he took office his main opponent, Yegor Ligachev, was made head of the party's Central Committee Secretariat, one of the two main centres of power (with the Politburo) in the Soviet Union. This prevented Gorbachev from using the party apparatus to implement his views on *perestroika*. By the summer of 1988, however, Gorbachev had become strong enough to emasculate the Central Committee Secretariat and to take the party out of the day-to-day running of the economy. This responsibility was to pass to the local soviets. A new Soviet parliament, the Congress of People's Deputies, was convened in the spring of 1989, with Gorbachev presiding. The new body superseded the Supreme Soviet as the highest organ of state power. The congress elected a new Supreme Soviet, and Gorbachev, who had opted for an executive presidency modeled on the U.S. and French systems, became the Soviet president with broad powers. This meant that all the republics, including first and foremost Russia, could have a similar type of presidency.

Gorbachev never made clear his vision of what *perestroika* meant. His goal was quite plain: to bring the Soviet Union up to economic par with such capitalist countries as Japan, Germany, and the United States. He traveled much abroad and was brilliantly successful in convincing foreigners that the U.S.S.R. was no longer a threat to anyone.

Gorbachev's radical economists, headed by Grigory A. Yavlinsky, told him that Western-style success required a market economy. Gorbachev could never bring himself to accept the market economy fully, although some elements of it were introduced. However, in the autumn of 1990 he lurched to the right. He was beset by ethnic problems, which were exacerbated by economic decline. He allowed the use of force in Georgia, Azerbaijan, and the Baltic states to suppress local strivings for sovereignty and independence. He desperately wanted to keep the Soviet Union together.

Boris Yeltsin

Gorbachev's difficulties opened up opportunities for other politicians in the various republics. In Russia the people turned to Boris Yeltsin. He had been brought to Moscow as part of the new team in 1985 and eventually became party boss of the capital. However, he found that his radical policies to eliminate corruption and to restructure were being blocked and that Gorbachev would not back him. He was removed as the Moscow party leader and from the Politburo in October–November 1987. Gorbachev told him that his political career was at an end. This was a traumatic experience for Yeltsin, and it led to bouts of depression and initially prevented him from developing his own political party or movement. However, Yeltsin set about building up a following among the common people in Moscow. Elections to the Congress of People's Deputies in March 1989 saw him score a stunning victory. When the congress elected the Supreme Soviet as a standing parliament, Yeltsin was not elected to it, because the congress had an overwhelmingly Communist majority. However, a Siberian deputy stepped down in his favour. Yeltsin for the first time had a national platform. In parliament he pilloried Gorbachev, the Communist Party, corruption, and the slow pace of economic reform. Yeltsin became the president of the Russian parliament despite the bitter opposition of Gorbachev. Yeltsin looked directly to the Russian people for support. In March 1991 when Gorbachev launched an all-union referendum about the future Soviet federation, Russia and several other republics added some supplementary questions. One of the Russian questions was whether the voters were in favour of a directly elected president. They were, and they chose Yeltsin. He used his newfound legitimacy to promote Russian sovereignty, to advocate and adopt radical economic reform, to demand Gorbachev's resignation, and to negotiate treaties with the Baltic republics, in which he acknowledged their right to independence and their right to secede from the union. Soviet attempts to discourage Baltic independence led to a bloody confrontation in Vilnius in January 1991, after which Yeltsin called upon Russian troops to disobey orders that would have them shoot unarmed civilians. In June 1990 the Russian republic declared sovereignty, establishing the primacy of Russian law within the republic. This effectively undermined all attempts by Gorbachev to establish a Union of Sovereign Socialist Republics. Yeltsin appeared to be willing to go along with this vision but, in reality, wanted Russia to dominate the new union, replacing the formal leading role of the Soviet Union. The Russian parliament passed radical reforms that would introduce a market economy, and Yeltsin also cut funding to a large number of Soviet agencies based on Russian soil. Clearly Yeltsin wished to rid Russia of the encumbrance of the Soviet Union and to seek the disbandment of that body.

COLLAPSE OF THE SOVIET UNION

A poorly executed coup attempt occurred Aug. 19–21, 1991, bringing an end to the Communist Party and accelerating the movement to disband the Soviet Union. The coup was carried out by hard-line Communist Party officials, attempting to avert a new liberalized union treaty and return to the old-line party values. The most significant anti-coup role was played by Yeltsin. He brilliantly grasped the opportunity to promote himself and Russia. He demanded the reinstatement of Gorbachev as U.S.S.R. president, but, when Gorbachev returned from house arrest in the Crimea, Yeltsin set out to demonstrate that he was the stronger leader. Yeltsin banned the Communist Party in Russia and seized all of its property. From a strictly legal point of view, this should have been done by court order, not by presidential decree. Russia systematically laid claim to most Soviet property on its territory.

POST-SOVIET RUSSIA

The Commonwealth of Independent States

The U.S.S.R. legally ceased to exist on Dec. 31, 1991. Russia and 10 other former Soviet republics, having declared themselves independent, had founded the Commonwealth of Independent States (CIS) on Dec. 21, 1991, with Minsk as its capital. This new loosely structured grouping is not a state entity, although technically the former Soviet armed forces are CIS forces. Control of these forces immediately came under dispute, especially between Russia and Ukraine. Russia retained the Soviet Union's nuclear superpower status, and, at independence, it controlled the largest conventional forces in the world. All Soviet embassies and consulates became Russian embassies and consulates. Russia also took over the Soviet Union's permanent seat on the UN Security Council. Russia dominates the successor states, which, to one degree or another, harbour historically founded suspicions of Russia's imperialist ambitions.

The new state, called the Russian Federation, set off on the road to democracy and a market economy without any clear conception of how to complete such a transformation in the world's largest country. Like most of the other former republics, it entered independence in a state of serious disorder and economic chaos. Russia was forced to seek aid from Western countries, and, when conditions did not improve, Yeltsin came under criticism.

The country emerged with formidable ethnic problems owing to the fact that it had formed an empire before 1917 and many of the autonomous ethnic regions that were part of the empire no longer wished to be under Russian hegemony. The Russian Orthodox church has reestablished itself as a force in the moral guidance of reborn Russia, but there are many other religions among the minority groups, particularly Islām, which are becom-

ing strong again after decades of Soviet repression. At the same time, many of Russia's republics, home to ethnic minorities, sought greater autonomy. One, the Republic of Chechnia (Chechnya), declared independence in 1991, while others, such as Tatarstan, negotiated additional rights and privileges. In 2000, however, the central government attempted to assert control over the autonomous republics, districts, and regions by subordinating them to seven new federal districts.

Yeltsin's government moved quickly to liberalize the Russian economy by privatizing state-owned enterprises and removing restrictions on market transactions. However, an economic crisis followed these reforms, because former Soviet enterprises were generally inefficient and uncompetitive on the global market, and their new owners often sought to liquidate their holdings and transfer the proceeds abroad. The Russian population suffered widespread economic hardship as a consequence. Differences between Yeltsin and the Russian parliament, particularly over economic reforms, led Yeltsin to dissolve the parliament in September 1993. Armed supporters of the rebellious parliament attempted to seize power in October, but the army rallied to Yeltsin's support and forced the rebels' surrender. After securing a new constitution in December 1993 that increased the powers of the president, Yeltsin worked to consolidate his political control of Russia during the mid-1990s.

In late 1994, Russian troops entered the secessionist Republic of Chechnia and in 1995 bombarded and occupied its capital, Grozny. In 1996 Chechen forces drove the Russians out of Grozny and won a cease-fire agreement that required Russia to withdraw completely from Chechnia. However, after a series of bombings in Russian cities in 1999 that the Russian government attributed to Chechen nationalists, Premier Vladimir Putin reasserted Russian sovereignty over Chechnia and sent Russian forces into the republic once again. The Russians recaptured Grozny in 2000, but the remaining Chechen forces took refuge in the hills to the south and continued to launch guerrilla attacks on Russian positions.

Meanwhile, Russia's economic difficulties continued, and a financial crisis in 1998 led to a currency devaluation, mounting bankruptcies, and a government default on foreign debt. Ongoing economic hardship and allegations of corruption caused Yeltsin's popularity to drop.

On Dec. 31, 1999, Yeltsin resigned and was succeeded by Putin, who promptly granted Yeltsin a pardon for any illegal activities. Putin was confirmed as president by an election in 2000. His government attempted to consolidate its control over Russia's fractious regions and, according to many observers, moved to muzzle media outlets critical of Putin's regime or its war in Chechnia. However, after nearly a decade of contraction, the economy finally began to show signs of growth in 2000, and there were hopes that the new millennium would bring greater economic and political stability to Russia. (Ma.McC./Ed.)

For later developments in the history of Russia, see the BRITANNICA BOOK OF THE YEAR.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia* sections 921, 924, 962, 963, 971, and 972, and the *Index*.

BIBLIOGRAPHY

Physical and human geography. A decade after the collapse of the Soviet Union, there was still a lack of books that dealt authoritatively with post-Soviet Russia as a separate entity. For detailed information, it is still sometimes necessary to refer to volumes covering the former U.S.S.R. as a whole. Comprehensive coverage of Russia is included in RAYMOND E. ZICKEL (ed.), *Soviet Union: A Country Study*, 2nd ed. (1991). Several other texts have chapters devoted to Russia or its various regions, including J.C. DEWDNEY, *The U.S.S.R. in Maps* (1982); I.S. KOROPECKYJ and GERTRUDE E. SCHROEDER (eds.), *Economics of Soviet Regions* (1981); and PAUL E. LYDOLPH, *Geography of the U.S.S.R.*, 5th ed. (1990).

MURRAY FESHBACH *et al.* (eds.), *Environmental and Health Atlas of Russia* (1995), explores the connection between public health and the quality of the environment, providing maps and explanatory essays. Ecological damage suffered during the Soviet period is discussed in MURRAY FESHBACH and ALFRED FRIENDLY, JR., *Ecocide in the USSR: Health and Nature Under Siege* (1992); D.J. PETERSON, *Troubled Lands: The Legacy of Soviet Environmental Destruction* (1993); ZE'EV WOLFSON (BORIS KOMAROV), *The Geography of Survival: Ecology in the Post-Soviet Era* (1994); and MURRAY FESHBACH, *Ecological Disaster: Cleaning Up the Hidden Legacy of the Soviet Regime* (1995).

The physical environments of Siberia and the Far East, which comprise three-quarters of the territory of the Russian Federation and make it an important Asian country, are covered in great detail in S.P. SUSLOV, *Physical Geography of Asiatic Russia* (1961; originally published in Russian, 2nd ed., 1954); R.V. KOVALEV (ed.), *Genesis of the Soils of Western Siberia*, trans. from Russian (1968); T. ROSSWALL and O.W. HEAL (eds.), *Structure and Function of Tundra Ecosystems* (1975); and V. SUKACHEV and N. DYLIS, *Fundamentals of Forest Biogeocoenology* (1968; originally published in Russian, 1964).

The ethnic composition of the population and its implications are discussed in EDWARD ALLWORTH (ed.), *Ethnic Russia in the USSR: The Dilemma of Dominance* (1980); VIKTOR KOZLOV, *The Peoples of the Soviet Union* (1988; originally published in Russian, 1975); and FARLEY MOWAT, *The Siberians* (1970, reissued 1982, also published as *Sibir: My Discovery of Siberia*, 1970). The Russians, as the dominant ethnic group, have received particular attention in HEDRICK SMITH, *The Russians*, updated ed. (1983). Population movements are treated in WOLFGANG LUTZ, SERGEI SCHERBOV, and ANDREI VOLKOV (eds.), *Demographic Trends and Patterns in the Soviet Union Before 1991* (1994); and DONALD W. TREADGOLD, *The Great Siberian Migration: Government and Peasant in Resettlement Emancipation to the First World War* (1957, reprinted 1976). Valuable additional material on many aspects of the Russian republic and its peoples is found in ARCHIE BROWN, MICHAEL KASER, and GERALD S. SMITH (eds.), *The Cambridge Encyclopedia of Russia and the Former Soviet Union*, 2nd ed. (1994); and STEPHEN WHITE (ed.), *Political and Economic Encyclopaedia of the Soviet Union and Eastern Europe* (1990).

The vast resources of Siberia, their development potential, and the problems involved in their exploitation have been the subject of numerous studies, including VIOLET CONOLLY, *Siberia Today and Tomorrow: A Study of Economic Resources, Problems, and Achievements* (1975); YVES LAULAN (ed.), *Exploitation of Siberia's Natural Resources* (1974); ALAN WOOD (ed.), *Siberia: Problems and Prospects for Regional Development* (1987); and ALAN WOOD and R.A. FRENCH (eds.), *The Development of Siberia: People and Resources* (1989). The special problems of the Far East region are the focus of ALLAN RODGERS (ed.), *The Soviet Far East: Geographical Perspectives on Development* (1990). The vital role of transport in Siberian development is the subject of ROBERT N. NORTH, *Transport in Western Siberia: Tsarist and Soviet Development* (1979); and THEODORE SHABAD and VICTOR L. MOTE, *Gateway to Siberian Resources: The BAM* (1977), discussing the Baikal-Amur Mainline Railway.

Historical studies of geopolitical aspects include W.A. DOUGLAS JACKSON, *The Russo-Chinese Borderlands: Zone of Peaceful Contact or Potential Conflict?*, 2nd ed. (1968); WILLIAM O. MCCAGG, JR., and BRIAN D. SILVER (eds.), *Soviet Asian Ethnic Frontiers* (1979); and JAMES H. BATER and R.A. FRENCH (eds.), *Studies in Russian Historical Geography* (1983).

Studies of urbanization in Russia include F.E. IAN HAMILTON, *The Moscow City Region* (1976); G. LAPPO, A. CHIKISHEV, and A. BEKKER, *Moscow, Capital of the Soviet Union* (1976); WRIGHT MILLER, *Leningrad* (1970); and JAMES H. BATER, *The Soviet City: Ideal and Reality* (1980).

In the early 1990s, numerous books were published on the problems leading to and connected with the breakup of the U.S.S.R. Economic issues are discussed in INTERNATIONAL MONETARY FUND *et al., A Study of the Soviet Economy*, 3 vol. (1991), an excellent, wide-ranging survey of the Soviet economy just before the country's dissolution; ANDERS ÅSLUND, *Gorbachev's Struggle for Economic Reform*, updated and expanded ed. (1991); ANDERS ÅSLUND (ed.), *Market Socialism or the Restoration of Capitalism?* (1992); and JOHN E. TEDSTROM (ed.), *Socialism, Perestroika, and the Dilemmas of Soviet Economic Reform* (1990). The ethnic question is the focus of RACHEL DENBER (ed.), *The Soviet Nationality Reader: The Disintegration in Context* (1992); and GAIL W. LAPIDUS, VICTOR ZASLAVSKY, and PHILIP GOLDMAN (eds.), *From Union to Commonwealth: Nationalism and Separatism in the Soviet Republics* (1992).

Current developments not yet discussed in monographic literature are discussed in the journal *Post-Soviet Geography* (monthly). (J.C.D.)

The best general history of Russian literature is VICTOR TERRAS, *A History of Russian Literature* (1991). Outstanding books on the interaction of literature and society in the 19th century include ISAIAH BERLIN, *Russian Thinkers*, ed. by HENRY HARDY and AILEEN KELLY (1978); and, for the Soviet period, RONALD HINGLEY, *Russian Writers and Soviet Society, 1917–1978* (1979). An excellent survey of Soviet culture as a whole is

ANDREI SINYAVSKY (ANDREI SINIAVSKII), *Soviet Civilization: A Cultural History,* trans. from Russian (1990). Important books on Russian art include CAMILLA GRAY, *The Great Experiment: Russian Art, 1863–1922* (1962, reissued as *The Russian Experiment in Art, 1863–1922,* 1971); and ANGELICA ZANDER RUDENSTINE (ed.), *Russian Avant-garde Art: The George Costakis Collection* (1981). The copiously illustrated work by KONSTANTIN RUDNITSKY, *Russian and Soviet Theater, 1905–1932,* trans. from Russian (1988), provides a good introduction to the golden age of Russian theatre. (A.B.Wa.)

History. *From the beginnings to c. 1700.* The best brief survey of early Russian history in English is S.F. PLATONOV, *History of Russia,* trans. from Russian (1925, reprinted 1964). A judicious broad survey is NICHOLAS V. RIASANOVSKY, *A History of Russia,* 5th ed. (1993). Other useful works include V.O. KLUCHEVSKY, *A History of Russia,* 5 vol., trans. from Russian (1911–31, reissued 1960); and PAVEL N. MILIUKOV, *Outlines of Russian Culture,* 3 vol., trans. from Russian (1942). GEORGE VERNADSKY, *Kievan Russia* (1948, reprinted 1973); and M.W. THOMPSON, *Novgorod the Great* (1967), trace the history of Kiev. The Mongol period is treated in GEORGE VERNADSKY, *The Mongols and Russia* (1953); and JOHN FENNELL, *The Crisis of Medieval Russia, 1200–1304* (1983). The history of Muscovy is chronicled by JOHN FENNELL, *The Emergence of Moscow, 1304–1359* (1968), and *Ivan the Great of Moscow* (1961); ROBERT O. CRUMMEY, *The Formation of Muscovy, 1304–1613* (1987); S.F. PLATONOV, *The Time of Troubles: A Historical Study of the Internal Crises and Social Struggle in Sixteenth- and Seventeenth-Century Muscovy* (1970, reissued 1985; originally published in Russian, 1923); GEORGE VERNADSKY, *Russia at the Dawn of the Modern Age* (1959), and *The Tsardom of Moscow, 1547–1682,* 2 vol. (1969); and PAUL DUKES, *The Making of Russian Absolutism, 1613–1801,* 2nd ed. (1990). (E.L.K.)

The 18th century: Interpretative surveys with significant treatment of the 18th century include RICHARD PIPES, *Russia Under the Old Regime* (1974, reissued 1992); and MARC RAEFF, *Understanding Imperial Russia: State and Society in the Old Regime* (1984; originally published in French, 1982). A brief summary of the reforms and reign of Peter I the Great is presented in BENEDICT H. SUMNER, *Peter the Great and the Emergence of Russia* (1950, reprinted 1972). REINHARD WITTRAM, *Peter I, Czar und Kaiser,* 2 vol. (1964), is still an essential comprehensive account, in German. Important insights in the Westernization of Russian culture are found in JAMES CRACRAFT, *The Petrine Revolution in Russian Architecture* (1988); and a chronicle of its impact on Russian consciousness is offered in NICHOLAS V. RIASANOVSKY, *The Image of Peter the Great in Russian History and Thought* (1985). Economic and social aspects of the Petrine revolution are glimpsed through an 18th-century work by a contemporary of the events, IVAN POSOSHKOV, *The Book of Poverty and Wealth,* trans. from Russian and ed. by A.P. VLASTO and L.R. LEWITTER (1987); discussed in PETER I. LYASHCHENKO (PETR I. LIASHCHENKO), *History of the National Economy of Russia, to the 1917 Revolution* (1949, reprinted 1970; originally published in Russian, 2 vol., 1947–48); and illuminated for the entire 18th century in ARCADIUS KAHAN, *The Plow, the Hammer, and the Knout: An Economic History of Eighteenth-Century Russia* (1985).

A critical analysis of the relationship between administration and society in the 18th century is given in JOHN P. LEDONNE, *Ruling Russia: Politics and Administration in the Age of Absolutism, 1762–1796* (1984), and *Absolutism and Ruling Class: The Formation of the Russian Political Order, 1700–1825* (1991). JAMES CRACRAFT, *The Church Reform of Peter the Great* (1971); and GREGORY L. FREEZE, *The Russian Levites: Parish Clergy in the Eighteenth Century* (1977), discuss ecclesiastical affairs. ROBERT O. CRUMMEY, *The Old Believers & the World of Antichrist: The Vyg Community & the Russian State, 1694–1855* (1970), studies the religious dissidence.

The reign and person of Catherine II the Great are magisterially analyzed by ISABEL DE MADARIAGA, *Russia in the Age of Catherine the Great* (1981), and *Catherine the Great: A Short History* (1990). Cultural, educational, and literary matters of the period are discussed in J.L. GARRARD (ed.), *The Eighteenth Century in Russia* (1973); and HAROLD B. SEGEL (ed.), *The Literature of Eighteenth-Century Russia,* 2 vol. (1967), an anthology with informative introduction and notes. Philosophical and political thought is presented in JAMES M. EDIE *et al.* (eds.), *Russian Philosophy,* 3 vol. (1965, reissued 1976); MARC RAEFF (ed.), *Russian Intellectual History* (1966, reissued 1978); and V.V. ZENKOVSKII, *A History of Russian Philosophy,* 2 vol. (1953, reprinted 1967; originally published in Russian, 1948–50). ALEXANDER VUCINICH, *Science in Russian Culture,* vol. 1: *A History to 1860* (1963), examines science and education. The emergence of critical social reflection is discussed in ALLEN MCCONNELL, *A Russian Philosophe, Alexander Radishchev, 1749–1802* (1964, reprinted 1981); and the history-making work itself is ALEKSANDR NIKOLAEVICH RADISHCHEV, *A Journey from St. Petersburg to Moscow,* trans. from Russian by LEO WIENER and ed. by RODERICK PAGE THALER (1958, reprinted 1966). (M.Ra.)

Russia from 1801 to 1904: General surveys of Russian history in the 19th century include J.N. WESTWOOD, *Endurance and Endeavour: Russian History, 1812–1986,* 3rd ed. (1987); and HUGH SETON-WATSON, *The Russian Empire, 1801–1917* (1967, reprinted 1990). The best work in English on the reign of Alexander I is MARC RAEFF, *Michael Speransky, Statesman of Imperial Russia, 1772–1839,* 2nd rev. ed. (1969). The reign of Nicholas I is explored by NICHOLAS V. RIASANOVSKY, *Nicholas I and Official Nationality in Russia, 1825–1855* (1959, reprinted 1969); and W. BRUCE LINCOLN, *Nicholas I, Emperor and Autocrat of All the Russias* (1978, reprinted 1989). Intellectual life during the reign of Nicholas I is described in NICHOLAS V. RIASANOVSKY, *A Parting of Ways: Government and the Educated Public in Russia, 1801–1855* (1976). A useful addition is JOHN S. CURTISS, *The Russian Army Under Nicholas I, 1825–1855* (1965), a general description of the army's performance. A highly detailed operational account as seen by the Russian high command is found in ALBERT SEATON, *The Crimean War: A Russian Chronicle* (1977). The general economic development of Russia in the 19th century is analyzed in G.T. ROBINSON, *Rural Russia Under the Old Régime: A History of the Landlord-Peasant World and a Prologue to the Peasant Revolution of 1917* (1932, reprinted 1969); and ARCADIUS KAHAN, *Russian Economic History: The Nineteenth Century* (1989). The industrialization drive of the 1890s is treated in THEODORE H. VON LAUE, *Sergei Witte and the Industrialization of Russia* (1963, reprinted 1974). An analysis of reform and counter-reform dynamics, specifically Russian, is given in THOMAS S. PEARSON, *Russian Officialdom in Crisis: Autocracy and Local Self-Government, 1861–1900* (1989).

Biographical studies that do much to explain the interplay of cultural and political factors in the 19th century are MARTIN E. MALIA, *Alexander Herzen and the Birth of Russian Socialism, 1812–1855* (1961); and RICHARD PIPES, *Struve: Liberal on the Left, 1870–1905* (1970). The classical description and interpretation of the earlier revolutionary movement and the development of so-called populism is FRANCO VENTURI, *Roots of Revolution: A History of the Populist and Socialist Movements in Nineteenth Century Russia* (1960, reprinted 1983; originally published in Italian, 1952).

Useful books on non-Russian peoples in this period include JAMES W. LONG, *From Privileged to Dispossessed: The Volga Germans, 1860–1917* (1988); JOHN DOYLE KLIER, *Russia Gathers Her Jews: The Origins of the 'Jewish Question' in Russia, 1772–1825* (1986); MICHAEL STANISLAWSKI, *Tsar Nicholas I and the Jews: The Transformation of Jewish Society in Russia, 1825–1855* (1983); and HANS ROGGER, *Jewish Policies and Right-Wing Politics in Imperial Russia* (1986).

Studies of important moments and problems in Russian foreign policy that thoroughly examine the Russian point of view are MARIAN KUKIEL, *Czartoryski and European Unity, 1770–1861* (1955, reprinted 1981); MICHAEL BORO PETROVICH, *The Emergence of Russian Panslavism, 1856–1870* (1956, reprinted 1985); BENEDICT H. SUMNER, *Russia and the Balkans, 1870–1880* (1937, reissued 1962); ANDREW MALOZEMOFF, *Russian Far Eastern Policy, 1881–1904: With Special Emphasis on the Causes of the Russo-Japanese War* (1958, reprinted 1977); and GEORGE F. KENNAN, *The Fateful Alliance: France, Russia, and the Coming of the First World War* (1984). (N.V.R.)

Russia from 1905 to 1917: The best general introduction to the period is HANS ROGGER, *Russia in the Age of Modernisation and Revolution, 1881–1917* (1983). Foreign policy is discussed in DIETRICH GEYER, *Russian Imperialism: The Interaction of Domestic and Foreign Policy, 1860–1914* (1987; originally published in German, 1977); and DOMINIC LIEVEN, *Russia and the Origins of the First World War* (1983). Lieven's *Russia's Rulers Under the Old Regime* (1989) gives a collective portrait of the people who determined the policy. The best introduction to the economy of the period is PETER GATRELL, *The Tsarist Economy, 1850–1917* (1986); while the study by MARGARET MILLER, *The Economic Development of Russia, 1905–1914: With Special Reference to Trade, Industry, and Finance,* 2nd ed. (1967), addresses a particular aspect in a more specialized way.

ABRAHAM ASCHER, *The Revolution of 1905,* 2 vol. (1988–92), covers the subject. The topic is also addressed in ANDREW M. VERNER, *The Crisis of Russian Aristocracy: Nicholas II and the 1905 Revolution* (1990). A more comparative socioeconomic approach to the revolution is demonstrated in TEODOR SHANIN, *The Roots of Otherness: Russia's Turn of Century,* 2 vol. (1986), concentrating especially on the peasantry. ROBERT SERVICE, *Lenin: A Political Life,* 3 vol. (1985–95), is a good introduction to the history of the Social Democratic Party. The reaction of the elites to the revolution is analyzed in ROBERTA THOMPSON MANNING, *The Crisis of the Old Order in Russia: Gentry and Government* (1982); and the reconstruction of the political sys-

tem in TERENCE EMMONS, *The Formation of Political Parties and the First National Elections in Russia* (1983). The politics of the new parliament, the Duma, are outlined in GEOFFREY A. HOSKING, *The Russian Constitutional Experiment: Government and Duma, 1907–1914* (1973); and the social dimension of the new politics is examined in LEOPOLD H. HAIMSON (ed.), *The Politics of Rural Russia, 1905–1914* (1979); and in VICTORIA E. BONNELL, *Roots of Rebellion: Workers' Politics and Organizations in St. Petersburg and Moscow, 1900–1914* (1983). Russia's problems during World War I are described in MICHAEL T. FLORINSKY, *The End of the Russian Empire: A Study in the Economic and Social History of the War* (1931, reprinted 1973); NORMAN STONE, *The Eastern Front, 1914–1917* (1975, reprinted 1985); and RAYMOND PEARSON, *The Russian Moderates and the Crisis of Tsarism, 1914–1917* (1977). (G.A.H.)

Soviet and post-Soviet Russia: For the Soviet period there are hardly any specific histories of Russia: it is always treated in the wider context of the Soviet Union. An overview of the revolution and its consequences is offered in SHEILA FITZPATRICK, *The Russian Revolution* (1982); and ROBERT SERVICE, *The Russian Revolution, 1900–1927,* 2nd ed. (1991). MARTIN MCCAULEY, *The Soviet Union: 1917–1991,* 2nd ed. (1993), provides a general survey. Relevant historical biographies include the one of Lenin cited in the previous paragraph and ADAM B. ULAM, *Stalin: The Man and His Era* (1973, reissued 1989). The Gorbachev era is analyzed in ANGUS ROXBURGH, *The Second Russian Revolution: The Struggle for Power in the Kremlin* (1991); ARCHIE BROWN (ed.), *New Thinking in Soviet Politics* (1992); STEPHEN WHITE, *Gorbachev and After,* 3rd ed. (1992), a solid narrative of the years of *perestroika:* RICHARD SAKWA, *Gorbachev and His Reforms, 1985–1990* (1990), the most detailed account of his reforms; and MIKHAIL GORBACHEV, *Perestroika: New Thinking for Our Country and the World,* new, updated ed. (1988), revealing insights into Gorbachev's thinking and the lack of clarity of vision on how to implement *perestroika.* Good introductions to political structure and situations are RICHARD SAKWA, *Soviet Politics* (1989); GORDON B. SMITH, *Soviet Politics: Struggling with Change,* 2nd ed. (1992); GEOFFREY PONTON, *The Soviet Era: Soviet Politics from Lenin to Yeltsin* (1994), a readable, wide-ranging survey; and GEOFFREY HOSKING, JONATHAN AVES, and PETER J.S. DUNCAN, *The Road to Post-communism: Independent Political Movements in the Soviet Union, 1985–1991* (1992), the most detailed, well-researched treatment of these movements. ALEC NOVE, *An Economic History of the USSR, 1917–1991,* 3rd ed. (1992), is a readable account. MARSHALL I. GOLDMAN, *What Went Wrong with Perestroika* (1991), written by an American economist and long-term student of the Soviet Union, is a good account of the weaknesses and contradictions of *perestroika.* GRAHAM SMITH (ed.), *The Nationalities Question in the Soviet Union* (1990), provides coverage of all the major nationalities. Foreign policy is discussed in ADAM B. ULAM, *Expansion and Coexistence: Soviet Foreign Policy, 1917–73,* 2nd ed. (1974); and JOSEPH L. NOGEE and ROBERT H. DONALDSON, *Soviet Foreign Policy Since World War II,* 4th ed. (1992). AMY W. KNIGHT, *The KGB: Police and Politics in the Soviet Union,* rev. ed. (1990), is the most thorough account; and an insider's story is CHRISTOPHER ANDREW and OLEG GORDIEVSKY, *KGB: The Inside Story of Its Foreign Operations from Lenin to Gorbachev* (1990).

(Ma.McC.)

Russian Literature

The unusual shape of Russian literary history has been the source of numerous controversies. Three major and sudden breaks divide it into four periods—pre-Petrine (or Old Russian), Imperial, post-Revolutionary, and post-Soviet. The reforms of Peter I (the Great; reigned 1682–1725), who rapidly Westernized the country, created so sharp a divide with the past that it was common in the 19th century to maintain that Russian literature had begun only a century before. The 19th century's most influential critic, Vissarion Belinsky, even proposed the exact year (1739) in which Russian literature began, thus denying the status of literature to all pre-Petrine works. The Russian Revolution of 1917 and the Bolshevik coup later in the same year created another major divide, eventually turning "official" Russian literature into political propaganda for the communist state. Finally, Mikhail Gorbachev's ascent to power in 1985 and the collapse of the U.S.S.R. in 1991 marked another dramatic break. What is important in this pattern is that the breaks were sudden rather than gradual and that they were the product of political forces external to literary history itself.

The most celebrated period of Russian literature was the 19th century, which produced, in a remarkably short period, some of the indisputable masterworks of world literature. It has often been noted that the overwhelming majority of Russian works of world significance were produced within the lifetime of one person, Leo Tolstoy (1828–1910). Indeed, many of them were written within two decades, the 1860s and 1870s, a period that perhaps never has been surpassed in any culture for sheer concentrated literary brilliance.

Russian literature, especially of the Imperial and post-Revolutionary periods, has as its defining characteristics an intense concern with philosophical problems, a constant self-consciousness about its relation to the cultures of the West, and a strong tendency toward formal innovation and defiance of received generic norms. The combination of formal radicalism and preoccupation with abstract philosophical issues creates the recognizable aura of Russian classics.

For coverage of related articles in the *Macropædia* and the *Micropædia,* see the *Propædia,* section 621, and the *Index.*

This article is divided into the following sections:

Old Russian literature (10th–17th centuries) 1003
The Kievan period
From the 14th to the 17th century
The 17th century
Imperial literature 1005
The Petrine reforms
The 18th century
The 19th century
The Silver Age
Post-Revolutionary literature 1009
Literature under Soviet rule
From the 1920s to *c.* 1985
Post-Soviet literature 1011
Bibliography 1011

Old Russian literature (10th–17th centuries)

The conventional term "Old Russian literature" is anachronistic for several reasons. The authors of works written during this time obviously did not think of themselves as "old Russians" or as predecessors of Tolstoy. Moreover, the term, which represents the perspective of modern scholars seeking to trace the origin of later Russian works, obscures the fact that the East Slavic peoples (of the lands then called Rus) are the ancestors of the Ukrainian and Belarusian as well as of the Russian people of today. Works of the oldest (Kievan) period also led to modern Ukrainian and Belarusian literature. Third, the literary language established in Kievan Rus was Church Slavonic, which, despite the gradual increase of local East Slavic variants, linked the culture to the wider community known as *Slavia orthodoxa*—that is, to the Eastern Orthodox South Slavs of the Balkans. In contrast to the present, this larger community took precedence over the "nation" in the modern sense of that term. Fourth, some have questioned whether these texts can properly be called literary, if by that term is meant works that are designed to serve a primarily aesthetic function, inasmuch as these writings were generally written to serve ecclesiastic or utilitarian purposes.

Slavia orthodoxa

THE KIEVAN PERIOD

The Kievan period (so called because Kiev was the seat of the grand princes) extends from the Christianization of Russia in 988 to the conquest of Russia by the Tatars (Mongols) in the 13th century. Russia received Christianity from Byzantium rather than from Rome, a fact of decisive importance for the development of Russian culture. Whereas Catholic Poland was closely linked to cultural developments in western Europe, Orthodox Russia was isolated from the West for long periods and, at times, regarded its culture as dangerous. Conversion by Byzantium also meant that the language of the church could be the vernacular rather than, as in the West, Latin; this was another factor that worked against the absorption of Western culture.

Russia was not the first Slavic culture to be converted to Christianity, and a standardized language, the Old Church Slavonic pioneered in the 9th century by Saints Cyril (or Constantine) and Methodius, was already available. Bulgaria, which had been Christianized a century earlier and had offered a home to the Cyrillo-Methodian community, became a conduit for the transmission of Greek culture, translated into Old Church Slavonic, to Russia, which in turn rapidly established its own scribal activities in copying and translating. Thus a significant literary activity of the Kievan period consisted of translating or adapting borrowed works. It is worth stressing that the enormous prestige accorded to translating has continued to be a distinctive characteristic of Russian culture. Even in the 18th, 19th, and 20th centuries, major Russian writers devoted their energies to the translation of foreign works, which in some cases constituted their most significant contribution—a literary fact reflecting Russia's status as a self-conscious cultural borrower for much of its history.

During the Kievan period the selection of translated foreign works circulating in Russia by and large reflected the interests of the church: almost all were from the Greek, and most were of ecclesiastical interest. *Ostromirovo evangeliye* (*The Ostromir Gospel*) of 1056–57 is the oldest dated Russian manuscript. Versions of the four Gospels, the Book of Revelation, guidebooks of monastic rules, homilies, hagiographic collections, and prayers reflect the religious interests of the clerical community. To be sure, translations of secular works also circulated, including Flavius Josephus' *The Jewish War* (which influenced Russian military tales), chronicles, and some tales. But, on the whole, translations offered a rather limited access to Greek culture aside from the ecclesiastical.

A celebrated monument of Old Russian literature is Hilarion's *Slovo o zakone i blagodati* (1037–50; "Sermon on Law and Grace"), an accomplished piece of rhetoric contrasting Old Testament law with New Testament grace. Other significant homiletic works were written by Clement of Smolensk, metropolitan of Russia from 1147 to 1154,

Hagiography

and by St. Cyril of Turov (1130–82). The central genre of Old Russian literature was probably hagiography, and a number of interesting saints' lives date from the earliest period. Both a chronicle account and two lives of Boris and Gleb, the first Russian saints, have survived to the present day. The sanctity of these two men, who were killed by their brother Svyatopolk in a struggle for the throne, consists not in activity but in the pious passivity with which, in imitation of Christ, they accepted death. This ideal of passive acceptance of suffering was to exercise a long-lasting influence on Russian thought.

The monk Nestor (*c.* 1056–after 1113), to whom a life of Boris and Gleb is ascribed, also wrote *Zhitiye prepodobnogo ottsa nashego Feodosiya* ("Life of Our Holy Father Theodosius") (d. 1074). The *Kievo-Pechersky paterik* (*The Paterik of the Kievan Caves Monastery*), closely related to hagiography, collects stories from the lives of monks, along with other religious writings. A saint's life of quite a different sort, *Zhitiye Aleksandra Nevskogo* ("Life of Alexandr Nevsky") (d. 1263), celebrates a pious warrior prince. The tradition of pilgrimage literature also begins in this period. Nestor was involved with compiling the *Povest vremennykh let* ("Tale of Bygone Years"; *The Russian Primary Chronicle*), also called the Primary Chronicle of Kiev (compiled about 1113), which led to the writing of other chronicles elsewhere.

From a literary point of view, the best work of Old Russian literature is the *Slovo o polku Igoreve* (*The Song of Igor's Campaign*), a sort of epic poem (in rhythmic prose, actually) dealing with Prince Igor's raid against the Polovtsy (Kipchak), a people of the steppes, his capture, and his escape. Composed between 1185 and 1187, the Igor Tale, as it is generally known, was discovered in 1795 by Count Musin-Pushkin. The manuscript was destroyed in the Moscow fire of 1812; however, a copy made for Catherine II the Great survived. The poem's authenticity has often been challenged but is now generally accepted. Its theme is the disastrous fratricidal disunity of the Russian princes.

FROM THE 14TH TO THE 17TH CENTURY

Moscow's ascendancy. Beginning in the 1230s, the Tatars conquered most of the Russian lands, thus destroying the hegemony of Kiev and initiating a period in which political and cultural power was dispersed among numerous principalities. Eventually the grand princes of Moscow succeeded in defeating the Tatars (1480) and subduing the principalities. (The exception was the lands under the rule of the Lithuanian-Polish kingdom, and this division initiated the development of separate Ukrainian and Belarusian cultural traditions.) Once the Russian lands were united, Tsar Ivan IV (Ivan the Terrible; reigned 1533–84) undertook a campaign against the remaining power of the old aristocracy (boyars). Reflecting these political facts, chronicles and saints' lives served the interests of particular local powers. A series of works in various genres, known as the Kulikovo cycle, celebrated the first (but by no means definitive) Russian victory over the Tatars in 1380 under the leadership of Grand Prince Dmitry Ivanovich ("Donskoy"). A rather weak imitation of the Igor Tale, the *Zadonshchina* (attributed to Sofony of Ryazan and composed no later than 1393) glorifies Dmitry Donskoy.

The Second South Slavic Influence. The Ottoman occupation of the Balkans at the end of the 14th century, and later the conquest of Constantinople in 1453, drove a number of prelates to Russia, thus initiating the "Second South Slavic Influence." Schooled in an Eastern Christian theological movement, Hesychasm, these men brought with them a style of writing closely linked to their theological doctrines. Known as "word weaving," this ornamental style played with phonic and semantic correspondences. It appears in the most notable hagiography of the period, *Zhitiye svyatogo Sergiya Radonezhskogo* ("Life of Saint Sergius of Radonezh") by Epifany Premudry (Epiphanius the Wise; d. between 1418 and 1422).

"Word weaving"

Possessors and Nonpossessors. A theological and political controversy of great significance took place between St. Joseph of Volokolamsk (1439–1515) and his followers, known as the "Possessors," or "Josephites," and Nil Sorsky (1433–1508) and his followers, known as the "Nonpossessors." Joseph justified the killing of heretics and the church's possession of lands (thus the name "Possessors"). These positions were disputed by Nil and his followers, especially Vassian Patrikeyev (d. before 1545) and Maximus the Greek (*c.* 1475–1556). The Nonpossessors called for greater tolerance and an inner, more spiritual religion, a view that left its mark on a tradition eventually embodied in Dostoyevsky's ideal monk, Father Zosima, in his novel *The Brothers Karamazov.* With the Josephites' triumph, the division between church and state dissolved; apostasy and treason became inseparably linked.

Works reflecting Muscovite power. Accompanying Moscow's rise were a series of writings on the theme of *translatio imperii* ("translation of empire"), which constructed genealogies and described the transmission of imperial and ecclesiastical regalia to Russia. Particularly important is the monk Philotheus' (Filofei's) epistle to Vasily III (written between 1514 and 1521), which proclaimed that, with the fall of Constantinople (the second Rome), Moscow became the third (and last) Rome. Along with the title tsar (caesar) and the claim that Orthodox Russia was the only remaining true Christian state, the doctrine of the Third Rome came to justify Russian imperial ambitions and to legitimize the idea that it was Russia's destiny to save and rule the world.

Reflecting the consolidation of Muscovite power were a series of encyclopaedic works, including the enormous *Velikiye Minei-Cheti* ("Great Martyrologue") of 1552, the *Ulozheniye* ("Code of Laws"), and other collections or codifications. Encyclopaedic writing also includes the famous *Domostroy,* or rules for household management, which later became a byword for oppressive narrow-mindedness. The 16th century also saw the first examples of polemical writing by laymen. Ivan Peresvetov (rather superfluously) urged Ivan the Terrible to inspire fear. From a literary point of view, the most remarkable work of this period is the correspondence between Andrey Mikhaylovich, Prince Kurbsky (1528–83) and Ivan the Terrible. In a series of letters Kurbsky, who escaped from Russia and entered the service of the Polish king, denounced Ivan's tyrannical rule and developed a theory justifying rebellion against unjust power. In a simple but polemically powerful style, which included citations from Cicero, he also denounced Russian cultural backwardness, thus earning a reputation as Russia's first "Westernizer" (as well as first "dissident" and first "émigré" writer). In his vituperative replies, Ivan exhibits the psychology of a victim (self-pitying in accounts of his childhood) turned victimizer.

Correspondence between Kurbsky and Ivan the Terrible

Among the other noteworthy works of this period are some tales of entertainment, including *Povest o Petre i Fevroni* (mid-16th century; "Tale of Peter and Fevroniya"). In his *Khozhdeniye za tri morya* ("Journey Beyond Three Seas") a merchant, Afanasy Nikitin, describes his travels to India and Persia during 1466–72. However, what is most striking about this period is what did not take place: Russia experienced no Renaissance and became quite isolated from the West. With nothing resembling Western secular literature, philosophy, or science, it remained a land remarkable for its lacks.

THE 17TH CENTURY

The 17th century began with a period of political chaos. The ruling Muscovite dynasty came to an end in 1598. Before Michael Romanov was at last proclaimed tsar in 1613, Russia was convulsed by struggles for power, peasant rebellions, and foreign invasions. This Time of Troubles became the topic of a number of historical or memoiristic works, including Avraamy Palitsyn's *Istoriya v pamyat sushchim predydushchim godom* (completed in 1620; "History to Be Remembered by Future Generations").

Western cultural influences gradually penetrated Russia in the 17th century. They entered the country through a number of channels, including the "German [foreign] quarter" in Moscow and through Ukraine, which was united with Russia in 1654. Ukrainian and Belarusian clerics, who had received a Polish-style education at the Kiev Academy, brought Western and Latin culture with them to Moscow. By the end of the 17th century, Russian

Simeon Polotsky

literature had changed in important ways. A key figure in producing these changes was Simeon Polotsky (1629–80), a monk educated at the Kiev Academy. He played the leading role in introducing syllabic poetry (verse that is measured by the number of syllables in each line), based on Polish models, into Russia. Old Russian literature had been dominated entirely by prose, and so Polotsky's verse marked a decisive break. So did the introduction of drama into Russia with Polotsky's school dramas (modeled on Jesuit Counter-Reformation plays having biblical or religious themes), the establishment of a court theatre by Tsar Alexis, and the production of *Artakserksevo deystvo* (1672; "Action of Artaxerxes"), the first court play (in prose), by Johann Gottfried Gregory. The change in literary culture is also evident in the beginnings of prose fiction. Translations of foreign adventure romances appeared, along with Russian stories, parodies, and satires, including the picaresque (and erotic) *Povest o Frole Skobeyeve* ("Tale of Frol Skobeyev") and *Kalyazinskaya chelobitnaya* ("The Kalyazin Petition"). *Povest o Gore-Zlochastii* ("Tale of Woe-Misfortune"), written in folk-epic verse, combines motifs of temptation, adventure, and salvation.

In the mid-17th century liturgical reforms undertaken by Patriarch Nikon split the Russian church. The dissenters (or Old Believers) produced some remarkable work, including the masterpiece of 17th-century Russian writing *Zhitiye protopopa Avvakuma* (1672–73; *The Life of the Archpriest Avvakum*). Avvakum, who eventually was burned at the stake, narrates his life in a powerful vernacular alternating with Church Slavonicisms. Written in prison, his narrative conveys a feel for his fanatic, earthy personality in a paradoxical form that is both autobiography and autohagiography.

The "Ukrainian hegemony" over Russian letters continued during the reign of Peter I the Great. St. Dmitry (Tuptalo) of Rostov, Stefan Yavorsky, and Feofan Prokopovich, the three most important writers of the period, were all educated at the Kiev Academy.

Imperial literature

THE PETRINE REFORMS

The Westernization of Russia. Peter the Great's radical and rapid Westernization of Russia altered the daily life of the upper classes and all high culture. The nobility was made to conform to Western models in its dress, customs, social life, education, and state service; women came out of seclusion; a European calendar was introduced; Russians were sent abroad to study; foreign languages were learned. Western culture was absorbed so rapidly in the course of the 18th century that by the 19th century the first language of the upper nobility was not Russian but French. As a result, a large cultural gap opened between the nobility and the peasantry, whose distance from each other became an important theme of Russian literature. In the context of world history, Russia may be seen as the first of many countries to undergo rapid modernization and Westernization while wrestling with a question capable of different answers: in adopting Western technology and science, is it also necessary to adopt Western culture and forms of living? Under Peter's autocratic will, Russia was forced into an uncompromisingly affirmative answer to this question, which has concerned Russian writers up to the present moment.

St. Petersburg

In 1703 Peter founded a new capital, St. Petersburg. It was built in Western architectural style and populated by his command on an inhospitable swamp. The city—Peter's "window to the West"—became a key theme of literary works, including Aleksandr Pushkin's poem *The Bronze Horseman,* Dostoyevsky's novel *Crime and Punishment,* and Andrey Bely's novel *St. Petersburg.* In contrast to Moscow, St. Petersburg came not only to symbolize the power of the state over the individual but also to stand for reason and planning divorced from tradition, individual human needs, and the nonrational elements of human nature. The hero of Dostoyevsky's *Notes from the Underground* calls the capital the world's "most artificial city," associating it with utopian contempt for tradition and experience. Like Peter's reforms generally, the city evoked the idea of historical change by sudden leaps rather than by a gradual, organic process.

The response of writers and critics. By the 19th century it became commonplace to regard Russia as a young country that had entered history only with the Petrine reforms. The very genres in which 19th-century literature was written had essentially no counterpart in medieval Russia, deriving instead from European literary history. Thus, in tracing their literary past, Russians often felt the necessity of "crossing borders." To be sure, it became common for Russian writers to appropriate old Russian themes, characters, and events, as is the case in Pushkin's *Boris Godunov,* Mikhail Lermontov's *Pesnya pro kuptsa Kalashnikova,* and Tolstoy's *Father Sergius.* But these works are recognizably conscious of overcoming a break. Some scholars have insisted that the idea of a radical break in Russian literary history is mistaken, but there is no doubt that the perception of discontinuity is a key fact of Russian literary history.

An aura of foreignness adhered to high culture, which is one reason why a tradition arose in which the sign of Russianness was the defiance of European generic norms. Justifying the self-consciously odd form of *War and Peace,* Tolstoy observed that departure from European form is necessary for a Russian writer: "There is not a single work of Russian artistic prose, at all rising above mediocrity, that quite fits the form of a novel, a poem, or a story." This (admittedly exaggerated) view, which became a cliché, helps explain the enormous popularity in Russia of those Western writers who parodied literary conventions, such as the 18th-century British novelist Laurence Sterne, as well as the development of Russia's most influential school of literary criticism, Formalism, which viewed formal self-consciousness as the defining quality of "literariness." The sense that culture, literature, and the forms of "civilized" life were a foreign product imported by the upper classes is also reflected in a tendency of Russian thinkers to regard all art as morally unjustifiable and in a pattern of Russian writers renouncing their own works. While English and French critics were arguing about the merits of different literary schools, Russian critics also debated whether literature itself had a right to exist—a question that reveals the peculiar ethos of Russian literary culture.

THE 18TH CENTURY

Formation of a literary language

The 18th century was a period of codification, imitation, and absorption of foreign models. The century's major contribution was the development of a literary language. Under the pressure of new subject matter and the influx of foreign expressions, Church Slavonic proved inadequate, and the resulting linguistic chaos required the standardization of literary Russian. In 1758 Mikhail Lomonosov published "Predisloviye o polze knig tserkovnykh v rossiyskom yazyke" ("Preface on the Use of Church Books in the Russian Language") in which he classified Russian and Church Slavonic words, assigning their use to three styles, and correlated these styles with appropriate themes, genres, and tones. Thus the Russian literary language was to be established by a combination of Russian and Church Slavonic.

Verse also changed decisively. The old syllabic verse, based on qualities of the Polish language, gave way to syllabotonic verse (*i.e.,* verse in which the number of stressed syllables in each line becomes the dominant prosodic element), more suitable to Russian. Theories of versification were advanced by Vasily Trediakovsky in 1735 and 1752 and, especially, by Lomonosov in 1739 (the date Belinsky chose as the beginning of Russian literature). It is also noteworthy that the Petrine assault on the church decisively ended the role of the clergy in Russian literature.

Throughout the 18th century Russian writers imitated, adapted, and experimented with a wide variety of European genres, thus grafting them onto the Russian tradition and making them available for later, more original, use. Much classical and western European literature was translated, read, and assimilated, thus producing a kind of telescopic effect, as works and movements that were centuries apart were absorbed at the same time. Four writers dominate the period from the death of Peter to the ascension

of Catherine II the Great in 1762. Antiokh Kantemir is best known for his verse satires. In addition to his treatises and poems in various genres, Trediakovsky produced a poetic psalter. Lomonosov, who was also a scientist and played a key role in founding Moscow State University (1755), achieved his greatest poetic success in panegyric and spiritual odes, especially "Oda na vzyatiye Khotina" (1739; "Ode on the Seizure of Khotin"), "Vecherneye razmyshleniye o Bozhiyem velichestve" (1743; "Evening Meditation on the Majesty of God"), and "Utrenneye razmyshleniye o Bozhiyem velichestve" (1743; "Morning Meditation on the Majesty of God"). Whereas Baroque poetics strongly influenced Trediakovsky and Lomonosov, the younger Aleksandr Sumarokov, a poet and dramatist, stood for a rigorous and lucid classicism.

Catherine II the Great. Catherine began her reign as an enlightened despot. She corresponded with Voltaire and Denis Diderot and sponsored the arts. Although her native language was German, she has to her credit a number of plays in Russian as well as a statement of legal principles, *Nakaz* (*Instruction*). In 1769 she established a satiric journal, *Vsyakaya vsyachina* ("All Sorts and Sundries"), which was soon followed by others, including the *Truten* ("Drone"), founded by Nikolay Novikov. In a curious exchange between journals, Novikov and Catherine disagreed with each other about the nature of satire—like the Kurbsky-Ivan correspondence in the 16th century, it was a case of a sovereign deigning to argue with a subject. Shocked by an uprising of Cossacks and peasants (1773–75), known from the name of its leader as the Pugachov Rebellion, and later by the French Revolution, Catherine turned increasingly conservative. Generally speaking, these events marked a turning point as the Russian autocracy switched from being a modernizing to a restraining force. When Aleksandr Radishchev published *Puteshestviye iz Peterburga v Moskvu* (1790; *A Journey from St. Petersburg to Moscow*), a work that was sharply critical of Russian society and serfdom, Catherine had him condemned to death, a sentence she commuted to Siberian exile. Offended by a posthumously published play by Yakov Knyazhnin (1742–91), *Vadim Novgorodsky* ("Vadim of Novgorod"), she had copies of the manuscript burned and the published text torn from the offending volume.

Satiric journals

Poetry. Catherine's reign saw real accomplishment in Russian poetry. Excellent verse was produced, and the canon as it is known today began to take shape. It is worth stressing the important role of tradition and the canon in Russian poetry. To a much greater extent than in many other traditions, including the English and American, Russian poetry typically relies on the reader's detailed knowledge of earlier poems. The poems of the past constitute a sort of literary bible, a common culture known in detail by the literate public. Poets count on their readers being sufficiently familiar with the tradition to detect even faint allusions to earlier poems. Moreover, Russian poets also rely on readers to appreciate the semantic associations that specific verse forms have acquired, which is perhaps one reason why free (unrhymed and unmetered) verse has played a relatively small role in Russian poetry.

Three poets—Ivan Khemnitser, Ivan Dmitriyev, and Ivan Krylov—are known for their fables. Krylov's fables rapidly became classics and some of his lines proverbial. *Rossiyada* (written 1771–79; "The Rossiad"), an epic by Mikhail Kheraskov, is a rather stilted effort that proved a literary dead end. It was the ode, rather than the epic, that was the successful high poetic genre of the age. But Vasily Maykov and Ippolit Bogdanovich wrote amusing mock epics. Maykov's *Elisey; ili, razdrazhenny Vakkh* (1769; "Elisei; or, Bacchus Enraged") cleverly parodies a Russian translation of the *Aeneid* with a narrative in which the Greek pantheon directs whores, drunks, and other low-lifes. In *Dushenka: drevnyaya povest v volnykh stikhakh* (1783; "Dushenka: An Ancient Tale in Free Verse"), Bogdanovich produced a light and witty updating of the Greek myth of Cupid and Psyche.

Gavrila Derzhavin

Gavrila Derzhavin is generally considered to be Russia's greatest 18th-century poet. He is best known for his odes, including his chatty panegyric "Oda k Felitse" (1782; "Ode to Felitsa"), in which praise for the prosaic virtues of Empress Catherine alternates with depictions of the low amusements of courtiers. His poems "Bog" (1784; "God") and "Vodopad" (1791–94; "The Waterfall") daringly make the metaphysical concrete and the specific poetic. Derzhavin, who also served as a governor and as Catherine's personal secretary, exemplifies the tendency of 18th-century writers to pursue government careers, a practice that was almost unthinkable a century later.

Drama and prose fiction. Although the theatrical repertoire in the late 18th and early 19th centuries continued to be dominated by translations and adaptations, numerous, if not very distinguished, tragedies were written by Sumarokov, Kheraskov, Vladislav Ozerov, and others. Of greater merit were two comedies by Denis Fonvizin, *Brigadir* (1769; *The Brigadier*), a satire on Gallomania, and *Nedorosl* (1783; "The Minor"). Prose fiction began to appear in print only in the mid-18th century. Mikhail Chulkov's picaresque *Prigozhaya povarikha* (1770; "The Comely Cook") is addressed to a popular audience. At the end of the 18th century, the dominant figure of Russian sentimentalism was Nikolay Karamzin, author of *Pisma russkogo puteshestvennika* (1792; *Letters of a Russian Traveler, 1789–1790*), describing a journey to western Europe in 1789–90, and of the very popular story "Bednaya Liza" (1792; "Poor Liza"), a tale of lovers separated because they belong to different social classes, which seems cloying to the modern reader. Appointed imperial historiographer, Karamzin later wrote the 12-volume *Istoriya gosudarstva rossiyskogo* (1818–26; "History of the Russian State"), which is a landmark of Russian literature. Karamzin's importance also lies in his contribution to the Russian literary language. His writing reflected the language of high society, using a Gallicized vocabulary and syntax at the expense of Church Slavonic.

THE 19TH CENTURY

The Russian 19th century is one of the most fruitful periods in world literature. Several features, in addition to those mentioned above, distinguish the literary culture of these years: (1) Literature enjoyed greater prestige in Russia than in the West, and its achievements were sometimes thought (as Dostoyevsky once declared) to be the justification for the Russian people's very existence. Literary critics were typically the leaders of Russian intellectual life and political thought. (2) Literature and criticism were expected to fulfill functions, such as philosophical, moral, and religious analysis, that in Europe were typically assigned to distinct disciplines. Thus Dostoyevsky's works are central to the histories of all these areas of Russian thought. One can see why the highest achievement of Russian literature was probably the philosophical novel. (3) In the 19th (still more, the 20th) century, politics and literature were intimately connected, and a writer or critic was often called upon to be a political prophet.

The "Golden Age" of poetry. Readers relying on translations usually think of Russian literature almost exclusively in terms of prose, but for Russians their tradition is also, and perhaps equally, one of poetry. The 19th century began with the "Golden Age" of Russian poetry. An aristocratic sensibility, the culture of salons, an aura of friendly intimacy, and genres suitable to this ethos marked the poetry of this period. The romantic poet Vasily Zhukovsky is celebrated for several translations or adaptations that are major poems in their own right, including versions of the English poet Thomas Gray's *An Elegy Written in a Country Church Yard* (1802 and 1839), Homer's *Odyssey* (completed 1847), and Lord Byron's "The Prisoner of Chillon" (1822). His "Svetlana" (1813) reworks the German poet Gottfried August Bürger's "Lenore." Konstantin Batyushkov was noted for playful and erotic as well as melancholy verse and for the elegy *Umerayushchy Tass* (1817; "The Dying Tasso"). The "Pushkin Pleiad," consisting of poets of Pushkin's generation and closely associated with him, included Anton Delvig, Prince Pyotr Vyazemsky, and, most important, Yevgeny Baratynsky, who was a superb philosophical "poet of thought."

Aleksandr Pushkin. Pushkin occupies a unique place in Russian literature. It is not just that Russians view him as their greatest poet; he is also virtually the symbol

of Russian culture. His life, as well as his work, has acquired mythic status. To criticize Pushkin, or even one of his characters—as, for example, Tatyana, the heroine of his novel *Yevgeny Onegin* (written 1823–31; *Eugene Onegin*)—has been taken as something akin to blasphemy. Pushkin's quasi-sacred status has itself been parodied by Russian authors, including the satirist Mikhail Zoshchenko, the absurdist Daniil Kharms, and, most recently, Andrey Sinyavsky in his *Progulki s Pushkinym* (1972; *Strolls with Pushkin*).

Even if one sets this mythic image aside, Pushkin is truly one of the world's most accomplished poets; his verse, however, which relies on the author's perfect control of form, tone, and language, does not read well in translation. Deeply playful and experimental, Pushkin adopted a vast array of conflicting masks and personae. He writes now seriously, now with irony, and now with irony at his own irony, on moral and philosophical themes. He is ultimately a philosophical fox, appreciating the limitations, as well as the virtues, of any set of ideas. A master parodist, Pushkin wrote a number of erotic and at times sacrilegious mock-epics, such as "Gavriiliada" (1821; "The Gabrieliad"), a scabrous retelling of the Annunciation, and *Ruslan i Lyudmila* (1820; *Ruslan and Ludmila*), which, after parodying epic, folk tale, literary ballad, and romance in a spirit of pure play, ends with a startlingly sombre epilogue. His diverse lyrics include a series on the poet's double identity as artist and man of the world; political poems that got him into trouble with the tsar as well as poems in praise of the tsar and his suppression of the Poles; some remarkable elegies ("Vospominaniye" [1828; "Remembrance"]; "Elegiya" [1830; "Elegy"]); love poems, including the "imageless" "Ya vas lyubil" (1829; "I Loved You Once"); and powerful meditations on human evil ("Anchar" [1828; "The Upas Tree"]; "Ne day mne Bog soyti s uma" [1833; "God Grant I Go Not Mad"]).

Pushkin's narrative poems include *Tsygany* (1824; "The Gypsies"), which considers the meaning of freedom. Plot is mere excuse for parody of literary forms and conventions in *Domik v Kolomne* (1830; *The Little House in Kolomna*). Most famously, *Medny vsadnik* (written 1833; *The Bronze Horseman*), which reflects on Peter the Great and the significance of St. Petersburg, examines the meaning of history in relation to individual lives. Concern with the nature of historical causation also led to complex formal innovations in the quasi-Shakespearean drama *Boris Godunov* (written 1824–25), which reworked Karamzin's *Istoriya gosudarstva rossiyskogo* and was in turn reworked by other artists, notably Modest Mussorgsky in his opera *Boris Godunov* (1869; revised 1874). Pushkin's greatest work was probably *Eugene Onegin,* the first truly great Russian novel and the source of countless themes and images in Russian fiction. Playful in the manner of Sterne's *Tristram Shandy* and Lord Byron's *Don Juan,* it far surpasses them for sheer brilliance of form, wit, and thought. Amid endless clever digressions, in which the poet adopts a dazzling array of tones and engages in myriad self-conscious self-parodies, it tells the story of Onegin, a "superfluous man"—that is, a man with no core or purpose to his life—and Tatyana, who stands for authenticity in a sea of literary or social clichés, which she somehow manages to transcend even when she accepts them. The work's serial publication over several years enabled both its own unpredictable creation and changes in the author's perspective to become themes of the poem itself.

A superfluous man

Each of Pushkin's four "little tragedies," all written in 1830, succinctly deals with a philosophical problem. The most remarkable, *Motsart i Salyeri* (*Mozart and Salieri*), based on a legend that Salieri poisoned Mozart, meditates on the nature of creativity while introducing, in brilliantly compressed speeches, what was to be one of the important Russian themes—metaphysical rebellion against God.

After 1830 Pushkin turned to prose. He wrote both a novella and a nonfictional account—*Kapitanskaya dochka* (1836; *The Captain's Daughter*) and *Istoriya Pugachovskogo bunta* (1834; *The History of Pugachev*)—about the same historical events, as if to illustrate that representing the historical truth requires more than one genre. *Povesti pokoynogo Ivana Petrovicha Belkina* (1830; *Tales of the Late Ivan Petrovich Belkin*) filters five narratives, each a parody of a received plot, through the minds of several narrators, collectors, or editors. *Pikovaya dama* (1834; *The Queen of Spades*) offers a suspenseful account of a man seeking mystic knowledge that would enable him to gamble without risk and, implicitly, to know the deepest forbidden truths. Dostoyevsky's *Crime and Punishment* may be seen as an expansion of Pushkin's brief story.

Lermontov and Griboyedov. Next to Pushkin, Mikhail Lermontov, who personifies Romanticism, is probably Russia's most frequently anthologized poet. His celebrated lyrics often recycle lines from his own and others' poems. "Smert poeta" (1837; "Death of a Poet"), which first earned him fame, deals with Pushkin's death shortly after a fatal duel in 1837. Among his narrative poems, *Demon* (1841; *The Demon*) describes the love of a Byronic demon for a mortal woman; *Pesnya pro tsarya Ivana Vasilyevicha, molodogo oprichnika i udalogo kuptsa Kalashnikova* (1837; *A Song About Tsar Ivan Vasilyevich, His Young Bodyguard, and the Valiant Merchant Kalashnikov*) is a stylized folk epic. Also an accomplished prose stylist, Lermontov wrote *Geroy nashego vremeni* (1840; *A Hero of Our Time*), which in form is something between a novel and a complexly framed cycle of stories about a single hero, a Byronic superfluous man. This work ranges from sketches of philosophical brilliance ("The Fatalist") to episodes of near puerility ("Princess Mary"). The theme of the superfluous man finds another important rendition in Aleksandr Griboyedov's classic work, *Gore ot uma* (completed 1824; *Woe from Wit*).

Nikolay Gogol. One of the finest comic authors of world literature, and perhaps its most accomplished nonsense writer, Gogol is best known for his short stories, for his play *Revizor* (1836; *The Inspector General,* or *The Government Inspector*), and for *Myortvye dushi* (1842; *Dead Souls*), a prose narrative that is nevertheless subtitled a "poem." "Nos" (1836; "The Nose"), a parable on the failure of all explanatory systems, relates an utterly inexplicable incident and the attempts to come to terms with it. Both "Shinel" (1842; "The Overcoat"), which is probably the most influential Russian short story, and "Zapiski sumasshedshego" (1835; "The Diary of a Madman") mix pathos and mockery in an amazing display. As in "Nevsky prospekt" (1835; "Nevsky Avenue") and "Povest o tom, kak possorilsya Ivan Ivanovich s Ivanom Nikiforovichem" (1835; "The Tale of How Ivan Ivanovich Quarrelled with Ivan Nikiforovich"), language itself seems to generate its own absurd content while the universe turns out to be a counterfeit of which there is no original. Characteristic of Gogol is a sense of boundless superfluity that is soon revealed as utter emptiness and a rich comedy that suddenly turns into metaphysical horror. *The Inspector General* develops a sequence of (witting and unwitting) confidence games within confidence games in a corrupt world of endless self-deception. The mock-epic, even mock-satiric, *Dead Souls* simultaneously allegorizes the timeless bureaucratic tendency to make official documentation more genuine than actual existence, the emptiness of the human soul, and the mind's absurd ways of grasping meaning or value. It is one of the most striking (and most Gogolian) ironies of Russian literary history that radical critics celebrated Gogol as a realist.

Other poets and dramatists. From the death of Lermontov until the end of the 19th century, Russian literature was dominated by prose, but some poets of lasting interest appeared. Fyodor Tyutchev, a member of Pushkin's generation, wrote nature, love, and political poetry but is probably best appreciated for his philosophical "poetry of thought," including "Silentium!" (1830). Afanasy Fet wrote delicate love lyrics remarkable for their absence of verbs. Violently attacked by radical critics as symbolizing pure art, he came to be appreciated by the Symbolist poets to follow. Nikolay Nekrasov, who was also a major figure in Russian journalism, wrote social satires, tendentious "civic" verse, and compassionate accounts of peasant life, including *Komu na Rusi zhit khorosho?* (1879; *Who Can Be Happy and Free in Russia?*), which he began writing in 1863 and left unfinished at the time of his death.

Among the dramatists of this period, Aleksandr Ostrov-

sky, who has proved much more popular in Russia than abroad, wrote many slice-of-life plays about the Russian merchantry. His plays *Svoi lyudi—sochtyomsya!* (1850; "It's a Family Affair—We'll Settle It Among Ourselves"; Eng. trans. *A Family Affair*) and *Groza* (1859; *The Thunderstorm*) were the subject of reviews by Nikolay Dobrolyubov (1836–61), one of Russia's most influential radical critics. Aleksandr Sukhovo-Kobylin wrote a macabre trilogy, whose third play, *Smert Tarelkina* (1869; *The Death of Tarelkin*), is a brilliant piece of grotesque humour about a man who fakes his own death. The theme of the faked suicide, a motif of Russian drama, later appeared in Leo Tolstoy's *Zhivoy trup* (written 1900; *The Living Corpse*) and Nikolay Erdman's *Samoubiytsa* (1928; *The Suicide*).

The "intelligentsia." Beginning about 1860, Russian culture was dominated by a group known as the "intelligentsia," a word that English borrowed from Russian but which means something rather different in its original Russian usage. In the word's narrow sense, the "intelligentsia" consisted of people who owed their primary allegiance not to their profession or class but to a group of men and women with whom they shared a common set of beliefs, including a fanatic faith in revolution, atheism, and materialism. They usually adopted a specific set of manners, customs, and sexual behaviour, primarily from their favourite book, Nikolay Chernyshevsky's utopian novel *Chto delat* (1863; *What Is to Be Done?*). Although appallingly bad from a literary point of view, this novel, which also features a fake suicide, was probably the most widely read work of the 19th century.

Generally speaking, the intelligentsia insisted that literature be a form of socialist propaganda and rejected aesthetic criteria or apolitical works. In addition to Chernyshevsky and Dobrolyubov, typical members of the intelligentsia came to include Lenin, Stalin, and other Bolsheviks who seized power in 1917. Thus it is not surprising that a gulf separated the writers from the intelligentsia. In an important anthology attacking the mentality of the intelligentsia, *Vekhi* (1909; *Landmarks*), the critic Mikhail Gershenzon observed that "an almost infallible gauge of the strengths of an artist's genius is the extent of his hatred for the intelligentsia." Typically, the writers objected to the intelligentsia's intellectual intolerance, addiction to theory, and belief that morality was defined by utility to the revolution. Tolstoy, Dostoyevsky, and Anton Chekhov were all sharply contemptuous of the intelligentsia.

Fyodor Dostoyevsky. Having been imprisoned in Siberia for political activity, Dostoyevsky parodied *What Is to Be Done?* in *Zapiski iz podpolya* (1864; *Notes from the Underground*), a novella that has had incalculable influence on Western literature for formal as well as thematic reasons. In a complex series of paradoxes, its hero argues against determinism, utopianism, and historical laws. In *Prestupleniye i nakazaniye* (1866; *Crime and Punishment*), a philosophical and psychological account of a murder, Dostoyevsky examines the tendency of *intelligents* (members of the intelligentsia) to regard themselves as superior to ordinary people and as beyond traditional morality. *Besy* (1872; *The Possessed*), a novel based on Russian terrorism, is famous as the work that most accurately predicted 20th-century totalitarianism. In *Idiot* (1868–69; *The Idiot*) and *Bratya Karamazovy* (1879–80; *The Brothers Karamazov*), Dostoyevsky, who is generally regarded as one of the supreme psychologists in world literature, sought to demonstrate the compatibility of Christianity with the deepest truths of the psyche.

Leo Tolstoy. Probably even more than Dostoyevsky, Tolstoy has been praised as being the greatest novelist in world literature. The 19th-century English critic and poet Matthew Arnold famously expressed the commonest view in saying that a work by Tolstoy is not a piece of art but a piece of life: his novels read as if life were writing directly, without mediation. Tolstoy's techniques reflect his belief that no theory is adequate to explain the world's complexity, which unfolds by "tiny, tiny alterations" fitting no pattern. He denied the existence of historical laws and insisted that ethics is a matter not of rules but of supreme sensitivity to the particular. "True life," he contended, is lived not at moments of grand crisis but at countless ordinary and prosaic moments, which human beings usually do not notice. All these ideas are illustrated and explicitly expressed in *Voyna i mir* (1865–69; *War and Peace*), set in the time of the Napoleonic wars, and in *Anna Karenina* (1875–77), which applies this prosaic view of life to marriage, the family, and work. *Anna Karenina* also contrasts romantic love, which is based on intense moments of passion and leads to adultery, with the prosaic love of the family, which is based above all on intimacy.

After completing *Anna Karenina,* Tolstoy underwent a religious crisis, which eventually led him to reject his two great novels, formulate a new religion that he thought of as true Christianity, and cultivate a different type of art. To outline his views, he wrote a number of tracts, including *Tsarstvo bozhiye vnutri vas* (1893; *The Kingdom of God Is Within You*) and *Chto takoye iskusstvo?* (1898; *What Is Art?*). His only long novel of this period, *Voskreseniye* (1899; *Resurrection*), is a tendentious failure. But he produced brilliant novellas, many of which were published posthumously, including *Otets Sergy* (written 1898; *Father Sergius*), in which he seems to reflect on his own quest for sainthood, and *Khadzhi-Murat* (written 1904; *Hadji-Murad*). *Smert Ivana Ilicha* (1886; *The Death of Ivan Ilyitch*), which is often considered the greatest novella in Russian literature, conveys the existential horror of sickness and mortality while describing civilization as a web of lies designed to distract people from an awareness of death.

Ivan Turgenev. The first Russian writer to be widely celebrated in the West, Turgenev managed to be hated by the radicals as well as by Tolstoy and Dostoyevsky for his dedicated Westernism, bland liberalism, aesthetic elegance, and tendency to nostalgia and self-pity. He first gained fame with his subtle descriptions of peasant life in *Zapiski okhotnika* (1852; *A Sportsman's Sketches*), which contributed to the climate leading to the abolition of serfdom. He is celebrated for his novels about *intelligents* and ideology: *Rudin* (1856), *Nakanune* (1860; *On the Eve*), and *Dym* (1867; *Smoke*). His most distinguished work, *Ottsy i deti* (1862; *Fathers and Sons*), offers both an even-handed portrait of the radical nihilists and an allegorical meditation on the conflict of generations.

Other prose writers. The mid-19th century produced a number of other fine prose writers. Sergey Aksakov wrote fictionalized reminiscences: *Semeynaya khronika* (1856; *The Family Chronicle*) and *Detskiye gody Bagrova-vnuka* (1858; *Years of Childhood*). Aleksandr Herzen wrote his greatest works in emigration. In *S togo berega* (written 1847–50; *From the Other Shore*), which combines essays and dialogues, he reflects with penetrating skepticism on the idea that history has knowable laws. Herzen's *Byloye i dumy* (written 1852–68; *My Past and Thoughts*) is regarded as the best Russian autobiography. Ivan Goncharov is the author of the comic masterpiece *Oblomov* (1859), a study of dreamy slothfulness: its hero spends a hundred pages getting out of bed. Nikolay Leskov is remembered for his short stories, including "Ledi Makbet Mtsenskogo uyezda" (1865; "Lady Macbeth of the Mtsensk District"), as well as for his novel *Soboryane* (1872; *The Cathedral Folk*). Like Gogol before and Mikhail Zoshchenko after him, he was a master of *skaz,* a written narrative imitating a spontaneous oral account in its use of dialect, slang, or a particular idiom. A radical satirist and (remarkably) a government official who attained general's rank, Mikhail Saltykov wrote (under the pseudonym N. Shchedrin) the extremely dark novel *Gospoda Golovlyovy* (1876; *The Golovlyov Family*), portraying the relentless decline of a family. The agony of an intellectual who wants to merge with the common people and the intimate link of utopianism to madness figure as prominent themes in the short stories of Vsevolod Garshin, including "Khudozhniki" (1879; "Artists") and "Krasny tsvetok" (1883; "The Red Flower").

Anton Chekhov. When Tolstoy abandoned the prosaic ethos, Chekhov, one of the greatest short story writers in world literature, remained loyal to it. Indeed, he reinterpreted it within his essentially bourgeois values, stressing the moral necessity of ordinary virtues such as daily

kindness, cleanliness, politeness, work, sobriety, paying one's debts, and avoiding self-pity. Replying to the intelligentsia's demand for political tendentiousness, which he equated with a stifling intellectual conformity, he maintained that his only "tendency" was a protest against lying in all its forms. In his hundreds of stories and novellas, which he wrote while practicing medicine, Chekhov adopts something of a clinical approach to ordinary life. Meticulous observation and broad sympathy for diverse points of view shape his fiction. In his stories, an overt plot subtly hints at other hidden stories, and so the experience of rereading his fiction often differs substantially from that produced by a first reading. Especially noteworthy are "Skuchnaya istoriya" (written 1889; "A Dreary Story"), "Duel" (written 1891; "The Duel"), "Palata No. 6" (written 1892; "Ward Number Six"), "Kryzhovnik" (written 1898; "Gooseberries"), "Dushechka" (written 1899; "The Darling"), "Dama s sobachkoy" (written 1899; "The Lady with the Lap Dog"), "Arkhiyerey" (written 1902; "The Bishop"), and "Nevesta" (written 1903; "The Betrothed").

Chekhov's plays

Along with Gogol's *The Inspector General,* Chekhov's plays are the high point of Russian drama. In his four great plays, *Chayka* (1896; *The Seagull*), *Dyadya Vanya* (1897; *Uncle Vanya*), *Tri sestry* (1901; *Three Sisters*), and *Vishnyovy sad* (1904; *The Cherry Orchard*), Chekhov's belief that life is lived at ordinary moments and that histrionics are a dangerous lie found expression in a major innovation, the undramatic drama—or, as it is sometimes called, the theatre of inaction.

THE SILVER AGE

The period from the 1890s to 1917 was one of intellectual ferment, in which mysticism, aestheticism, Neo-Kantianism, eroticism, Marxism, apocalypticism, Nietzscheanism, and other movements combined with each other in improbable ways. Primarily an age of poetry, it also produced significant prose and drama. Russian Symbolism, which was influenced by French Symbolist poetry and the philosophy of Vladimir Solovyov (1853–1900), is usually said to have begun with an essay by Dmitry Merezhkovsky, "O prichinakh upadka i o novykh techeniyakh sovremennoy russkoy literatury" (1893; "On the Reasons for the Decline and on the New Trends in Contemporary Russian Literature"). A poet and propagator of religious ideas, Merezhkovsky wrote a trilogy of novels, *Khristos i Antikhrist* (1896–1905; *Christ and Antichrist*), consisting of *Yulian otstupnik* (1896; *Julian the Apostate*), *Leonardo da Vinchi* (1901; *Leonardo da Vinci*), and *Pyotr i Aleksey* (1905; *Peter and Alexis*), which explores the relation of pagan and Christian views of the world.

Symbolists. The Symbolists saw art as a way to approach a higher reality. The first wave of Symbolists included Konstantin Balmont (1867–1942), who translated a number of English poets and wrote verse that he left unrevised on principle (he believed in first inspiration); Valery Bryusov (1873–1924), a poet and translator of French Symbolist verse and of Virgil's *Aeneid,* who for years was the leader of the movement; Zinaida Gippius (1869–1945), who wrote decadent, erotic, and religious poetry; and Fyodor Sologub, author of melancholic verse and of a novel, *Melky bes* (1907; *The Petty Demon*), about a sadistic, homicidal, paranoid schoolteacher.

Three writers dominate the second wave of Symbolism. Eschatology and anthroposophy shaped the poetry and prose of Andrey Bely, whose novel *Peterburg* (1913–22; *St. Petersburg*) is regarded as the masterpiece of Symbolist fiction. Aleksandr Blok, who wrote the lyric drama *Balaganchik* (1906; "The Showbooth"), is best known for his poem *Dvenadtsat* (1918; *The Twelve*), which describes 12 brutal Red Guards who turn out to be unwittingly led by Jesus Christ. The principal theoretician of the Symbolist movement, Vyacheslav Ivanov (1866–1949), wrote mythic poetry conveying a Neoplatonist philosophy.

Acmeists and Futurists. In the second decade of the 20th century, Symbolism was challenged by two other schools, the Acmeists, who favoured clarity over metaphysical vagueness, and the brash Futurists, who wanted to throw all earlier and most contemporary poetry "from the steamship of modernity." Among the Acmeists, Nikolay Gumilyov (1886–1921), who stressed poetic craftsmanship over the occult, was executed by the Bolsheviks. Already an accomplished creator of superb love lyrics in these years, Anna Akhmatova produced densely and brilliantly structured poems in the Soviet period, including *Poema bez geroya* (written 1940–62; *A Poem Without a Hero*) and *Rekviyem* (written 1935–40; *Requiem*), which was inspired by Soviet purges and was therefore unpublishable in Russia. From 1923 to 1940 she was forced into silence, and in 1946 Akhmatova and Zoshchenko became the target of official abuse by the Communist Party cultural spokesman Andrey Zhdanov (1896–1948). Some consider Osip Mandelshtam (1891–1938), who died in a Soviet prison camp, to be the greatest Russian poet of the 20th century. Many of his difficult, allusive poems were preserved by his wife, Nadezhda Mandelshtam (1899–1980), whose memoirs are themselves a classic.

Anna Akhmatova

The two most important Futurist poets were Velimir Khlebnikov and Vladimir Mayakovsky. Khlebnikov hoped to find the laws of history through numerology and developed amazingly implausible theories about language and its origins. His verse, which is characterized by neologisms and "trans-sense" language, includes "Zaklyatiye smekhom" (1910; "Incantation by Laughter") and *Zangezi* (1922). Mayakovsky epitomized the spirit of romantic bohemian radicalism. Humour, bravado, and self-pity characterize his inventive long poems, including *Oblako v shtanakh* (1915; *A Cloud in Trousers*). After the Russian Revolution in 1917, which he ardently supported initially, Mayakovsky "stepped on the throat" of his song to produce propaganda poems. But he also satirized Soviet bureaucracy in the witty "Razgovor s fininspektorom o poezii" (1926; "Conversation with a Tax Collector about Poetry"). As a dramatist, he is best known for *Vladimir Mayakovsky* (1913), in which he played the lead role, and *Klop* (1929; *The Bedbug*), in which a philistine, along with a bedbug, is resurrected into the banal communist future of 1979. Having written a poem about the suicide of the peasant poet Sergey Yesenin (1895–1925), Mayakovsky later shot himself, leaving a brilliantly ironic suicide note with a poem explaining that "love's boat has smashed against daily life."

Others. Celebrated in their day, the fiction writers Leonid Andreyev (1871–1919), Aleksandr Kuprin (1870–1938), and Vladimir Korolenko (1853–1921) now have faded reputations. But Ivan Bunin, who became the first Russian to win the Nobel Prize for Literature (1933), wrote superb works both before the Revolution and as an émigré after it. Especially noteworthy are his dark novella *Derevnya* (1910; *The Village*), which is relentlessly critical of Russians, and his *Zhizn Arsenyeva* (1930; *The Life of Arseniev,* or *The Well of Days*), a fictionalized autobiography. Maksim Gorky became the official founder of Socialist Realism. Western readers now appreciate his three-volume autobiography *Detstvo* (1913–14; *My Childhood*), *V lyudyakh* (1915–16; *In the World,* or *My Apprenticeship*), and *Moi universitety* (1923; *My Universities*) and his *Vospominaniya o Lve Nikolayeviche Tolstom* (1919; *Reminiscences of Leo Nikolaevich Tolstoy*). His highly tendentious novel *Mat* (1906; *Mother*), a model for Socialist Realism, and many other works divide characters simplistically into two groups—progressive and virtuous or reactionary and vicious.

Post-Revolutionary literature

LITERATURE UNDER SOVIET RULE

The Bolshevik seizure of power in 1917 radically changed Russian literature. After a brief period of relative openness (compared to what followed) in the 1920s, literature became a tool of state propaganda. Officially approved writing (the only kind that could be published) by and large sank to a subliterary level. Censorship, imprisonment in labour camps, and mass terror were only part of the problem. Writers were not only forbidden to create works that were dissident, formally complex, or objective (a term of reproach), but they were also expected to fulfill the dictates of the Communist Party to produce propaganda on specific, often rather narrow, themes of current interest

to it. Writers were called upon to be "engineers of human souls" helping to produce "the new Soviet man."

The fragmentation of the literary tradition

As a result of Bolshevik rule, the literary tradition was fragmented. In addition to official Soviet Russian literature, two kinds of unofficial literature existed. First, a tradition of émigré literature, containing some of the best works of the century, continued until the fall of the Soviet Union. Second, unofficial literature written within the Soviet Union came to include works circulated illegally in typewritten copies ("samizdat"), works smuggled abroad for publication ("tamizdat"), and works written "for the drawer," or not published until decades after they were written ("delayed" literature). Moreover, literature publishable at one time often lost favour later; although nominally acceptable, it was frequently unobtainable. On many occasions, even officially celebrated works had to be rewritten to suit a shift in the Communist Party line. Whereas pre-Revolutionary writers had been intensely aware of Western trends, for much of the Soviet period access to Western movements was severely restricted, as was foreign travel. Access to pre-Revolutionary Russian writing was also spotty. As a result, Russians periodically had to change their sense of the past, as did Western scholars when "delayed" works became known.

From a literary point of view, unofficial literature clearly surpasses official literature. Of Russia's five winners of the Nobel Prize for Literature during the Soviet period, Bunin emigrated after the Revolution, Boris Pasternak had his novel *Doctor Zhivago* (1957) published abroad, Aleksandr Solzhenitsyn (b. 1918) had most of his works published abroad and was expelled from the Soviet Union, and Joseph Brodsky (1940–96) published all his collections of verse abroad and was forced to emigrate in 1972. Only Mikhail Sholokhov (1905–84) was clearly an official Soviet writer. In the early years following the Revolution, writers who left or were expelled from the Soviet Union included Balmont, Bunin, Gippius, Vyacheslav Ivanov, Kuprin, and Merezhkovsky. Émigrés also included the poets Vladislav Khodasevich (1886–1939) and Georgy Ivanov (1894–1958). Marina Tsvetayeva (1892–1941), regarded as one of the great poets of the 20th century, eventually returned to Russia, where she committed suicide. Vladimir Nabokov, who later wrote in English, published nine novels in Russian, including *Dar* (published serially 1937–38; *The Gift*) and *Priglasheniye na kazn* (1938; *Invitation to a Beheading*).

FROM THE 1920S TO C. 1985

Experiments in the 1920s. Within Russia the 1920s saw a wide diversity of literary trends and works, including those by mere "fellow travelers" (Leon Trotsky's phrase) of the Revolution. Isaak Babel wrote a brilliant cycle of linked stories, collected as *Konarmiya* (1926; *Red Cavalry*), about a Jewish commissar in a Cossack regiment. Formally chiseled and morally complex, these stories examine the seductive appeal of violence for the intellectual.

Dystopia

A modern literary genre, the dystopia, was invented by Yevgeny Zamyatin in his novel *My* (1924; *We*), which could be published only abroad. Like Aldous Huxley's *Brave New World* and George Orwell's *Nineteen Eighty-four,* which are modeled on it, *We* describes a future socialist society that has turned out to be not perfect but inhuman. Yury Olesha's *Zavist* (1927; *Envy*) is a satire in the tradition of *Notes from the Underground.* Like Chekhov, Zoshchenko was a master of the comic story focusing on everyday life. Pasternak, who had been a Futurist poet before the Revolution, published a cycle of poems, *Sestra moya zhizn* (1922; *My Sister—Life*), and his story "Detstvo Lyuvers" (1918; "Zhenya Luvers's Childhood"). Other important novels include Boris Pilnyak's "ornamental" *Goly god* (1922; *The Naked Year*); Andrey Platonov's deeply pessimistic *Kotlovan* (*The Foundation Pit*), which was written in the late 1920s and published in the West in 1973; Ilya Ilf and Yevgeny Petrov's clever satire *Dvenadtsat stulyev* (1928; *The Twelve Chairs*); Konstantin Fedin's novel *Goroda i gody* (1924; *Cities and Years*); and Leonid Leonov's *Vor* (1927; *The Thief*).

The Russian Formalists were a school of critics closely tied to the Futurists. They developed a vibrant, comprehensive theory of literature and culture that inspired structuralism, an influential critical movement in the West. Two of them, Viktor Shklovsky and Yury Tynyanov, wrote significant fiction illustrating their theories: Shklovsky's *Zoo; ili, pisma ne o lyubvi* (1923; *Zoo; or, Letters Not About Love*) and Tynyanov's "Podporuchik kizhe" (1927; "Second Lieutenant Likewise"). Their respectful opponent, Mikhail Bakhtin, whom some consider the most original, far-ranging, and subtle theorist of literature in the 20th century, wrote *Problemy tvorchestva Dostoyevskogo* (1929, 2nd ed., 1963; *Problems of Dostoevsky's Poetics*) and essays about the relation of novelistic form to time, language, psychology, and ethics. The 1920s also produced novels that became classics of official Soviet literature, including Dmitry Furmanov's *Chapayev* (1923) and Aleksandr Serafimovich's *Zhelezny potok* (1924; *The Iron Flood*). Fyodor Gladkov's *Tsement* (1925; *Cement*) became a model for the "industrial production" novel. Also in this period, Sholokhov began writing the best-known official work, a four-part novel published as *Tikhy Don* (1928–40; "The Quiet Don"; translated in two parts as *And Quiet Flows the Don* and *The Don Flows Home to the Sea*).

The Stalin era. The decade beginning with Stalin's ascendancy in the late 1920s was one of unprecedented repression. The "war in the countryside" to enforce the collectivization of agriculture cost more than 10 million lives, about half of them by starvation. Purges took the lives of millions more, among them Babel, Kharms, Mandelshtam, Pilnyak, the peasant poet Nikolay Klyuyev (1887–1937), and the director Vsevolod Meyerhold (1874–1940). In 1932 all independent literary groupings were dissolved and replaced by an institution that had no counterpart in the West, the Union of Soviet Writers.

The Union of Soviet Writers

The union became the state's instrument of control over literature, and expulsion from it meant literary death. In 1934 Socialist Realism was proclaimed the only acceptable form of writing. Henceforth, literature was to be governed by a series of official directives regarding details of style and content in order to ensure that each work offered a "truthful" depiction "of reality in its revolutionary development." Literature had to be "party-minded" and "typical" (that is, avoiding unpleasant, hence "atypical," aspects of Soviet reality), while showing the triumph of fully "positive heroes."

Some talented writers turned to the safer areas of children's literature and translation. Others, such as Valentin Katayev in his production novel *Vremya, vperyod!* (1932; *Time, Forward!*) and Fedin in *Pervyye radosti* (1946; *Early Joys*), sought to infuse official writing with some interest. Quite popular was Nikolay Ostrovsky's fictionalized autobiography *Kak zakalyalas stal* (1932–34; *How the Steel Was Tempered*). In his unfinished novel *Pyotr Pervy* (1929–45; *Peter the Great*) and his play *Ivan Grozny* (1941–43; "Ivan the Terrible"), Aleksey Tolstoy, an émigré who returned to become one of Stalin's favourite writers, praised tyrannical tsars admired by Stalin. The moral nadir of Soviet literature was reached in a collaborative volume, *Belomorsko-Baltiski kanal imeni Stalina: istoriya stroitelstva* (1934; *Belomor: An Account of the Construction of the New Canal Between the White Sea and the Baltic Sea*). With Gorky as an editor and 34 contributors, including Gorky, Katayev, Shklovsky, Aleksey Tolstoy, and Zoshchenko, the volume praised a project (and the secret police who directed it) using convict labour and costing tens of thousands of lives. During these dark years the work now generally regarded as the finest post-Revolutionary novel, Mikhail Bulgakov's *Master i Margarita* (*The Master and Margarita*), was written "for the drawer" (1928–40); it appeared (expurgated) in Russia only in 1966–67 and unexpurgated in 1973. It tells of the Devil and his retinue visiting Soviet Russia, where they play practical jokes of metaphysical and political significance. A novel within the novel gives the "true" version of Christ's encounter with Pilate. The result is a joyful philosophical comedy of enormous profundity.

The need to rally support in World War II brought a loosening of Communist Party control. The war itself created the opportunity for a large "second wave" of emigration, thus feeding émigré literature. The period from 1946 until

the death of Stalin in 1953 was one of severe repression known as the *zhdanovshchina,* or Zhdanovism. During this campaign, attacks on "rootless cosmopolitans" involved anti-Semitism and the rejection of all foreign influences on Russian literature. The Soviet practice of *samokritika* (public denunciation of one's own work) was frequent.

Thaws and freezes. The years from the death of Stalin until the fall of Nikita Khrushchev in 1964 saw several "thaws" separated by "freezes." Ilya Ehrenburg's novel *Ottepel* (1954; *The Thaw*) provided this term for a period of relative liberalism. In 1956 Khrushchev delivered a famous speech denouncing certain Stalinist crimes. From that time on, it was possible for Russians to perceive orthodox communists as people of the past and to regard dissidents not as holdovers from before the Revolution but as progressives. The harsher years under Leonid Brezhnev following Khrushchev's fall opened with the arrest, trial, and imprisonment of two writers, Andrey Sinyavsky (whose pseudonym was Abram Terts) and Yuly Daniel (pseudonym Nikolay Arzhak), for publishing "anti-Soviet propaganda" abroad. In the years that followed, well-known writers were arrested or, in one way or another, expelled from the Soviet Union, thus generating the third wave of émigré literature. Among those who found themselves in the West were Brodsky, Sinyavsky, Solzhenitsyn, Vasily Aksyonov, Georgy Vladimov, Vladimir Voynovich, and Aleksandr Zinovyev.

Third wave of émigré literature

Significant literary works written in the post-Stalin years include Pasternak's poetic novel set at the time of the Revolution, *Doctor Zhivago* (first published in Italy in 1957), which sees life's meaning as transcending politics. Sinyavsky's book-length essay *Chto takoye sotsialisticheskiy realizm?* (1956; *On Socialist Realism*), attacking Socialist Realist aesthetic doctrine and advocating the use of fantasy, and a number of "phantasmagoric works," including *Lyubimov* (1961–62; *The Makepeace Experiment*), were published abroad. Charged with being the author of these works, Sinyavsky was tried and imprisoned in 1966. Some have considered the transcripts of his trial to be one of his most interesting "works." After his emigration to France in 1973 he published the novel *Spokoynoy nochi* (1984; *Goodnight!*) under the name Terts and *Osnovy sovetskoy tsivilizatsii* (1989; *Soviet Civilization: A Cultural History*) under the name Sinyavsky.

A movement called "village prose" cultivated nostalgic descriptions of rural life. Particularly noteworthy is Valentin Rasputin's elegiac novel *Proshchaniye s Matyoroy* (1976; *Farewell to Matyora*) about a village faced with destruction to make room for a hydroelectric plant. The novel's regret for the past and suspicion of the new dramatically marks the difference between village prose and the Socialist-Realist collective farm novel. Yury Trifonov wrote about what he called "the ordeal of ordinary life" in *Dom na naberezhnoy* (1976; *The House on the Embankment*) and *Starik* (1978; *The Old Man*). Lyudmila Petrushevskaya's plays portray family life; her collection of stories *Bessmertnaya lyubov* (1988; *Immortal Love*) could be published only under Mikhail Gorbachev. Works first published in full in the West and in fundamental ways critical of Soviet ideology and culture include Andrey Bitov's experimental novel *Pushkinsky dom* (1978; *Pushkin House*), Venedikt Yerofeyev's alcoholic, hallucinatory novel *Moskva-Petushki* (1977; *Moscow to the End of the Line*), Zinovyev's *Ziyayushchiye vysoty* (1976; *The Yawning Heights*), and Voynovich's satire *Zhizn i neobychaynyye priklyucheniya soldata Ivana Chonkina* (1975; *The Life and Extraordinary Adventures of Private Ivan Chonkin*).

Solzhenitsyn first earned fame with *Odin den Ivana Denisovicha* (1963; *One Day in the Life of Ivan Denisovich*), an understated novel about the horrors of a Soviet camp. As part of his de-Stalinization campaign, Khrushchev personally saw to its publication. Under Brezhnev, Solzhenitsyn was expelled from the U.S.S.R. Solzhenitsyn's *Arkhipelag GULag, 1918–1956: opyt khudozhestvennogo issledovaniya,* 3 vol. (1973–75; *The Gulag Archipelago, 1918–1956: An Experiment in Literary Investigation*) is arguably the greatest work of Soviet prose. It narrates the history of the Soviet camp system with controlled fury and in an ironic mode reminiscent of the 18th-century English historian Edward Gibbon.

Solzhenitsyn

Post-Soviet literature

Almost no one expected the Soviet Union to come suddenly to an end. The effects of this event on literature have been enormous. The period of glasnost (verbal openness) under Gorbachev and the subsequent collapse of the U.S.S.R. led first to a dramatic easing and then to the abolition of censorship. Citizenship was restored to émigré writers, and Solzhenitsyn returned to Russia. *Doctor Zhivago* and *We* were published in Russia, as were the works of Nabokov, Solzhenitsyn, Voynovich, and many others. The divisions between Soviet and émigré and between official and unofficial literature came to an end. Russians experienced the heady feeling that came with absorbing, at great speed, large parts of their literary tradition that had been suppressed and with having free access to Western literary movements. A Russian form of postmodernism, fascinated with a pastiche of citations, arose, along with various forms of radical experimentalism. During this period, readers and writers sought to understand the past, both literary and historical, and to comprehend the chaotic, threatening, and very different present.

BIBLIOGRAPHY. General histories and handbooks of Russian literature include D.S. MIRSKY, *A History of Russian Literature, from Its Beginnings to 1900,* ed. by FRANCIS J. WHITFIELD (1958); VICTOR TERRAS (ed.), *Handbook of Russian Literature* (1985); VICTOR TERRAS, *A History of Russian Literature* (1991); and CHARLES A. MOSER (ed.), *The Cambridge History of Russian Literature,* rev. ed. (1992).

Histories of specific periods or genres are found in the following texts: DMITRIJ ČIŽEVSKIJ (DIMITRIJ TSCHIŽEWSKIJ), *History of Russian Literature from the Eleventh Century to the End of the Baroque* (1960, reprinted 1981); N.K. GUDZY (N.K. GUDZIĬ), *History of Early Russian Literature,* ed. by SUSAN WILBUR JONES (1949, reissued 1970; originally published in Russian, 2nd ed., 1941); SIMON KARLINSKY, *Russian Drama from Its Beginnings to the Age of Pushkin* (1985); WILLIAM EDWARD BROWN, *A History of Seventeenth-Century Russian Literature* (1980), *A History of 18th-Century Russian Literature* (1980), and *A History of Russian Literature of the Romantic Period,* 4 vol. (1986); WILLIAM MILLS TODD III, *Fiction and Society in the Age of Pushkin* (1986); WILLIAM MILLS TODD III (ed.), *Literature and Society in Imperial Russia, 1800–1914* (1978); MARC SLONIM, *The Epic of Russian Literature from Its Origins Through Tolstoy* (1950, reissued 1975), and *Modern Russian Literature from Chekhov to the Present* (1953); VLADIMIR MARKOV, *Russian Futurism: A History* (1968); EDWARD J. BROWN, *Russian Literature Since the Revolution,* rev. and enlarged ed. (1982); GLEB STRUVE, *Russian Literature Under Lenin and Stalin, 1917–1953* (1971); BORIS GROYS (BORIS GROÎS), *The Total Art of Stalinism: Avant-Garde, Aesthetic Dictatorship, and Beyond* (1992; originally published in German, 1988); JOHN GARRARD and CAROL GARRARD, *Inside the Soviet Writers' Union* (1990); VICTOR ERLICH, *Russian Formalism: History, Doctrine,* 3rd ed. (1981); KATERINA CLARK, *The Soviet Novel,* 2nd ed. (1985); KATHLEEN F. PARTHÉ, *Russian Village Prose: The Radiant Past* (1992); and HAROLD B. SEGEL, *Twentieth-Century Russian Drama: From Gorky to the Present,* updated ed. (1993).

Studies providing interesting general ideas about the tradition include ROBERT LOUIS JACKSON, *Dostoevsky's Underground Man in Russian Literature* (1958, reprinted 1981); RUFUS W. MATHEWSON, JR., *The Positive Hero in Russian Literature,* 2nd ed. (1975); CARYL EMERSON, *Boris Godunov: Transpositions of a Russian Theme* (1986); GARY SAUL MORSON (ed.), *Literature and History: Theoretical Problems and Russian Case Studies* (1986); and ANDREW BARUCH WACHTEL, *An Obsession with History: Russian Writers Confront the Past* (1994).

Among the useful anthologies of Russian literature are SERGE A. ZENKOVSKY (ed. and trans.), *Medieval Russia's Epics, Chronicles, and Tales,* rev. and enlarged ed. (1974), which contains many useful commentaries; HAROLD B. SEGEL (ed. and trans.), *The Literature of Eighteenth-Century Russia,* 2 vol. (1967), with an excellent history in vol. 1 and useful commentaries throughout; GEORGE GIBIAN (ed.), *The Portable Nineteenth-Century Russian Reader* (1993); CLARENCE BROWN (ed.), *The Portable Twentieth-Century Russian Reader,* rev. and updated ed. (1993); HELENA GOSCILO and BYRON LINDSEY (eds.), *Glasnost: An Anthology of Russian Literature Under Gorbachev* (1990); HELENA GOSCILO (ed.), *Balancing Acts: Contemporary Stories by Russian Women* (1989); and DIMITRI OBOLENSKY (ed.), *The Penguin Book of Russian Verse,* rev. ed. (1965, reprinted as *The Heritage of Russian Verse,* 1976). (G.S.M.)

Rutherford

Ernest Rutherford, Baron Rutherford of Nelson, nuclear physicist and Nobel Prize winner, is to be ranked in fame with Sir Isaac Newton and Michael Faraday. Indeed, just as Faraday is called the "father of electricity," so a similar description might be applied to Rutherford in relation to nuclear energy. He contributed substantially to the understanding of the disintegration and transmutation of the radioactive elements, discovered and named the particles expelled from radium, identified the alpha particle as a helium atom and with its aid evolved the nuclear theory of atomic structure, and used that particle to produce the first artificial disintegration of elements. In the universities of McGill, Manchester, and Cambridge he led and inspired two generations of physicists who—to use his own words—"turned out the facts of Nature," and in the Cavendish Laboratory his "boys" discovered the neutron and artificial disintegration by accelerated particles.

By courtesy of the National Portrait Gallery, London



Rutherford, oil painting by J. Dunn, 1932. In the National Portrait Gallery, London.

Education in New Zealand

Rutherford was born in Spring Grove, New Zealand, on Aug. 30, 1871, the fourth of the 12 children of James, a wheelwright at Brightwater near Nelson on South Island, and Martha Rutherford. His parents, who had emigrated from Great Britain, denied themselves many comforts so that their children might be well educated. In 1887 Ernest won a scholarship to Nelson College, a secondary school, where he was a popular boy, clever with his hands, and a keen footballer. He won prizes in history and languages as well as mathematics. Another scholarship allowed him to enroll in Canterbury College, Christchurch, from where he graduated with a B.A. in 1892 and an M.A. in 1893 with first-class honours in mathematics and physics. Financing himself by part-time teaching, he stayed for a fifth year to do research in physics, studying the properties of iron in high-frequency alternating magnetic fields. He found that he could detect the electromagnetic waves—wireless waves—newly discovered by the German physicist Heinrich Hertz, even after they had passed through brick walls. Two substantial scientific papers on this work won for him an "1851 Exhibition" scholarship, which provided for further education in England.

Before leaving New Zealand he became unofficially engaged to Mary Newton, a daughter of his landlady in Christchurch. Mary preserved his letters from England, as did his mother, who lived to age 92. Thus, a wealth of material is available that sheds much light on the nonscientific aspects of his fascinating personality.

Arrival at the Cavendish Laboratory

On his arrival in Cambridge in 1895, Rutherford began to work under J.J. Thomson, professor of experimental physics at the university's Cavendish Laboratory. Continuing his work on the detection of Hertzian waves over a distance of two miles, he gave an experimental lecture on his results before the Cambridge Physical Society and was delighted when his paper was published in the *Philosophical Transactions* of the Royal Society of London, a signal honour for so young an investigator.

Rutherford made a great impression on colleagues in the Cavendish Laboratory, and Thomson held him in high esteem. He also aroused jealousies in the more conservative members of the Cavendish fraternity, as is clear from his letters to Mary. In December 1895, when Röntgen discovered X rays, Thomson asked Rutherford to join him in a study of the effects of passing a beam of X rays through a gas. They discovered that the X rays produced large quantities of electrically charged particles, or carriers of positive and negative electricity, and that these carriers, or ionized atoms, recombined to form neutral molecules. Working on his own, Rutherford then devised a technique for measuring the velocity and rate of recombination of these positive and negative ions. The published papers on this subject remain classics to the present day.

In 1896 the French physicist Henri Becquerel discovered that uranium emitted rays that could fog a photographic plate as did X rays. Rutherford soon showed that they also ionized air but that they were different from X rays, consisting of two distinct types of radiation. He named them alpha rays, highly powerful in producing ionization but easily absorbed, and beta rays, which produced less radiation but had more penetrating ability. He thought they must be extremely minute particles of matter.

In 1898 Rutherford was appointed to the chair of physics at McGill University in Montreal. To Mary he wrote, "the salary is only 500 pounds but enough for you and me to start on." In the summer of 1900 he traveled to New Zealand to visit his parents and get married. When his daughter Eileen, their only child, was born the next year, he wrote his mother "it is suggested that I call her 'Ione' after my respect for ions in gases."

Toward the end of the 19th century many scientists thought that no new advances in physics remained to be made. Yet within three years Rutherford succeeded in marking out an entirely new branch of physics called radioactivity. He soon discovered that thorium or its compounds disintegrated into a gas that in turn disintegrated into an unknown "active deposit," likewise radioactive. Rutherford and a young chemist, Frederick Soddy, then investigated three groups of radioactive elements—radium, thorium, and actinium. They concluded in 1902 that radioactivity was a process in which atoms of one element spontaneously disintegrated into atoms of an entirely different element, which also remained radioactive. This interpretation was opposed by many chemists who held firmly to the concept of the indestructibility of matter; the suggestion that some atoms could tear themselves apart to form entirely different kinds of matter was to them a remnant of medieval alchemy.

Nevertheless, Rutherford's outstanding work won him recognition by the Royal Society, which elected him a fellow in 1903 and awarded him the Rumford medal in 1904. In his book *Radio-activity* he summarized in 1904 the results of research in that subject. The evidence he marshaled for radioactivity was that it is unaffected by external conditions, such as temperature and chemical change; that more heat is produced than in an ordinary chemical reaction; that new types of matter are produced at a rate in equilibrium with the rate of decay; and that the new products possess distinct chemical properties.

Rutherford, a prodigious worker with tremendous pow-

ers of concentration, continued to make a succession of brilliant discoveries—and with remarkably simple apparatus. For example, he showed (1903) that alpha rays can be deflected by electric and magnetic fields, the direction of the deflection proving that the rays are particles of positive charge; he determined their velocity and the ratio of their charge (E) to their mass (M). These results were obtained by passing such particles between thin metal plates stacked closely together, the size of a matchbox, each plate charged oppositely to its neighbour in one experiment, and in another experiment putting the assembly in a strong magnetic field; and in each experiment he measured the strengths of the fields which just sufficed to prevent the particles from emerging from the stack.

Rutherford wrote 80 scientific papers during his seven years at McGill, made many public appearances, among them the Silliman Memorial Lectures at Yale University in 1905, and received offers of chairs at other universities. In 1907 he returned to England to accept a chair at the University of Manchester, where he continued his research on the alpha particle. With the ingenious apparatus that he and his research assistant, Hans Geiger, had invented, they counted the particles as they were emitted one by one from a known amount of radium; and they also measured the total charge collected from which the charge on each particle could be detected. Combining this result with the rate of production of helium from radium, determined by Rutherford and the American chemist Bertram Borden Boltwood, Rutherford was able to deduce Avogadro's number (the constant number of molecules in the molecular weight in grams of any substance) in the most direct manner conceivable. With his student Thomas D. Royds he proved in 1908 that the alpha particle really is a helium atom by allowing alpha particles to escape through the thin glass wall of a containing vessel into an evacuated outer glass tube and showing that the spectrum of the collected gas was that of helium. Almost immediately, in 1908, came the Nobel Prize—but for Chemistry, for his investigations concerning the disintegration of elements.

The nuclear theory of the atom

In 1911 Rutherford made his greatest contribution to science with his nuclear theory of the atom. He had observed in Montreal that fast-moving alpha particles on passing through thin plates of mica produced diffuse images on photographic plates, whereas a sharp image was produced when there was no obstruction to the passage of the rays. He considered that the particles must be deflected through small angles as they passed close to atoms of the mica, but calculation showed that an electric field of 100,000,000 volts per centimetre was necessary to deflect such particles traveling at 20,000 kilometres per second, a most astonishing conclusion. This phenomenon of scattering was found in the counting experiments with Geiger; Rutherford suggested to Geiger and another student, Ernest Marsden, that it would be of interest to examine whether any particles were scattered backward—*i.e.,* deflected through an angle of more than 90 degrees. To their astonishment, a few particles in every 10,000 were indeed so scattered, emerging from the same side of a gold foil as that on which they had entered. After a number of calculations, Rutherford came to the conclusion that the requisite intense electric field to cause such a large deflection could occur only if all the positive charge in the atom, and therefore almost all the mass, were concentrated on a very small central nucleus some 10,000 times smaller in diameter than that of the entire atom. The positive charge on the nucleus would therefore be balanced by an equal charge on all the electrons distributed somehow around the nucleus.

Although in 1904 Hantaro Nagaoka, a Japanese physicist, had proposed an atomic model with electrons rotating in rings about a central nucleus, it was not taken seriously, because, according to classical electrodynamics, electrons in orbit would have a centripetal acceleration toward the centre of rotation and would thus radiate away their energy, falling into the central nucleus almost immediately. This idea is in marked contrast with the view developed by J.J. Thomson in 1910; he envisaged all the electrons distributed inside a uniformly charged positive sphere of atomic diameter, in which the negative "corpuscles" (electrons) are imbedded. It was not until 1913 that Niels Bohr, a Danish physicist, postulated that electrons, contrary to classical electrodynamics, do not radiate energy during rotation and do indeed move in orbits about a central nucleus, thus upholding the convictions of Nagaoka and Rutherford. A knighthood conferred in 1914 further marked the public recognition of Rutherford's services to science.

Conferral of knighthood

During World War I he worked on the practical problem of submarine detection by underwater acoustics. He produced the first artificial disintegration of an element in 1919, when he found that on collision with an alpha particle an atom of nitrogen was converted into an atom of oxygen and an atom of hydrogen. The same year he succeeded Thomson as Cavendish professor. Although his experimental contributions henceforth were not as numerous as in earlier years, his influence on research students was enormous. In the second Bakerian lecture he gave to the Royal Society in 1920, he speculated upon the existence of the neutron and of isotopes of hydrogen and helium; three of them were eventually discovered by workers in the Cavendish Laboratory.

His service as president of the Royal Society (1925–30) and as chairman of the Academic Assistance Council, which helped almost 1,000 university refugees from Germany, increased the claims upon his time. But whenever possible he worked in the Cavendish Laboratory, where he encouraged students, probed for the facts, and always sought an explanation in simple terms. When in 1934 Enrico Fermi in Rome successfully disintegrated many different elements with neutrons, Rutherford wrote to congratulate him "for escaping from theoretical physics."

Rutherford read widely and enjoyed good health, the game of golf, his home life, and hard work. He could listen to the views of others, his judgments were fair, and from his many students he earned affection and esteem. In 1931 he was made a peer, but any gratification this honour may have brought was marred by the death of his daughter. He died in Cambridge on Oct. 19, 1937, following a short illness, and was buried in Westminster Abbey.

BIBLIOGRAPHY. JAMES CHADWICK (compiler), *The Collected Papers of Lord Rutherford of Nelson,* 3 vol. (1962–65), is a useful resource. A.S. EVE, *Rutherford* (1939), is the official biography sanctioned by Lady Rutherford. Further studies of his life and work include E.N. DA C. ANDRADE, *Rutherford and the Nature of the Atom* (1964, reprinted 1978), a biographical treatment that emphasizes the development of Rutherford's ideas; J.B. BIRKS (ed.), *Rutherford at Manchester* (1962), a discussion of his work at the university there; N. FEATHER, *Lord Rutherford,* new ed. (1973), a discussion of his research work at Cambridge University; MARIO BUNGE and WILLIAM R. SHEA (eds.), *Rutherford and Physics at the Turn of the Century* (1979), a discussion of developments in physics between 1895 and 1905; THADDEUS J. TRENN, *The Self-splitting Atom* (1977), an account of the Rutherford-Soddy collaboration; LAWRENCE BADASH (ed.), *Rutherford and Boltwood* (1969), containing the correspondence between the two scientists concerning the question of radioactivity; and DAVID WILSON, *Rutherford: Simple Genius* (1983), a full account incorporating much new material. (T.E.A./Ed.)

Sacred Offices and Orders

The concept of the religious or sacred personage, as distinguished from the lay practitioner, is a central structural element in most organized religions and in many societies in which religion cannot be regarded as a separate organized entity per se. Although the status and functions of religious personages vary considerably worldwide and throughout history, certain major categories are recognizable for the purposes of discussion.

The role of the founders of religions is obviously of primary importance, but the uniqueness of such positions precludes cross-cultural examination. The concepts of prophecy and sainthood, likewise centring upon the singularity of an individual's relationship to the divine, are treated in the *Macropædia* article DOCTRINES AND DOGMAS, RELIGIOUS. The less strictly religious functions of magician, diviner, and seer are treated in the *Macropædia* article OCCULTISM.

This article treats four highly institutionalized types of religious personage and the institutions, offices, or orders in which such personages function.

For coverage of related topics in the *Macropædia* and the *Micropædia*, see the *Propædia*, sections 811, 812, 821, and 827, and the *Index*.

The article is divided into the following sections:

Priesthood 1014
  Nature and significance 1014
  The priest and his office 1014
  Priesthood in the religions of the world 1015
    Nonliterate cultures
    The ancient Middle East
    Ancient Greece and Rome
    Ancient Judaism
    Christianity
    Hinduism
    Jainism and Buddhism
    Buddhism, Taoism, and Shintō in China and Japan
  The modern situation 1019
Shamanism 1019
  Nature and significance of classic shamanism 1019
  Worldview 1020
  Social role, personality structure, and functions of the shaman 1020
    Social role
    Personal characteristics and selection
    Types and functions
  Symbolism in objects and actions 1021
  Contemporary residues and reconstructions of shamanism 1022
  Shamanistic activity in other cultures 1022
    Eskimo
    American Indians
    Southeast Asia and Oceania
Monasticism 1023
  Nature and significance 1023
  Purposes of monasticism 1024
    Discovery of the true self
    Emancipation of the self
    Social and institutional purposes
  Types of monasticism 1025
    Organizational or institutional types
    Hierarchical and status types
  Varieties of monasticism in the religions of the world 1028
    Religions of the East
    Religions of the West
  Monasticism in the 20th century 1030
Sacred kingship 1031
  Principal schools of interpretation 1031
  Status and functions 1031
    The sacred status of kings, leaders, and chieftains
    Functions of the sacred king: the king as the source of cosmic power, order, and control
    Regal ceremonies
  Conclusion 1035
Bibliography 1035

Priesthood

Throughout the long and varied history of religion, the priesthood has been the official institution that has mediated and maintained a state of equilibrium between the sacred and the profane aspects of human society and that has exercised a stabilizing influence on social structures and on cultic organizations. The term priest is derived etymologically from the Greek word *presbyteros* ("elder"), of which it is a contraction, and it is equated with the Latin word *sacerdos* (the Roman officiant at the sacrifices and sacred rites).

NATURE AND SIGNIFICANCE

The role of the priest

The primary role of the priest is that of the ritual expert, the one who has a special and sometimes secret knowledge of the techniques of worship, including incantations, prayers, sacrificial acts, songs, and other acts that are believed to bridge the separation between the divine or sacred and the profane realms. The priest gains such knowledge through the institution known as the priesthood, which may be composed of various groups or guilds devoted to all or only a few aspects of the priestly craft. Because the priest gains his special knowledge from a school for priests, he is differentiated from other religious and cultic leaders, such as the magician, shaman (healer and visionary), diviner, or prophet, who obtain their positions by means of individual efforts (*e.g.*, learning from a master magician or diviner; individual ecstatic experiences that are publicly recognized). As a member of the institution that regulates the relationship between the divine or sacred and the profane realms through the various rituals of a particular religion, the priest is the accepted religious and spiritual leader in his society.

At various times in the history of a culture or society the priestly institution may be attacked by other institutions or groups that vie for the religious leadership (and thus sometimes the social, political, and economic leadership) of a people. Such anticlericalism is a phenomenon not only of modern society (as noted in the Russian Revolution of 1917, the Mexican Civil War that began in 1857, and other less dramatic movements) but also in the ancient world, such as in Egypt in the 14th century BC, when the priesthood of the god Amon and the priesthood of the god Aton changed positions. Anticlericalism may be fostered by battles for religious leadership between two or more opposing priestly groups, or by prophets and others who are concerned with religious experiences in their personal rather than in their institutional forms. Among Protestants the doctrine of the priesthood of all believers (*i.e.*, all believers have direct access to God) militates generally against strong anticlerical tendencies within their own ranks. In Islām there is, technically, no priesthood, though there are local spiritual and community leaders, such as the imam, the mullah, the mufti, the qadi, and others.

THE PRIEST AND HIS OFFICE

The function of the priest as the mediator and maintainer of the equilibrium between the sacred and the profane in human society, and as the stabilizer of the social structures and the cultic organizations, determines the various criteria for holding the priestly office. In preliterate soci-



A Shintō priest blessing children during the Shichi-go-san (Seven-Five-Three) festival at the Meiji Shrine, Tokyo, Japan. The priest waves a *haraigushi,* a wooden wand with folded-paper pendants, symbolically purifying the children and assuring them of good health and prosperity.
© Cameramann International, Ltd.

Functions and criteria

ety the functions are accomplished by ritual experts who are trained in the special knowledge and techniques of magico-religious disciplines in which sacred power is believed to be inherent. They also are trained in disciplines that enable them to gain a supernormal psychic knowledge and in the techniques of mystical experience. As agents of the sacred power, they are believed to have the ability to control and manipulate through ritual natural processes and events and to engage in a sacramental relationship between man and the sacred order on which the community and its members depend for their sustenance, survival, and well-being.

Because religious institutions and beliefs are intimately connected with other social institutions, those who are set apart to establish efficacious relations with the transcendental powers that are believed to control the universe, as well as human affairs and destinies, occupy a key position at the dynamic centre in the social structure. Thus, in certain types of societies the office and functions of the priest may be limited to those having a particular ancestry, those belonging to certain tribes (such as the Levites in Judaism), families (such as the Eumolpids of the Eleusinian mystery religion of Greece), or castes (such as the Brahmans in Hinduism), and those initiated into certain professional orders (such as the cure doctors among the Maya).

Besides such sociocultural criteria, there are also certain personal requirements in various cultures for those who would become members of a priesthood. Celibacy (as in Roman Catholicism and the Arcakas of the Digambara sect in Jainism), asceticism (as in various Buddhist groups), and religious experiences (as among some holiness Protestant sects) are among the personal requirements for those who aspire to or are chosen to assume the priestly office.

The religious functions of priests are quite varied. In his specific role as the officiant of the rites that unite the sacred and the profane realms, the priest as a pontifex (from the Latin word that means maker of a bridge) celebrates or administers at the rituals of initiation into the cult or church, presides over ritual reenactments of creative, redemptive, or salvatory (salvation-working) events, and offers sacrifices to the gods or to one God. He also functions as a perpetuator of the sacred traditions, practices, and beliefs and as a teacher, healer, counselor, and diviner.

In all their respective offices, functions, and capacities, those who have exercised and manipulated sacred power have attained a uniquely prestigious position as the spiritual and social leaders par excellence. When perplexing and emotional situations that appear to be beyond human control, knowledge, skill, or techniques have arisen in a community, the people have a recourse in the priest, who has the special knowledge of the relationship between the divine or sacred and the profane realms. The priest is often called upon at critical junctures in the lives of individuals (such as birth, puberty, marriage, and death) and in the life of a community (such as seasonal changes or at times of flood, drought, and famine).

PRIESTHOOD IN THE RELIGIONS OF THE WORLD

Medicine men and cult leaders

Nonliterate cultures. The office of priesthood in nonliterate society ranges from the medicine man, who magically manipulates the sacred power as a quasi-impersonal force, to the shaman, who, as an agent of divine or spiritual beings, may be at once a medicine man, a visionary, an occult diviner, and a genuine sacerdotal cult leader. Not infrequently, chiefs or headmen of a clan or village may become sacred men or ritual experts displaying supernatural insight and knowledge as the agents of spiritual beings. As such, they may eventually become prophets or sacral kings occupying a unique status in the community. This status may be acquired by accepted claims of descent from a mythical divine ancestor or god on whom the fertility of the crops and the welfare of the community in general are thought to depend. Occult power and insight may be derived from spirits with which the prophet or diviner is related in a trance or ecstasy, or else in dreams, visions, or auguries and oracles that make known the divine will. To occupy such a position, a strenuous course of training is frequently required involving some knowledge of therapeutics (healing), leechcraft (use of leeches in healing), trephining (operating on the skull), herbs, poisons, and perhaps sleight of hand and similar techniques, together with the development of psychical and occult qualities. While charlatanry (pretensions to medical knowledge) is often practiced, the office also demands an understanding of the technical equipment calculated to bring about the results that are sought, and the proper type of temperament, conditions of mind, and state of emotion.

Unlike the shaman and the medicine man, both of whom exercise their respective functions while relying largely on their own initiative and psychic and occult powers of

divination, healing, and direct access to the spirit world at their command, the priest supplicates and conciliates supernatural forces superior to himself, guards the sacred tradition in his care, and acts as the master of sacrifice. The shaman or magician officiates in his own name and by his own methods and techniques; the priest serves the altar, in the temple or shrine, as the representative of the community in his relations with the gods and the sacred order by virtue of the status and its functions that have been conferred upon him at his ordination, bestowing its sacredness and attendant taboos.

The ancient Middle East. In the Nile valley the occupant of the throne (the pharaoh) was regarded as a god incarnate, the sacred king. Because he was believed to be the epitome of all that was divine, he alone was the intermediary between mankind and the gods whom he summed up in his complex personality. Before Menes, founder of the 1st dynasty (*c.* 3100–2890 BC), centralized the rule of Upper and Lower Egypt, the high priest and king "Scorpion" was traditionally considered to be the incarnation of Horus (the sky god). Under the powerful influence of the priesthood of Heliopolis, the sacred kingship was given a solar significance; the pharaohs were represented as the sons of the sun god Re, who was identified with the god Atum. When Memphis became the imperial city, Re-Atum was brought into relation with its god Ptah (the creator) and subsequently with Osiris, a fertility god who also was regarded as a god of the dead. Osiris was equated with the life-giving waters of the Nile and believed to be the first civilizing king, reigning in the person of the pharaoh as his posthumous son Horus. Therefore in Egypt the divine origin and status of the monarchy was so firmly established that it became the stabilizing force of the civilization in the Nile valley.

Sacred kingship and the priesthood in Egypt

In theory, the king of Egypt was the high priest of every god, and in all important ceremonies he alone was depicted in the temple scenes as the officiant. For practical purposes, however, he delegated his functions to the particular professional local priesthoods. In due course the priests, courtiers, and officials shared in his divine powers and privileges. Even though they only acted on behalf of the divine ruler, his representatives shared in some measure his personality and potency. As the priest par excellence, the pharaoh remained unique, and it was from him that the priesthood and nobility exercised their sacerdotal functions when the solar theology of Heliopolis was established in the 5th dynasty (*c.* 2494–*c.* 2345 BC). After the 5th dynasty the pharaoh was accredited exclusively with having built all the temples, and on their walls he was portrayed officiating as the mediator between mankind and the gods in all important rites. In fact, nevertheless, his place was taken by one of the retinue of priests attached to the royal household or to the local temples. Only very occasionally did he perform his sacerdotal duties in person. Every temple had its high priest who, in addition to his sacred offices, might also be a high judicial official, thereby combining sacred and profane roles, varying with the actions performed.

From the number of titles assigned to priestly officials in the Old Kingdom (*c.* 2686–*c.* 2160 BC), it appears that one priest must have served simultaneously a multiplicity of cults in the shrine to which he was attached and received a share of the offerings and emoluments. In wealthy great temples in the New Kingdom (*c.* 1567–*c.* 1085 BC) there was an elaborate organization, and a large staff engaged in administrative, civil, and educational duties, as well as in their sacerdotal and mortuary functions. There were also a number of priestesses associated with the mother-goddess Hathor (wife of the sun god Re), who were mainly concerned with playing the sacred sistra and other musical instruments. From the 18th to the 21st dynasties (1567–1085 BC), however, under the Theban priesthood of Amon-Re, the priestly hierarchy was able to create an absolute "state within a state."

In Mesopotamia, where kingship occupied a less prominent position than it did in the Nile valley, powerful priesthoods and highly organized temples were firmly established in and after the 4th millennium BC. The temples were centres of sacred learning of the content and methods of incantation, prognostication, exorcism, and political and economic administration, and their attendent priests were divided into classes with special sacerdotal and secular functions. The valleys of the Tigris and Euphrates rivers contained a series of city-states loosely bound together under the rule of a governor (*patesi* or *ensi*) and of a high priest (*sangu mah*). Their main functions were to integrate the temple communities of the city, their revenues and boundaries, and the *unigallu,* or head of the priesthood, was responsible for the performance of the seasonal rites, with the assistance of those who sang the liturgies and directed the temple music and those who uttered the lamentations, invoked the oracles, and interpreted the astrological signs and omens. The duties of other members of the priesthood included making incantations and pronouncing spells and exorcisms. The seers, or *baru,* foretold the future by means of soothsaying, interpreting dreams and visions, and communicating revelations.

Priestly classes and functions in Mesopotamia

An important function of the priesthood in the ancient Middle East was that of exorcism. All down the ages good and evil have been viewed as two contending forces in perpetual conflict, and in the dual task of the riddance of evil and the induction of good the exorcist, the shaman, and the seer have exercised complementary functions hardly distinguishable in the higher cultural levels in Babylonia and Assyria. When the function of the driving out of demons from human beings and buildings by incantations and ritual expulsions was separated from the interpretation of omens and astrological portents, the office of the exorcists was in ever increasing demand. According to cuneiform texts from about the 3rd millennium BC, exorcists held a position of supreme importance, assisting at the consecration of temples, at funerals, and at seasonal ceremonies. In the temple schools, exorcists, diviners, and astrologers, together with physicians, scribes, and judges, fostered the study of astronomy, medicine, and jurisprudence. In their sacerdotal functions and capacities they were regarded as the agents and exponents of the gods, having a monopoly on sacred knowledge, controlling for good or evil the destinies of the community and its members. Therefore, their power was enormous, and they maintained the relation between religion and culture by their all-embracing hierarchy that determined the present welfare and future destinies of both human beings and the social order.

Similar beliefs and practices occurred in northern Syria in the middle of the 2nd millennium BC, before the Israelite settlement in Canaan. Here, again, the priesthood was responsible for the dramatic rituals on which the social structure and the well-being of mankind were believed to depend, especially in the climax of the autumnal festival that culminated in the enthronement of the year god. The priestly movement was centred in the temple of Ugarit, and the religious texts discovered there (Ras Shamra) were essentially liturgical documents devised to make the sacred drama enacted there efficacious. Baal, the weather god, was regarded as "the lord of the furrows of the field," responsible for the rain and the production of the fruits of the earth. A series of temples was erected in his honour in Syria and Palestine. The temple at Ras Shamra was a considerable structure with a lattice, the opening of which was supposed to produce the autumnal rains, and its liturgical ritual was celebrated by the priests and priestesses in their official capacities as those who personified the gods and goddesses responsible for providential bounty and beneficence.

Canaanite priesthood

With the advent of Persian sovereignty in Mesopotamia, a new approach to cosmology and mythology was introduced. In the religion founded by Zoroaster (*c.* 7th century BC), the ultimate ground of the universe was reduced to a single supreme deity, Ahura Mazdā, the All-wise Lord. Astrology became widely practiced by the Iranian Magi (priests of the Zoroastrian religion) who were highly skilled in the knowledge of astral lore and the signs of the zodiac. In the three centuries after the Persian occupation of Mesopotamia (6th century BC), the Babylonian priests under Iranian influence made remarkable strides in astronomical learning that ultimately dethroned astrology from its exalted position in sacred tradition. Beyond the

Euphrates valley, however, it survived until after the Muslim rule, when the astrologer and the astronomer became indistinguishable.

Ancient Greece and Rome. The ancient Greeks were devoid of hierarchical institutions composed of men and women through whom the gods were approached, though priests and priestesses could be found in many places engaging in specific sacerdotal functions and ritual acts. Some of them attained considerable social and civic prestige and importance and were attached to particular temples or shrines such as the oracle at Delphi, which was consulted on private and state matters. Their duties, however, were generally those of members of a household engaged in everyday affairs, rather than of a caste or sacerdotal order set apart and consecrated for the performance of sacrificial and other rites, functions, and practices. Though not regarded as mediators between the gods and men, they did act in such ritualistic capacities in certain civic and administrative areas. There was no specific distinction, however, between them and lay members of society. On the contrary, such officials as magistrates might be priests and vice versa. Some exercised considerable influence if they were regarded as outstandingly efficient, wise, or distinguished in their respective civic or religious capacities.

Civic importance of the priesthood

Similarly, in ancient Rome when the agricultural religion of Numa (the legendary second king) was transformed into an institutional state cult in the republic, it was organized as a hierarchy with the *rex sacrorum* ("king of the sacred things") inheriting the office and attributes of the former priest-king. The *rex sacrorum* had to be a patrician and was chosen for life, subordinate only to the *pontifex maximus,* who was the head of the college of *pontifices* ("advisors on the sacred law") and *flamines* ("priests devoted to a particular god"), 3 of whom were assigned to the gods Jupiter, Mars, and Quirinus, the remaining 12 to other deities. The *flamen* Dialis, dedicated to the supreme sky god, Jupiter, occupied a unique position socially, politically, and sacerdotally and was subject to strict taboos and regulations because of his sacred office. The *flaminica,* the wife of the *flamen* Dialis, participated in his sacredness and official status, and so vital was her association with him and his office that if she died he ceased to perform his functions.

Attached to the temple of the goddess Vesta on the Forum in Rome were the six Vestal Virgins dedicated from childhood to the service of the sacred fire in the *atrium Vestae* (hearth temple) and to the care of the storehouse (*penus*). Originally they were selected from patrician families by the *pontifex maximus,* but later plebians were eligible for election. During their 30 years of service to the goddess, beginning in childhood, chastity had to be strictly observed on pain of death by starvation, but after the completion of the period of service the virgins were free to marry. The duties assigned to them at very ancient festivals, such as the Lupercalia (a fertility festival) on February 15, and at other occasions indicate their unique position and significance in the state cult going back, in all probability, to their origins in the family tradition, associated with the *pontifices* as the officiating priests. Augurs (divinatory personages) had a powerful influence on state religious beliefs and practices, especially in divination to ascertain the will of the gods and the blessing of the crops. They also interpreted signs in the sky as good or bad for the guidance of the magistrates. At the end of the republic (in the 1st century BC) this practice led to abuses that were ridiculed by the politician and orator Cicero. Among other groups of Roman priests were the Salii on the Palatine Hill (the 12 Leapers of the god Mars), and the Luperci, whose sacerdotal functions were confined to the Lupercalia. The *fetiales* were Roman officials employed in making treatises or declarations of war, whose work gradually fell into disuse at the beginning of the empire (late 1st century BC) when the state cult was in decline and losing its vitality.

Ancient Judaism. When Christianity became the legal religion of the Roman Empire after AD 313, it had already inherited from its Jewish background a concept of an organized priesthood. The Jewish priesthood had been centralized in the Temple at Jerusalem (destroyed by the Romans in AD 70) since the 10th century BC. The Hebrew designations for those who exercised oracular, divinatory, and ecstatic functions in ancient sanctuaries that were prominent cultic centres prior to the building of the temple, such as Mamre, Hebron, Bethel, Shechem, and Gilgal, were *kohen* (cohen), *levi, navi',* and *ro'e,* corresponding to priest, Levite, prophet, and seer, respectively. *Kohen* is the equivalent of the Arabic word *kāhin* ("diviner"), and in Hebrew it has the meaning of "priest," denoting the occupant of the office concerned with obtaining oracles by the aid of the ephod (an apronlike garment) containing the Urim and Thummim (sacred lots) and by inspiration, as well as with officiating at a sanctuary. After the 7th century BC, when worship was concentrated in Jerusalem, the capital, the priesthood was restricted to the Levitical house of Aaron (brother of Moses, the 13th-century-BC lawgiver) after having been previously drawn from other lines of descent, such as those of David, Nathan, Micah, and Abinadab (royal, prophetic, and priestly families).

Priests, Levites, prophets, and seers

Whether in fact the Levites ever were members of a sacerdotal tribe is open to debate, but in any case they represented a special fraternity set apart to be guardians of the sanctuary and to engage in oracular and prophetic function, over against the rival priestly *kohanim* in their respective independent confraternities. It was not until after the exile of the Jews to Babylon in 586 BC, when the Priestly Code was drawn up, that the distinction between priests and Levites became absolute. The priesthood was confined exclusively to those claiming succession from Aaron, in spite of the Zadokites claiming priestly descent from Eleazar as an "everlasting covenant" (Num. 18:2–7, 25:13; I Chron. 24:37). The Zadokites may have represented the survival of an ancient Jebusite (Canaanite) royal priesthood, giving them special duties and privileges in the Temple worship above those of the Levites. Later, when the priesthood became reserved for the descendants of the family of Aaron alone, the title was restricted to members of the non-Aaronic families of the tribe acting as the servants of the Temple.

The oracle given by the priests as the inspired word of the Law, called the Torah, which was referred back to Moses in postexilic Judaism, acquired a new significance, involving a rigid observance of its ritual and legal commands that permeated every aspect of life, worship, and conduct. The cessation of the daily sacrifice and other Levitical priestly ministrations in the Temple after the fall of Jerusalem (AD 70) gave a new emphasis to and interpretation of the Torah in the synagogue and in domestic rituals. The prerogatives of the high priest, and those of the priesthood in general, with its exclusive lineage, were maintained after the revolt of the Jews under the leadership of the Maccabees against the Hellenistic Syrians in the 2nd century BC, and the priestly blessing (*dukhan*) in the synagogue remained the exclusive right of the *kohanim* claiming descent from Aaron. They also have had the right to be the first called upon to read the Torah in the synagogue, followed by a Levite. Their privileges, however, have been questioned by some rabbinical authorities (nonpriestly Torah scholars and religious leaders). The Sadducees (deriving their name from the Zadokites) were the high priests in Jerusalem during and after the time of the Hasmoneans, the descendants of the Maccabees (135–104 BC). They exercised considerable influence in the Jewish Sanhedrin (supreme rabbinic court) as the conservative class of the religious aristocracy, favoured accommodations to Greek culture, and maintained the importance of the letter of the Written Law over against the oral tradition of the rival Pharisees. The high priesthood, however, was declining in status under the increasing control of the Roman authorities.

Emphasis on interpretation of the Torah

Christianity. At this critical juncture in Judaism, Christianity, with its own particular conception of priesthood and sacrificial redemption, began in Palestine and rapidly spread throughout the surrounding regions in the Greco-Roman world. In the New Testament the imminent destruction of the Temple in Jerusalem and its worship is predicted, and the culmination of its high priesthood in the person of Christ, after the order of Melchizedek (the priest-king of Jerusalem revered by Abraham in the Old Testament), is proclaimed. The Jewish Aaronic priesthood

and its ritual are represented in the Letter to the Hebrews as imperfect shadows, in a Platonic sense, of the archetypal order of the eternal sacrifice of Christ. Only Christ, who was described as "beyond the veil" (referring to the veil that separated the "Holy of Holies" section from the other areas of the Temple), was believed to be able to save those who came to God through him, since he had removed the barrier of sin that separated man and God for those in a state of grace. The writer of the Letter to the Hebrews stated that Christ, in his reconciling offering as both priest and victim on the cross, accomplished the removal of the barrier in the heavenly tabernacle. This was interpreted in terms of the Hebrew Day of Atonement, a ceremony in which the Jewish high priest made expiation annually for himself, the priesthood, and the whole congregation of Israel. The view of Christ as king, high priest, mediator, and victim influenced the establishment and gradual development of the Christian priesthood in the church, which shares in and makes accessible to its baptized members the all-sufficient priesthood of Christ.

The relationship between presbyter and bishop

Originally the terms *presbyteros* ("elder") and *episkopos* ("overseer"), current in the New Testament and the early church, were probably identical. From the 2nd century on, however, the sacerdotal hierarchy developed along the lines of the Hebrew priesthood, the title *episcopus,* or bishop, becoming reserved for those who presided over the presbyterate, then called *sacerdotes* because they shared in the episcopal *sacerdotium* ("priesthood"), which included the offering of the eucharistic sacrifice of bread and wine. But the conferring of holy orders (ordination of presbyters) and administering the sacrament of confirmation, together with administration of the diocese (jurisdictional area), were confined to the episcopate. In due course the threefold ministry of bishops, priests, and deacons (administrative and liturgical assistants in a parish) became organized on a diocesan basis. This remained the norm in the Western church until the Reformation in the 16th century, when it was repudiated by the continental Reformers (*e.g.,* Luther, Calvin, and Zwingli). In Roman Catholicism, Anglicanism, Swedish Lutheranism, and Eastern Orthodoxy, apostolic succession and jurisdiction has been maintained, especially in the Roman Catholic papacy and in Eastern Orthodox patriarchates.

In European Christianity in the Middle Ages (*c.* 6th to 14th centuries), the deeply laid tradition inherited from the theocracies and priesthoods in the ancient Middle East of a common social, judicial, political, and religious structure was sustained and gave expression to medieval civilization, especially in the Latin West. Church and state became so closely associated that they were virtually identical, as in the cases of the sacred and secular in preliterate societies. As Dante (1265–1321) contended in *De Monarchia* ("Concerning the Monarchy"), the pope, as the head of the spiritual aspects of society, and the emperor, as the ruler of the temporal areas of concern, were equally ordained by God to exercise their functions in their respective spheres of power and influence for the welfare of mankind. If this duality of control did not endure, because the church gradually usurped more and more of the civil jurisdiction and dominated emperors, kings, and other ecclesiastical rulers, the unification of the body politic was rendered more complete as an integrated whole, and the life and character of medieval civilization was determined through the papacy and its priesthood.

The role of the Brahman caste

Hinduism. In Vedic India, the early period of Hinduism, when the priestly caste (Brahman, or Brāhmaṇa) was vested in a particular tribe or special class, it occupied the primary place of importance in the segmentation of Hindu society. The king was subordinate in some respects to the Brahmans, though at one time both sometimes were chosen from the Kṣatriya, or warrior caste. Nevertheless, because the existence of the universe and all cosmic processes were made to depend upon sacrificial offerings, the king delegated such functions to the priests before the end of the 7th century BC, the priests having usurped that position previously held by the kings. The priesthood then exercised supreme control over the fortunes of the gods and men, of heaven and earth, and of the state, though not infrequently priests worked in the service of princes. The Brahmans, however, were so firmly established in the caste system as the twice-born masters of sacrifice and of exclusive sacred knowledge that they were viewed as holding the universe in their grasp. As "lords of creation" by divine right they were divided into 10 tribes, above all other castes. They were required to pass through four ashramas (the celibate religious student, married householder, forest hermit, and wandering ascetic), or conditions of life, prior to and after marriage, to become anchorites (or hermits), and to attain the plenitude of their status, vocation, and authority, thus renewing the creative process by the due performance of the sacrificial offering.

Against this Brahmanic sacerdotalism and its caste organization, the reaction noted in the Upanishads (writings representing the end of the Vedic period), about 1000–500 BC, introduced a mystical conception of the priesthood in Hinduism, and subsequently in Buddhism. Living as hermits in the forest, groups of mystics, reacting to Brahmanic ritualism, gathered around them disciples to learn and propagate a philosophical doctrine based upon the quest to discover a new function for the priesthood, in which the identification of the inner eternal self of man (atman) with the divine ground of the universe (*brahman*) was achieved by asceticism, renunciation of the world, and mystical experience and realization, rather than by the sacrificial offering. This quest was eventually associated with the meditative techniques of yoga (mental and physical exercises) and opened the way for the rejection of the exclusive claims of the Brahmans in favour of the mystical insights and esoteric knowledge accessible to all who adopted the Upanishadic teaching and way of life.

Revival of sacrifice and Vedic study

With the establishment of the four ashramas, sacrifice and Brahmanic study of the Vedas were rehabilitated and brought into relationship with the Upanishadic tradition. The Brahman was thus regarded as occupying the highest state of life; he was

> one who has sensed the deepest self and acts out of that consciousness, [communicating it to others], . . . gives moral guidance, . . . lays down the science of values, draws out the blueprints for social reconstruction, and persuades the world to accept the high ends of life.
>
> (From S. Rādhākrishnan, *Eastern Religions and Western Thought.*)

This view has lacked, however, a conception of the godhead with whom personal relations are possible, thereby making the priesthood an aspect of an impersonal divine power (*brahman*), of a pantheistic principle executing its functions with automatic precision by virtue of its sacerdotal equipment, of an alternating mystical and ascetic priestly sublimation.

Jainism and Buddhism. In Jainism and Buddhism, arising within Brahmanism as nontheistic sectarian movements, the Brahman priesthood was sublimated and the Vedic caste and sacrifice eliminated. In their place a monastic system was evolved, with monks and nuns devoted primarily to rigorous asceticism in the quest of perfection and in the pursuit of chastity and truthfulness. Complete detachment from all phenomenal possessions and connections in Jainism (founded by Mahāvīra in the 6th century BC) made paramount the mendicant life of meditation and spiritual exercises dependent upon the fulfillment of vows of poverty. The functions of the priesthood were sublimated in a process of self-salvation, centred around the purpose of the deliverance of a suffering humanity from the cycles of rebirth. Since in Buddhism *taṇhā* ("desire") was regarded as the fundamental cause of *dukkha* ("the burden of existence"), priestly intervention and the sacrificial offerings were considered to be of no avail in the pursuit of the Eightfold Path leading to the passionless peace of nirvana (the state of bliss).

Priestly traditions in Tantric and Mahāyāna Buddhism

In the absence of any conception of a deity in Buddhism, the question of sacerdotal mediation could be ignored, though in the Mahāyāna ("Greater Vehicle") school and in the Tantric (esoteric, magical) school some elements of the priestly tradition survived. The earliest converts to Buddhism were Brahmans for the most part, and a religious organization in monasteries developed, with various prescribed roles for their inhabitants, with daily routines, certain periods devoted to alms and quests, and periods

for sacred learning and the translation of literary and theological works. To these activities were added other functions, such as recitations of the sacred texts at births, marriages, and in sicknesses to keep evil influences at bay. In the temples, shrines were erected to the honour of the eternal Buddha, and the image of the Blessed One on a lotus bedecked with flowers has become the central object of worship in certain Buddhist groups. The recitation of the ancient Pāli sutras (divine revelations) is believed to transmit the merit inherent in the texts, as is the endless repetition of the sacred formula *Namu O-mi-to* (*Amida-butsu*), "Homage be to the Buddha of infinite light." This, however, is not a sacerdotal devotion performed by, or requiring the presence of, a priest. Buddhism, in fact, has never been able to produce a strong and lasting ecclesiastical organization or a hierarchical segmentation (as in Hinduism), because it has interpreted unity in terms of Becoming, instead of in terms of Being.

Buddhism, Taoism, and Shintō in China and Japan. In the Mahāyāna Buddhist sects, the monks, and those who are popularly known as bonzes, can hardly be said to exercise definitely sacerdotal functions in the temples, monasteries, and shrines. For the most part these functions have been confined to recitations and invocations, which all of the believers share. In China the Taoist "priesthood" emerged as an organized institution at the beginning of the Christian Era. Some were celibates and others were married, living ordinary domestic lives. A number were mendicants and some engaged in alchemy and astrology; others were illiterate. There were also those who assisted in ceremonies and collecting the revenues. In the 6th century AD, in imitation of Buddhism, the Taoist celibates lived in monasteries with a patriarch as the head and interchanged facilities with their Buddhist counterparts. In the Zen contemplative sect in Japan, an attempt was made to attain a state of enlightenment (satori) by a strict discipline and training in quasi-yoga intuitive methods, without priestly intervention or divine grace. The Zen temples and halls of meditation have become centres of learning, art, and education as well as of vigorous austerity and contemplative mysticism, which has not been without attractions to various persons of the West in recent times.

When Buddhism reached China, Japan, and Tibet in the opening centuries of the Christian Era, it came under the influence of the indigenous faiths, cults, and social structures, and, reciprocally, it became a most important influence, adapting its beliefs and customs to those already established in these regions. In the second half of the 6th century AD, after Buddhism had acquired official recognition, pagodas, temples, and monasteries were erected with ornamentations of Buddhist origin. Buddhism adapted itself to Shintō, the native religion of Japan, and to its shrines, festivals, and rites. The functions of the four priestly classes (*e.g.,* as ritual experts, diviners, musicians, female dancers, and "abstainers" to ward off pollution) that emerged from the family or tribal cults of Shintō were absorbed by Buddhism.

When Shintō was restored as the national religion of Japan in the 19th century, after a period of decline, the Shintō and Buddhist priests were assigned their respective duties and offices by the State Department of Religion without discrimination, for the maintenance of reverence for the gods and love of country (the Truth of Heaven and the Way of Humanity) and proper respect for the sacral emperor (the mikado). This dual sacerdotal combination lasted only until 1875, because Buddhism and Shintō were basically incompatible. This resulted in Shrine Shintō becoming the national faith under the Imperial family, maintaining its divine status, cultic practices, and priesthood, but leaving Buddhism free to propagate its dharma (or law) in its own way. New rituals and ceremonies were composed by the government for use at the Shintō shrines, and the duties and grades of the priests were fixed.

THE MODERN SITUATION

In recent years, in Christianity especially, notwithstanding the doctrinal divergencies and modes of expression in the nature and function of the priesthood, new approaches have been made by both Catholic and Protestant theologians, liturgical scholars, and laypersons. This is particularly apparent in the ecumenical and liturgical movements in Western Christendom, in the administration of the sacraments, and in the participation of the laity in the liturgy and in the other offices. In Roman Catholicism, especially under the influence of the second Vatican Council (1962–65), and in the contemporary Anglican Convocations, the "priesthood of the laity" has been more widely recognized and practiced, though this has been a cardinal doctrine of Protestantism since the Reformation. Laypersons trained in liturgical functions have assumed functions, such as reading of the Scriptures and administering the eucharistic elements of bread and wine, in various Protestant churches and, in some instances, in Roman Catholic celebrations that in the past have been the prerogatives of priests. In the Eastern Orthodox churches, such movements and influences have been less effective and operative, largely because of the heavy pressure they endured from governments that were hostile in the formerly communist countries where Orthodox membership is concentrated. They have always, however, preserved a living unity of faith, worship, and organization. Priesthood is inherent in these institutions; it has proved to be the unifying, stabilizing force in Eastern Orthodoxy, the second largest Christian body (after Roman Catholicism), beset as it has been by so many hazards and hostilities in its long and checkered history. (E.O.J.)

Participation of the laity in liturgical and other functions

Shamanism

Shamanism is a religious phenomenon centred on the shaman, an ecstatic figure believed to have power to heal the sick and to communicate with the world beyond. The term applies primarily to the religious systems and phenomena of the northern Asian, Ural-Altaic (*e.g.,* Mansi, Khanty, Samoyed, Tungus), and Paleo-Asian (*e.g.,* Yukaghir, Chukchi, Koryak) peoples.

The term shamanism comes from the Manchu-Tungus word *šaman.* The noun is formed from the verb *ša-* ("to know"); thus, "shaman" literally means "he who knows." Various other terms are used by other peoples among whom shamanism exists.

There is no single definition of shamanism that applies to the elements of shamanistic activity found in North and South America, in southeastern India, in Australia, and in small areas all over the world as well as to the phenomena among the northern Asian, Ural-Altaic, and Paleo-Asian peoples. It is generally agreed that shamanism evolved before the development of class society in the Neolithic Period and the Bronze Age, that it was practiced among peoples living in the hunting-and-gathering stage, and that it continued to exist, somewhat altered, among peoples who had reached the animal-raising and horticultural stage. According to some scholars, it originated and evolved among the more developed societies that bred cattle for production. Opinions differ as to whether the term shamanism may be applied to all religious systems in which the central personage is believed to have direct intercourse through an ecstatic state with the transcendent world that permits him to act as healer, diviner, and psychopomp (escort of souls of the dead to the other world). Since ecstasy is a psychosomatic phenomenon that may be brought about at any time by persons with the ability to do so, the essence of shamanism lies not in the general phenomenon but in specific notions, actions, and objects connected with the ecstatic state.

NATURE AND SIGNIFICANCE OF CLASSIC SHAMANISM

Among the peoples of northern Asia, shamanism developed into a more definitely articulated and specialized form than among other peoples. Shamanism as practiced there is distinguished by its special clothing, accessories, and rites as well as by the specific worldview connected with them. North Asiatic shamanism in the 19th century, which may be taken as the classical form, was characterized by the following traits:

The basic characteristics of shamanism

(1) A specialist (man or woman) is accepted by the society as being able to communicate directly with the transcendent world and thereby also possessed of the abil-

ity to heal and to divine; this person is held to be of great use to society in dealing with the spirit world. (2) This figure has special physical and mental characteristics: he is neurasthenic or epileptic, with perhaps some minor defect (*e.g.,* six fingers or more teeth than normal), and with an intuitive, sensitive, mercurial personality. (3) He is believed to have an active spirit or group of spirits to assist him and also may have a passive guardian spirit present in the form of an animal or a person of the other sex—possibly as a sexual partner. (4) The exceptional abilities and the consequent social role of the shaman are believed to result from his being the "choice" of the spirits, though the one who is chosen—often an adolescent—may resist his selection, sometimes for years. Torture by the spirits, appearing in the form of illness, breaks the resistance of the shaman candidate and he (or she) has to accept the vocation. (5) The initiation of the shaman, depending on the belief system, may happen on a transcendent level or on a realistic level, or sometimes on both, one after the other. While the candidate lies as if dead, in a trance state, the body is cut into pieces by the spirits of the Yonder World or is submitted to a similar trial. The reason for cutting up his body is to see whether he has more bones than the average person. After awakening, the rite of symbolic initiation, climbing the World Tree, is occasionally performed. (6) By falling into ecstasy at will, the shaman is believed to be able to communicate directly with the spirits either by his soul leaving the body to enter their realm or by acting as their mouthpiece, like a medium. (7) One of the distinguishing traits of shamanism is the combat of two shamans in the form of animals, usually reindeer or horned cattle. The combat rarely has a definite purpose but rather seems to be a deed the shaman is compelled to do. The outcome of the combat means well-being for the victor and destruction for the loser. (8) In going into ecstasy, as well as in his mystical combat, certain objects are used: drum, drumstick, headgear, gown, metal rattlers, and staff. (The specific materials and shapes of these instruments are useful for identifying the types and species of shamanism and following their development.) (9) Characteristic folklore texts and shaman songs have come into being as improvisations on traditional formulas in luring calls and imitations of animal sounds.

As an ethnological term, shamanism is applied primarily to the religious systems of those regions in which all these traits are present together. In addition, there are primitive religions in which some of the above criteria are missing but which are still partially shamanistic; *e.g.,* among the Chukchi of northeastern Siberia, the specialist chosen by the spirits does not fall into ecstasy. Such religious systems may be regarded as marginally shamanistic.

Peripheral or marginal shamanism

Phenomena similar to some of the traits of shamanism may be found among primitive peoples everywhere in the world. Such detached traits, however, are not necessarily shamanistic. The central personalities in such systems—sorcerers, medicine men, and the like—may communicate with the other world through ecstasy, but, unlike the shaman, they have attained their position through deliberate study and the application of rational knowledge. Although they perform ceremonies as priests, hold positions of authority, and possess magical abilities, the structure and quality of their transcendental activities are entirely different from that of the shaman.

WORLDVIEW

The universe. The classic worldview of shamanism is found among the peoples of northern Asia. In their view the universe is full of heavenly bodies peopled by spiritual beings. Their own world is disk-shaped—saucerlike—with an opening in the middle leading into the Netherworld; the Upper World stands over the Central World, or Earth, this world having a manyfold vault. The Earth, or Central World, stands in water held on the back of a colossal monster that may be a turtle, a huge fish, a bull, or a mammoth. The movement of this animal causes earthquakes. The Earth is surrounded by an immense belt. It is connected with the Upper World by the Pillar of the World. The Upper World consists of several strata—3, 7, 9, or 17. On the navel of the Earth stands the Cosmic Tree, which reaches up to the dwelling of the upper gods.

Gods, spirits, and souls. All three worlds are inhabited by spirits. Among the Mongolian and Turkish peoples, Ülgen, a benevolent deity and the god of the Upper World, has seven sons and nine daughters. Among the Buryat of southern Siberia, Tengri (often identified with Ülgen) also has children, the Khat's—the western ones being good and the eastern ones wicked. The gods of the Buryats number 99 and fall into two categories: the 55 good gods of the west whose attribute is "white," and the 44 wicked gods of the east whose attribute is "black." The leader of the latter is Erlen khan, a figure equivalent to Erlik khan of the Altai Kizhi people, who is the ruler of the Underworld. Besides gods and the progeny of gods—both sons and daughters—other spirits also inhabit all three worlds. Fire is also personified, as is the Earth itself. Such personifications are represented in idols as well. Man, besides his body, consists of a soul, even of several souls. Man also has a mirror soul, which can be seen when looking into water, and a shadow soul, which is visible when the sun is shining.

SOCIAL ROLE, PERSONALITY STRUCTURE, AND FUNCTIONS OF THE SHAMAN

Social role. The extraordinary profession of the shaman naturally distinguishes him socially. The belief that he communicates with the spirits gives him authority. Furthermore, the belief that his actions may not only bring benefit but also harm makes him feared. Even a good (white) shaman may do harm, and a wicked (black) shaman, who is in contact with the spirits of the Lower World, is very alarming. In consequence of his profession, the shaman cannot go hunting and fishing and cannot participate in productive work; therefore, he must be supported by the community, which considers his professional activity necessary. Some shamans make use of their special position for economic gain. Among the reindeer-raising Evenk of northern Siberia, poor families have to pay yearly one animal, and rich ones pay two, three, or even four, to the shaman for his activities. A saying of the Altai Kizhi illustrates this situation: "If the beast becomes ill, the dogs fatten; if man becomes ill, the shaman fattens."

Special economic position of the shaman

Among the Evenk, it was the duty of every member of the clan to aid the shaman economically. When distributing the fishing spots in the spring and summer, the part of the river most abundant in fish was given to the shaman, and even the fishing devices were set up for him. He was aided in grazing and herding the reindeer in autumn, and in winter the members of the clan went hunting in his stead. Even furs were presented to the shaman occasionally. The social authority of the shaman was shown through the honours bestowed on him and the practice of always giving him the best food. Generally, the shaman was never contradicted, nor was any unfavourable opinion expressed about him behind his back.

Such an economic and social position resulted in the shaman attaining political power. As early as 1752, for instance, it was noted that the Tungus shaman was also the leader of his clan. Along the Yenisey River, shamans led armed groups of the Evenk on the left and the right banks who fought against each other. In the northern forest regions of Mongolia the shamans stood at the head of the tribes and clans. In the fight of the Buryat against the Russians in the 17th and 18th centuries, the shaman always led the fight. The ruler of one domain among the Vadeyev Samoyed in northern Siberia was a shaman as well as a reigning prince. Among the Eskimo of North America and Asia, the positions of leader and of shaman often are occupied by the same person: indeed, the two Eskimo terms *angakok* ("shaman") and *angajkok* ("leader") have the same root.

Personal characteristics and selection. Scholars generally agree that the shaman acquires his profession through inheritance, learning, or an inner call, or "vocation," but each of these terms requires some qualification. "Inheritance" means that the soul of a dead shaman or the so-called shaman illness is inherited. "Learning" here does not usually mean the study of exact knowledge and explicit dogma, for the shaman, it is believed, is taught by the spir-

The compulsory nature of shamanism as a vocation

its. The inner "call" is in reality not the call of the person but of the spirit who has chosen him and who forces him to accept his vocation. This compulsion is unavoidable. "Had I not become shaman, I would have died," said a Gilyak (southeastern Siberia). The future shaman of the Altai Kizhi was subjected to terrible torture until, finally, he grasped the drum and began to act as a shaman.

According to the abundant literature on the subject and the experience of investigators in the field, no one voluntarily ventures into the shaman role, nor does a candidate have time to study the role. Such study, however, is not necessary, because peoples born into a culture with shamanistic beliefs know them thoroughly, and when the call arrives, the future shaman can learn specific practices by close observation of active shamans, even the technique of ecstasy. The shamanistic view of the world and spirits is already familiar to him. The various qualitative categories by which shamans are distinguished—small, intermediate, and great—are explained by the category of the spirit who chose the shaman. It is evident, however, that this depends on the personal abilities of the shaman himself, his mental capacities, his dramatic talent, and his power to make his will effective. All these elements add to the quality of his shaman performance, the art expressed in it.

The shaman is born to his role, as is evident in certain marks distinguishing him from ordinary men. He may be born with more bones in his body—*e.g.,* teeth or fingers—than other people. Therefore, he does not become a shaman simply by willing it, for it is not the shaman who summons up the spirits, but they, the supernatural beings, who choose him. They call him before his birth. At the age of adolescence, usually at the period of sexual ripening, the chosen one suddenly falls into hysterics with faintings, visions, and similar symptoms, being tortured sometimes for weeks. Then, in a vision or a dream, the spirit who has chosen him appears and announces his being chosen. This call is necessary for the shaman to acquire his powers. The spirit who has chosen him first lavishes the unwilling shaman-to-be with all sorts of promises and, if he does not win his consent, goes on to torment him. This so-called shaman illness will anguish him for months, perhaps for years, as long as he does not accept the shaman profession. When the candidate finally gives way to the compulsion and becomes a shaman, he falls asleep and sleeps for a long time—three days, seven days, or thrice three days. During the "long sleep," the candidate, according to belief, is cut into pieces by the spirits, who count his bones, determining whether he truly has an "extra bone." If so, he has become a shaman. Some people, such as the Mongols and the Manchu-Tungus, still initiate the shaman. They introduce him to the supernatural beings, and he symbolically ascends the "tree-up-to-the-heavens"—that is, the pole representing it.

Shamanism and psychoses

The central activity of the shaman is ecstasy at the wish of his clients, and some have inferred from this that he is psychotic. A person becomes a shaman at puberty, according to this view, when, especially in subarctic and Arctic climatic conditions, changes in his constitution and nervous system may result in the loss of mental balance and in various mental disorders. Social and ethnic factors also may be seen to support the role of psychoses. A person who was born with certain marks knows he is destined to the vocation and becomes apprehensive of the call of the spirits. His fears of the event, according to this theory, create the hallucinations, and the hallucinations reinforce the belief.

Types and functions. *Differences in quality and degree.* Shamans differ greatly in quality and in degree. Difference of quality is manifest in the kind of spirits the shaman communicates with. "White" shamans, for example, apply to a benevolent deity and the good spirits, while "black" shamans call on a wicked deity and the wicked spirits.

The difference in degree is exemplified in the Yakut belief (in northeastern Siberia) that the souls of the future shamans are reared upon an immensely high tree in the Upper World, in nests at various heights. The greatest shamans are brought up close to the top of the tree, the intermediate ones toward the middle, and the smaller ones on the lower branches. Hence, shamans may be classified into three groups: great, intermediate, and last, according to their powers.

Basic tasks. It is the obligation of the shaman to know all matters that human beings need to know in everyday life but are unable to learn through their own capacities. He foresees events distant in time and space, discovers the place of a lost animal, forecasts prospects for fishing and hunting, and assists in increasing the gain. Besides these everyday functions, he is a healer and a psychopomp. He fulfills all these obligations by communicating with the spirits directly whenever he pleases.

The shaman's assistance is necessary at the three great events of life: birth, marriage, and death. If a woman bears no child, for instance, then, according to the belief of the Nanai (Golds), in the Amur region of northeastern Asia, the shaman ascends to heaven and sends an embryo soul (*omija*) from the tree of embryos (*omija muoni*). Among the Buryat, the shaman performs libations after birth to keep the infant from crying and to help him develop more quickly. Among the Nanai, when death occurs the shaman is necessary to catch the soul of the deceased floating in the universe and to escort it to the Yonder World. Illness is believed to be caused by the spirits, who must be appeased for a cure to be effected. Among the Khanty of northern Siberia, the shaman decides how many reindeer should be sacrificed to appease the spirit who causes an illness. Among the Altai Kizhi, he states which *körmös* (soul of the dead) caused the disaster and what to do to conciliate it. Illness might be caused by the soul having left the patient's body and fallen into the hands of spirits who are angry with it and therefore torment it; the shaman liberates the strayed soul. Illness also may be caused by spirits entering into a man; the shaman cures it by driving the spirits out.

The shaman as healer

Forms of ecstasy. The shaman may fulfill his obligations either by communicating with the spirits at will or through ecstasy. The latter has two forms: possession ecstasy, in which the body of the shaman is possessed by the spirit, and wandering ecstasy, in which his soul departs into the realm of spirits. In passive ecstasy the possessed gets into an intense mental state and shows superhuman strength and knowledge: he quivers, rages, struggles, and finally falls into an unconscious trancelike condition. After accepting the spirit, the shaman becomes its mouthpiece—"he becomes him who entered him." In active ecstasy the shaman's life functions decrease to an abnormal minimum, and he falls into a trancelike condition. The soul of the shaman, it is believed, then leaves his body and seeks one of the three worlds, or strata, of heaven. After awakening, he relates his experiences, where he wandered, and with whom he spoke. There are cases of possession ecstasy and wandering ecstasy combined, when the spirit first enters the shaman and then leads his soul to the world of supernatural beings. Scholars differ as to which is the original and which the derivative form; *e.g.,* the historian of religions Mircea Eliade did not consider possession ecstasy to be essential to shamanism.

SYMBOLISM IN OBJECTS AND ACTIONS

The shaman attains the ecstasy necessary for communicating with the spirits through the performance of the shaman rite, which requires certain appurtenances.

Dress. He wears a ritual gown, which usually imitates an animal—a deer, a bird, a bear. Similarly, the headdress is a crown made of antlers or a band into which feathers of birds have been pierced. The footwear is also symbolic—iron deer hooves, birds' claws, or bears' paws. The clothing of the shamans among the Tofalar (Karagasy), Soyet, and Darhat are decorated with representations of human bones—ribs, arm and finger bones. The shamans of the Goldi-Ude tribe perform the ceremony in a singular shirt and in a front and back apron on which there are representations of snakes, lizards, frogs, and other animals.

Drums, sticks, and other objects. An important device of the shaman is the drum, which always has only one membrane. It is usually oval but sometimes round. The outer side of the membrane, and the inside as well among some peoples, is decorated with drawings; *e.g.,* the Turks, or Tatars, of Abakan mark the Upper and Lower World.

The handle is usually in the shape of a cross, but sometimes there is only one handle in a vertical direction or in the shape of the letter Y or X. The drumstick is made of wood or horn, and the beating surface is covered with fur. In some cases the drumstick is decorated with human and animal figures, and rattling rings often hang down from it. During the ecstasy brought on by the sound of the drum, the spirits move to the shaman—into him or into the drum—or the soul of the shaman travels to the realm of the spirits. In the latter case the shaman makes the journey on the drum as if riding on an animal, the drumstick being his lash. Sometimes the shaman makes the journey on a river and the drum is his boat, the drumstick his oar. All this is revealed in the shaman song. Besides the drum, the Buryat shaman sometimes makes the journey with sticks ending in the figure of a horse's head. The shaman of the Tungus people, who raise reindeer, makes the journey on a stick ending in the figure of a reindeer's head. Among some people, the shaman wears a metal disk, a "shaman-mirror."

The drum as symbolic bearer of the shaman

National Museum of Finland



A Mongol shaman (photographed *c.* 1909) wearing a ritual gown and holding a drum, the cross-shaped handle of which has the image of a spirit helper.

Drama and dance. Shamanic symbolism is impressively presented in dramatic enactment and dance, as observed among the peoples where the shamanistic rites survived longest (such as the Samoyed, Tofalar, Buryat, and Tungus). The shaman, garbed in his ritual robes, lifts his voice in song to the spirits. This song is always improvised, with certain obligatory images and similes, dialogue, and refrains. The performance always takes place in the evening. The theatre is the conical tent, or yurt; the stage is the space around the fire where the spirits are invoked. The audience consists of the invited members of the clan, awaiting the spirits in awe. The stage lighter and decorator, the shaman's assistant, tends the fire so as to throw fantastic shadows onto the wall. All these effects help those present to visualize everything that the recited action of the shaman narrates. The shaman is an actor, dancer and singer, and a whole orchestra. This restless figure is a fascinating sight, with his cloak floating in the light of a fire in which anything might be imagined. The ribbons of his gown flit around him, his round mirror reflects the flames, and his trinkets jingle. The sound of his drum excites not only the shaman but also his audience. An integral characteristic of this drama is that those who are present are not mere objective spectators but rather faithful believers, and their belief enables the shaman to achieve results, as in healing mental illnesses. Among some people—the Altai Kizhi, for instance—a tall tree is set into the smoke opening at the top of the tent, symbolizing the Tree of the World. The shaman ascends the tree to the height of the Upper World, which is announced to his audience through the text of his song.

CONTEMPORARY RESIDUES AND RECONSTRUCTIONS OF SHAMANISM

The residues of shamanism may be found among peoples who have been converted to religions of a later stage of culture—*e.g.,* Finno-Ugric peoples who became Christians, Turkic peoples in Central Asia and Asia Minor who became Muslim, and Mongols who became Buddhists. Among the Finns, the *tietäjä,* a figure equivalent to the shaman, also is born with one more tooth than normal. Among the Osmanlı Turks of Asia Minor, the horned headwear of the shaman is remembered in popular belief. Among people who formerly believed in shamanism but later were converted to various world religions (*e.g.,* Christianity or Islām), former shamanism may be revealed through an analysis of their folklore and folk beliefs. An example of such a case is the discovery of former Hungarian shamanism. In northern Asia shamanism appears in various forms that may be attributed to differences in cultural phases. In the most northern parts, among the Chukchi, Koryak, and Itelmen, the shaman does not exist as a member of a special profession: a suitable member of the family—often an old woman—performs the activity of the shaman. Often the shamans are of "changed sex"—effeminate men who have adopted feminine clothing and behaviour at the command of their "spirit." Among the Yukaghir of Arctic Siberia, shamanism is part of the cult of the clan; so also among pockets in the Ob-Ugrian peoples and among all three Altaic peoples: Turkic, Mongol, and Manchu-Tungus. These are instances of definitely professional shamanism, which, however, have been excluded by so-called higher religions. Shamanism was excluded among the Khalkha-Mongolian and eastern Buryat, who became Buddhists, and among the Kazakh and Kyrgyz who adopted Islām, and it was greatly changed and developed into an atypical form by the Manchurians.

Certain scholars have investigated ecstatic actions that may be adjudged outside the area of shamanism in the strictest sense. Mircea Eliade studied North and South America, Southeastern Asia and Oceania, Tibet, and China (see below *Shamanistic activity in other cultures*), and S.P. Tokarev has also studied Africa. Some scholars suppose that the phenomena of shamanism spread to the two American continents when the first settlers migrated from Asia. The shamanistic phenomena in the Shintō religion of Japan are attributed to the migration of nomadic peoples from the territory bordering northern Korea. No such theory of migration has yet been developed to explain the "shamanism" of Southeast Asia and Oceania. Those who oppose this broad usage of the term shamanism argue that an apparent structural similarity among phenomena in widely separated areas does not justify an assertion of a common source or that typological similarity must be distinguished from a genetic connection. For them, shamanism may be attributed only to a precise pattern of cultural phenomena in a specific, well-defined territory, one that forms a concrete, systematic whole, such as the religious systems of the peoples mentioned at the beginning of this section. (V.D.)

Migration and the spread of shamanism

SHAMANISTIC ACTIVITY IN OTHER CULTURES

Although the classic model and most complete expression of shamanism is found in the Arctic and central Asian regions, the phenomenon must not be considered as limited to those countries. It is encountered, for example, in Southeast Asia, Oceania, and among many North Amer-

ican aboriginal tribes (shamanism does not play a role of the first order in Africa). A distinction is to be made, however, between the religions dominated by a shamanistic ideology and by shamanistic techniques (as is the case with Siberian and Indonesian religions) and those in which shamanism constitutes rather a secondary phenomenon.

Eskimo. Shamanism predominates in the religious life of the Eskimos. The chief prerogatives of the Eskimo shaman (*angakok;* plural, *angákut*) are healing, the ecstatic underwater journey to the Mother of Animals for the purpose of assuring an abundance of game, and the aid he brings to barren women. Sickness is brought on by the violation of taboos or results from the capture of the soul by a ghost. In the first case the shaman strives to drive out the impurity by collective confessions; in the second case he undertakes the ecstatic journey to heaven or to the depths of the sea to retrieve the sick person's soul and restore it to his body. The *angakok* is also a specialist in magic flight. Some shamans are reputed to have visited the moon; others claim to have flown around the earth. The *angákut* also know the future, make prophecies, predict changes in the weather, and excel at magic feats.

American Indians. Among many North American tribes shamanism constitutes the most important aspect of the religious life. The shaman is characterized by the supernatural power he acquires as the result of a direct personal experience. This power is obtained either spontaneously or after a voluntary quest, but in either case the future shaman has to undergo certain initiatory trials. In general, the power is utilized in such a way as to affect the whole society. The shaman's principal function is healing, but he also plays an important role in other magico-religious rites such as communal hunting and, where they exist, secret societies or mystical movements (*e.g.,* of the Ghost Dance religion type). North American shamans, like all their fellows, claim to control the weather (bring on or stop the rain, etc.), know future events, expose the perpetrators of thefts, and so on. Furthermore, they defend men against sorcerers. But the magico-medical powers held by North American shamans do not exhaust their ecstatic abilities. There is every reason to suppose that modern secret societies and mystical movements among the Indians have appropriated in large part the ecstatic activity that once characterized shamanism.

Legacy of ecstatic activity

In the tribes of South America the shaman enjoys considerable prestige and authority. Not only is he the healer par excellence and, in certain regions, the guide of souls of the dead to their new abode, he is also the intermediary between men and the gods or spirits, substituting himself for the priest at times. He guarantees the respect for ritual observances, defends the tribe against evil spirits, points out places for fruitful hunting and fishing, increases the wildlife, controls the weather, eases childbirth, reveals future events, and so on. Of course, the South American shaman can also fill the role of sorcerer; he can, for example, change himself into an animal and drink the blood of his enemies. Yet it is rather to his ecstatic abilities that the South American shaman owes his magico-religious position and social authority.

It is probable that a certain form of shamanism was diffused on the two American continents with the first waves of immigrants from Asia; later contacts between northern Asia and North America made Asian influence possible well after the penetration of the first immigrants.

Southeast Asia and Oceania. Shamanism is prevalent in the Malay Peninsula and in Oceania. Among the Negritos of the Malay Peninsula, the shaman heals with the help of celestial spritis or by using crystals of quartz. But the influence of Indo-Malayan beliefs is noticeable, too (the shaman changing into a tiger, trance achieved by dancing, etc.). In the Andaman Islands the shaman gets his power from contact with spirits. The most common method is to "die" and return to life, the traditional pattern of shamanic initiation. The shamans gain their reputation through their acts of healing and their meteorological magic (they are thought to bring on storms).

The distinctive marks of Malayan shamanism are the calling forth of the tiger's spirit and the achievement of the trance (*lupa*), during which the spirits seize the shaman, possess him, and reply to questions asked by the audience. Mediumistic qualities also are characteristic of different forms of shamanism in Sumatra, Borneo, and Celebes. Among the Ngadju-Dayak of Borneo there even exists a special class of shamans, the *basirs* (literally, "incapable of procreation"), hermaphrodites who dress and act like women. These are considered to be intermediaries between heaven and earth because they unite in their own person the feminine element (earth) and the masculine element (heaven).

The shaman as medium

Possession by gods or spirits is a peculiarity of Polynesian ecstatic religion. The extreme frequency of possession in that region has made possible a proliferation of healers. Priests, inspired persons, medicine men, and sorcerers may all perform magical cures. For this reason it is not possible to speak of shamanism *stricto sensu* in Polynesia.

In Australia a person becomes a medicine man through a ritual of initiatory death, followed by a resurrection to a new and superhuman condition. But this initiatory death, like that of the Siberian shaman, has two specific marks not found elsewhere in combination: first, a series of operations performed on the candidate's body (opening of the abdomen, renewal of the organs, washing and drying of the bones, insertion of magical substances); second, an ascent to heaven, sometimes followed by other ecstatic journeys into the other world. The revelations concerning the secret techniques of the medicine men are obtained in a trance, a dream, or in the waking state before, during, or after the initiatory ritual proper. (M.Ee.)

Monasticism

The word monasticism is derived from the Greek *monachos* "living alone," but the etymology indicates only one of the elements of monasticism as a force in history and society. The etymological method of arriving at an understanding of monasticism is, at any rate, misleading because a large section of the world's monastics live in cenobite (common life) communities. The term monasticism does, however, indicate what later became a socially and historically highly significant feature—*i.e.,* living alone in the sense of being unmarried or celibate, though this feature is not directly related to its etymology.

Still, even this aspect of monasticism does not extend beyond the cultures and languages within which was formulated the religious terminology that originated in the eastern Mediterranean—*i.e.,* the Judeo-Christian and Islāmic religions. In the Islāmic world, terms that can be translated by "monk," "monastic," and similar words do not mean "single" in the Arabic and Persian terminologies as in the Greek. Other aspects (*e.g., zuhd,* "asceticism") of the monastic life in Islām provided the etymological and definitional sets denoting "monasticism." None of the many Indic (Sanskrit, Pāli, Apabhraṃśa, Prākrit) terms for monk means "single" or "living alone" in these languages, although monastics within those traditions—Brahman-Hindu, Buddhist, and Jaina—do indeed live alone or in groupings that are set off from the rest of their societies, analogously to Jewish, Christian, and Islāmic monastics. The etymologies of the Indian as well as of some of the Arabic and Persian terms connote poverty, certain ecstatic states of mind, dress conventions, and so on, while some (by historical rather than semantic connection) imply single, celibate living.

NATURE AND SIGNIFICANCE

Within a cross-cultural perspective, monastics can be seen to have been instrumental in creating, preserving, and augmenting institutions of learning, religious as well as secular, and in transmitting cultural goods, artifacts, and intellectual skills through the generations. Monastic institutions also have had medical, political, and military-related functions, though the latter two have all but disappeared in most societies.

A definition of monasticism that would cover all its forms, Eastern and Western, must necessarily be wide; particulars must be relegated to the analysis of specific monastic systems. A universal definition might thus be: "monasticism" is a term covering religious institutions,

Definition and characteristics

ritual, and belief systems whose agents, members, or participants attempt to practice religious works that are above and beyond those required by the religious teachings of their society or of exceptional individual religious and spiritual leaders in their society—*i.e.,* those who have interpreted radically the tenets that apply to all believers or to the whole society. Beyond such a statement, one can speak only of the major characteristics of the monastic life and its institutions, since none of them is universal. Celibacy is fundamental to the majority of the world's religious orders, but it is by no means universal; asceticism is universal, provided the term is defined widely enough so as to include all supererogatory religious practices. The truly universal characteristic of monasticism follows from its definition: the monastic separates himself from his society, either to be by himself as a hermit or an anchorite (religious recluse) or to join a society of others who have separated themselves from their surroundings with similar intentions—*i.e.,* the full-time pursuit of the religious life in its most radical and usually its most fundamentalistic interpretation.

There is no monasticism in societies that do not have a written and transmitted lore. Nonliterate societies do not have monastic institutions, since the monastic takes his departure from an established corpus, or written body, of religious doctrine, which has undergone criticism and has generated countercriticism as well as a dialectic that presupposes a literate, codified manipulation of the doctrine. The monk and the monastic founders may either support or oppose the official religious tradition, but the presence of such a tradition is essential as the matrix of all monastic endeavour.

PURPOSES OF MONASTICISM

Discovery of the true self. *Overcoming imperfections.* All monasticism has its mainstay in theologically based convictions that the present state of things leaves much to be desired—that life in society cannot generate the spiritual consummation stipulated by the religious worldview that was or is recommended overtly or covertly by its founder. In some traditions, especially in those of South Asian provenance, the true "self" is clogged and covered by imperfections—by sin, ignorance, or other theologically suggested impediments. The ego with which the layperson and the seeking neophyte identifies is not his true self—it is the latter that has to be discovered, uncovered. The sundry barriers—differently conceived as matter, individuated mind, or a soul-mind aggregate defiled by sin, ignorance, and perversion—must be broken through, or a veil lifted, so that the true self, the primordial spirit, may shine forth. In most traditions this breakthrough is held to be unattainable through the ordinarily acceptable good social life. A new, hardened, and disillusioned approach must be sought. The body and the mind, which are part or the whole of the impediment, have to be controlled, disciplined, and chastised; hence either asceticism or a set of psychophysical experimentations that differ radically from the normal acts of life is espoused.

Spiritual perfection. The quest for spiritual perfection is elitistic, even when, as within Christian monastic orders, humility is essential. Withdrawal from society is necessary, since the instrumentalities of perfection cannot normally be acquired and augmented in the surroundings of everyday life. The operational basis is meditation on a set of spiritual concepts that either represent the supreme value or provide crutches to support the body and the mind on their tedious journey toward whatever their supreme consummation may be. Processes of intense contemplation, often accompanied by physical exercise, constitute the mystical practice—*i.e.,* prayer, worship, incantation, propitiation, and various forms of self-abasement or self-inflation. All these are pursued in enormously variegated forms and degrees.

Forms of mystical practice

Emancipation of the self. *Salvation.* Seen cross-culturally, the ultimate purpose of the monastic endeavour is the attainment of a state free from bondage, both bondage and freedom in such endeavours being defined in theological terms. Most languages of cultures with monastic traditions have special terms to denote bondage and freedom; others use terms of common parlance that are then understood by members of society as referring to theologically adumbrated types of bondage and freedom. Thus, the term salvation in the Christian context means deliverance from the powers of evil besetting the person's body, mind, and soul. Varying concepts of salvation, liberation, and emancipation are generated by, or closely related to, the society's identification of an individual, the extensions of his body, mind, soul, and spirit, and his status within a larger universe.

The idea of salvation, like other such concepts, presupposes a specific cosmological view against which to frame the answers to the question—formulated or unformulated—"What is it that is bound and that can, should, or must be freed to achieve the most desirable state within or vis-à-vis the totality of things—*e.g.,* the cosmos, God, and other absolutes?" The question has spatial and temporal parameters; in some of the indigenous Indian religions, salvation can be achieved during one's lifetime, but whether this happens or whether the achievement is delayed into another existence is actually accidental to Indian notions of liberation. In the Judeo-Christian and Muslim world, salvation proper cannot be achieved as long as the body continues to exist. "Salvation" and its semantic equivalents, in other words, refer to the present *or* future in the South Asian religions but refer only to the future in the eastern Mediterranean creeds. The life of the monastic is a full-time seeking of salvation, in contradistinction to that of the "part-time" general believer.

Redemption. Redemption as deliverance from the spiritual effect of past transgressions may or may not be identical with salvation, though the terms are in many cases synonymous. In any event, as part of his monastic ambition, the monk seeks redemption from his sins or he intercedes, takes upon himself and works out the redemption of others. This is accomplished through personal sacrifice or many forms of self-mortification. The processes of mortification augment or stabilize the austerities required of the monastic and are part of the picture in all monastic traditions. The emphasis on the autocentric or the vicarious aspect of the quest for redemption depends entirely on the doctrinal framework with which the monastic identifies. In either case the dialectic of autocentric and vicarious emphases cuts both ways—in mortifying his own body and mind for the benefit of others, the monk also helps his own advancement along the spiritual path. When a Jaina monk (a follower of a 6th-century-BC Indian religious reformer, Mahāvīra) volunteers to lie upon a bed infested with vermin that suck his blood, he may do so for a client or a patron with a view to diminishing that person's karmic load (involving the doctrine that every deed, good or bad, receives due retribution), but at the same time he practices the monastic virtues prescribed for him as a monk. When a Franciscan monk (a follower of Francis of Assisi, the 12th–13th-century Christian monastic leader) serves the poor and the sick, he also improves his own virtues of service and humility, all of which are instruments for his own redemption.

Social and vicarious acts as means of spiritual advancement

Liberation. When liberation from cycles of birth and death constitutes the foundation of a belief system, as in the basic pattern of samsara (*i.e.,* an inevitable metempsychotic chain that can be broken only through supererogatory efforts of the meditational order) in the Indian traditions, the monks are the harbingers of the methods of liberation. In India and Tibet as well as in Southeast Asia, the monk always has been the centre of attention of the religious life far more than in the Christian world, in which he was and is marginal to the main ritualistic and ideological thrust. The importance of the monastic life in a religious system is related to its ideological content. Thus, if the state of life after salvation is transformed, but is yet basically a continuous type with the present life, then the monastic has less importance than he does in belief systems in which salvation means a totally different state, one that cancels finitude and eradicates all traces of separate individual existence.

Limited personal goals—*e.g.,* power or wealth—are aimed at by some Hindu monastics, but these are infrequent and atypical for the total monastic situation.

Social and institutional purposes. *Conquest of the spiritual forces of evil.* Social goals are interwoven with those of salvation in most monastic traditions, and there is a vacillating emphasis on one or the other depending on the founders' interpretation of the theological framework. The first Christian hermits of the Egyptian desert (*c.* AD 250–500), Anthony of Egypt, Paul of Thebes, Pachomius of the Thebaid, and others, were referred to as the "Desert Fathers." They presaged later monastic institutions and were the inceptors of cenobitism—literally "lying [*i.e.,* eating, sleeping, and living] together"—which was to be fundamental to all later Christian monastic institutions. Though the early hermits, mostly native Egyptian (Coptic) peasants, were inspired by the example of famous recluses, their rigorous monastic asceticism soon projected an orientation toward communal life that was related to paramilitary models encountered at one time or another in virtually all monastic traditions—the community that was viewed as composed of soldiers of the spirit. The Desert Fathers believed that they were to combat the forces of the devil tempting them in the desert. Much of this might have been antedated by the Jewish Qumrān community, which was located near the Dead Sea and is often identified with the Essenes, a religious group that flourished in the Judaean desert between 150 BC and AD 70. The Qumrān ascetics considered themselves to be the true, unpolluted priests of orthodox Judaism, decrying the Jerusalem priesthood that they characterized as defiled, spurious, and unclean, sullied by Hellenism and potentially heretical. This may have been the first instance of a monastic elite versus an urban sacerdotal establishment in which the interpretation of the canonical teachings was under dispute. Rigorous asceticism, communal prayer, and common work were the rule, though celibacy might not as yet have been part of that community.

The Qumrān community

Improvement of society. By and large, monastic institutions may have aided the progress of civilization, even though they often, and perhaps rightly, have been blamed for obstructing and retarding it. Monasticism as an instrument for the creation, preservation, and transmission of secular and religious traditions was pervasive in all those cultures that gave a special status to the cenobite institution. Its function as a propagating or proselytizing agent of the religious tradition, however, is by no means universal, nor even regionally determined. In the Mediterranean religions, as well as in those that originated in South Asia, there are discrete, uninterrupted, ordered sequences from the totally contemplative, hence nonproselytizing, to the teaching and preaching orders with virtually no scope for contemplation. The role of monks and their orders in the arts, sciences, letters, as well as in the pedagogical and the therapeutic social services, is thus discussed under the headings of the diverse monastic systems (see below).

Institutional centres for religious leadership. In some religions, monasteries become training centres for institutional religious leaders. There is, however, a clear dichotomy between training centres for ecclesiastical and for monastic leaders. Even though the distinction may seem to be blurred in the Roman Catholic and Eastern Orthodox traditions, the fact that monastic training could also be priestly should be viewed in a cross-cultural survey as accidental. In all Indian religions, by contrast, there is a radical and exclusive division between the priestly and the monastic careers and their concomitant institutions. The common denominator lies in the supererogatory status of the monastic life—if churches and seminaries prepare ecclesiastical leaders, teachers, and intellectuals, monasteries may train people to whom the same epithets apply, but with a difference: the monk is usually said to be more radical and less compromising than the ecclesiastic.

Priestly and monastic training

Other purposes. Apart from the ubiquitous redemptory, spiritual, and social goals of monastic systems, most of them condone peripheral goals of more or less mundane types. Thus, a Tibetan lamasery (monastic religious centre) is not only a centre of spiritual counsel but also a bank, a judicial court, a school, and a gossip centre for the laity. Some of the most specialized operations are found in monasteries, including coaching in wrestling, in some Hindu orders, and the preparation of perfumes by the Muslim Sanūsīyah (a conservative, rational, and missionary order established in the 19th century).

TYPES OF MONASTICISM

Organizational or institutional types. *Eremitic.* A taxonomy of institutional types includes several varieties: first, historically and in terms of simplicity, is the eremitic (religious recluse) type, including the early Christian hermits or anchorites; the actual or legendary *ṛṣis* ("seers") of Vedic India (pre-800 BC); some of the earliest Jaina *śramaṇas* ("ascetics"), particularly the semihistorical founders of Jainism (Mahāvīra and Pārśvanātha); the Taoist recluses of early southwestern China; and, down even to today, sporadic hermits in the various religious areas of the world—such as Gauribala in Ceylon (now Sri Lanka), La Mère in Pondicherry, India, and other Western converts to Asian belief systems without organized monastic trappings. Some of the European and American neomystics also should be included in this class.

Common to all true hermits and eremitic institutions is an emphasis on living alone, on a highly regularized contemplative life (with individually generated, often experimental spiritual disciplines), and on frequently idiosyncratic and even heretical interpretations of underlying scriptural or disciplinary codes. Incipient self-mortification and individual austerities can be traced, but these are incidental to the eremitic style.

Quasi-eremitic. The lauras (communities of anchorites) of early Christianity in Greece and Cyrenaica (exemplified by the Mount Athos tradition that exists even today), the small-scale ashrams (religious retreats) of monastic Hinduism from at least 300 BC to this day, and perhaps the semiformal congregations of the early Buddhist monks and nuns, preceding the establishment of the sangha (monastic order or community), are best called quasi-eremitic. Common elements in this type of monastic assembly are loose organizational structures with no administrative links to mother institutions and no external hierarchies. The category forms a transition between the eremitic and the cenobitic; in many cases, such as that of the medieval Indian *panthā* organizations (sectarian), groups may display eremitic and cenobitic features consecutively, either during different annual seasons or on the occasion of special conventions. In the 20th century some Nepalese followers of Gorakhnāth (8th century AD) live as recluses during most of the time but form themselves into a quasi-military association on certain occasions, such as during the all-Indian monastic assemblies (*kumbhamelā*) every sixth year at certain centres of pilgrimage. During these periods they become organizationally indistinguishable from the most highly structured cenobitic units present at the conventions.

Cenobitic. It is probably not wrong to identify proper "monasticism" with cenobitism. There seems to be a correlation between a formulated rule, or set of rules (regula in the Christian orders, vinaya and *śīla* as part of the Buddhist canon), and cenobitic institutions; eremitic and quasi-eremitic settings lack formulated rules and give more scope to the individual's self-imposed disciplines. A Christian ascetic, Pachomius of the Thebaid (*c.* AD 290–346), was reputedly the founder of organized cenobitism in the Western world and is said to have built nine monasteries for men and two for women. These monasteries may have been the model for the monasteries founded by the Greek theologian St. Basil the Great (*c.* 330–379), who set down what was the first monastic rule. The basis for all subsequent Eastern Christian (Greek) monastic institutions, it was simple and primitive compared to the regulae (rules) of the orders founded in later centuries in Europe. Avoiding the extreme austerities of the Desert Fathers, St. Basil's rule was strict but not severe. Its asceticism was an instrument in the consummate service of God; it was to be pursued in community life and in obedience. Liturgical prayer and manual and mental work were obligatory. Germinally, the Rule of St. Basil enjoined or implied chastity and poverty, though these were far less explicit than in the later regulae. What Basil's rule was for Eastern monachism, St. Benedict's was for early Western monasticism. Benedict of Nursia (*c.* 480–543/547) was

Rules of St. Basil and St. Benedict

a practical Roman, and his regulae (enjoining poverty, chastity, obedience, and stability), which was used—for the most part—by most of the paramilitary aristocratic orders, such as the Knights Templars, until the 13th century, and which is the rule of the Benedictine order today, incorporates instructions on institutionally held property. It also set the model for all subsequent Roman Catholic monastic orders—*i.e.,* the requirement that the individual monk does not own property, but that it is held by the order through its appointed trustees.

One-third of the Theravāda (southern, or so-called Way of the Elders) Buddhist canonical literature is vinaya (comportment), the Buddha's own statement of more than 200 rules for his monks. Compared to these, the later Brahmanic (Hindu) orders, such as the Sannyasi order founded by the Hindu reformer Śaṅkara (8th century AD), contain hardly any "rules" except an implicit renunciation of worldly desires, a detachment from society, and an indifference toward the "opposites" such as pleasure and pain. The 6th-century-BC founder of Jainism, Mahāvīra, about 80 years senior to the Buddha, formulated the nucleus of the Jaina doctrine and also established the core Jaina order, giving it a very elaborate rule that goes into minute regulations for monastic residence: restricting the itinerant monk's sojourn to one week at a time in a village and one month in a town; requiring that the monk not sleep more than three hours and that he spend the day and the rest of the night in expiation, meditation, studying Jaina scripture, and begging for alms. Some scholars believe that the Jaina rule provided the model for all monastic rules in India and thus for the monastic traditions in all the Asian countries that came under India's religious tutelage.

Indian and other monastic orders and groups

The Essenes of Judaism, regardless of whether they were identical with the Qumrān settlement, probably had a written rule; certainly, they were highly formalistic, emphasizing ritualistic purity, with ablutions prescribed for the members, and with a fundamentalistic adherence to the letter of the Jewish ritualistic and legal books Leviticus and the Deuteronomy written into their discipline. The so-called hippie communes of the 1960s and '70s, insofar as they sought religious experience, should be included in a historical list of cenobitic organizations; growing food, preparing and consuming it jointly, and sharing common dormitory facilities are essential elements of the cenobitic structure.

Quasi-monastic. Paramilitary, or quasi-monastic, associations are another type of monastic group. Most Christian orders of this type also had medical or healing commitments; non-Christian monastic orders of this type did not cater to the sick. The Knights Templars, whose order was founded in the 12th century during the Crusades—the most prestigious and most thoroughly defamed aristocratic organization—styled themselves "poor fellow soldiers of Christ and the Temple of Solomon"; their foundational commitment was the protection and the guidance of pilgrims en route to and in the Holy Land. Though the Templars, following the rules of St. Basil and St. Benedict, took vows of poverty, chastity, and obedience, their corporate wealth and the secrecy of their initiation and of all their internal affairs helped bring about their extermination under the French king Philip IV the Fair in the early 14th century. The military model was evident in their hierarchical structure—there were chaplains, knights, and sergeants under a grand master of the temple.

Christian military orders

The Templars were inspired by the Knights Hospitallers (or Order of the Hospital of St. John of Jerusalem), founded in the 11th century. The Hospitallers were probably the first to generate genuine medical and hospital services, at first for pilgrims to Jerusalem. They were the classical nursing order, and their first foundation was the Hospital Saint-Antoine-de-Viennois (*c.* 1100). There were other houses in Spain, Italy, and Germany. Their chief officers were ordained priests, but the majority of the members were nonsacerdotal "hospitallers," or lay brothers and sisters. They adopted the Benedictine rule until 1231, meeting under an elected master, and an annually rotating chapter-general assembling the "commanders"; the order switched to the "modern" rule of St. Augustine in 1247.

The Teutonic Knights (Deutscher Ritterorden), founded in Jerusalem in 1189, had a feudally independent relationship to Rome and the papal administrative bureaucracy (Curia) specially defined by more than 100 papal bulls; their grand master, who had the same rights as a prince of the Holy Roman Empire, was assisted by five "grand commanders." The organization was composed of knights (usually noblemen), middle-class priests, and sisters (mostly noblewomen), well trained in hospital services. After the fall of Acre to the Muslims in 1291, the order moved its headquarters to various places in Europe. But in the 13th century and after, the order revived its erstwhile function, when European rulers called upon it to war against the Altaic and the Prussian pagan peoples.

The popular, and wrong, identification of Tibetan monks as "lamas" has obscured the highly segmented institution of the Tibetan Buddhist clergy. Among the Khamba (*khams pa*) in eastern Tibet, men with minimal monastic initiation (*lung*) organized themselves as a military or police force to protect the unarmed higher initiated clergy as well as the monastic territory. These were not defenders of, or fighters for, Buddhism, but were a type of monastic police force, sartorially indistinguishable from the actual clerics. They were very evident in Tibet's confrontation with the Chinese communists (1959–65).

Buddhist, Muslim, and Sikh military orders

In the Islāmic world the mystical orders (Ṣūfī) and the partially overlapping dervish (*darvīsh*) assemblies constituted a living critique, as it were, of the formalistic, fundamentalist, and Qur'ān-oriented (scriptural) orthodoxy controlled by the *'ulamā'* ("teachers") from Cairo, the assemblage of Muslim learned ones who set standards for orthodox faith and practice. The Ṣūfīs' direct approach to divinity through such meditative or ecstatic practices as the *dhikr* (the chanting of the names of Allāh, or God)—that was accompanied by variegated physical expressions such as dances and songs and by the ingesting of drugs, usually of the genus *Cannabis* (such as hashish)—became symbolic representations of the criticism of officially sanctioned behaviour. The Turkish Bektashi (Bektaşi) excelled in poetry and in humorous repartee, which even now is recalled in references to the art of the Bektashi. The Sanūsīyah (Senussi) order of Ṣūfīs, in Libya and other northeastern African countries, not only antagonized the Wahhābīyah (a generic name for fundamentalistic orthodoxy in Islām rather than a term denoting any specific group) but also achieved impressive political and military corporate stature in very recent times, opposing the Italian colonialist forces in Libya with some measure of success. These orders sought communion with Allāh through mystical practices rather than salvation by righteousness. "Not I and God but only God" was one of their mottoes.

Though the religion of Sikhism is historically Hindu, the early "Pure" (Khālsā) did not encourage monasticism; Gurū Nānak, the founder of Sikhism (1469–1539), was a married man, and so were most of the other nine Gurūs ("teachers"). In the late 17th century, however, the Nirmal-akhāḍā was created in complete analogy to the celibate monastic orders of Hinduism and organized on the same principles. Underlying this development is a universal Hindu tendency to create monastic corollaries to lay teachings; the process has been repeated in India much more recently. The Arya Samaj (a Hindu reform society), founded by Dayananda Sarasvati in 1875, is a good case in point. Although Dayananda was a monk in the Daśanāmī Sannyasi order (Holy Men of the Ten Names), he discouraged monasticism—yet, bending to an all-Hindu cultural pressure, monks have been ordained in his organization since the early decades of this century. An older quasi-monastic organization among the Sikhs is the Nihaṅg Sāhibs, a basically military organization within Sikhism. Created to fight Muslim incursions into the Sikh communities of the Punjab, the Nihaṅg Sāhibs wear robes that are blue and yellow military uniforms that have been unchanged since the 17th century. The Nihaṅg Sāhibs are married, but during their temporary active service as *nihaṅgs* they abstain from sexual intercourse and live in a cenobitic manner.

Mendicant monks and orders. By a strict definition mendicancy (living by begging) would preclude cenobitism. In actual fact, however, there are many orders

that are mendicant and cenobitic at different times. The Hindu and Buddhist official orders are really both. During his training the neophyte lives in a strictly cenobitic setting; on subsequent peregrinations he eats the food he obtains by begging, which is part of his advanced discipline, and he eats alone. The Myanmar (Burmese), Thai, and Sinhalese Buddhist clergy could be termed mendicant stationary—the monks wander out in the early morning to collect food in their alms bowls, but they consume it jointly at the house in a cenobitic fashion.

The Christian mendicant orders

In the Christian world the Franciscans, founded by St. Francis of Assisi (1181/82–1226), with their numerous subsections, and the Dominicans, founded by Domingo de Guzmán (1170–1221), were and are the most powerful statutory mendicant orders. The synthesis of contemplation and the apostolic ministry is strong in these orders; the Dominican motto "to contemplate and to give the fruits of contemplation to others" is significant. The Sanskrit term *parivrājaka* ("walking around") quite literally connotes mendicant status and as a title is carried by a large number of Hindu monastic organizations. It has canonical sanction—the Hindu scriptural definition of a monk is "[one who] having renounced the desire for sons, for wealth, the fear of social opprobrium and the craving for social approval, he sallies forth, begging for food." Here also there is a blend between the contemplative and the preaching life; the various Hindu orders place varying emphases on the one or the other, a distribution that seems quite parallel to that of the Christian orders. The vow of chastity is spelled out for the Hindu mendicants, but poverty and obedience are implied rather than enjoined. The Hindu monastic organization is much looser than either the Buddhist or the Christian, and in this sense it resembles the earliest eremitic and quasi-eremitic types in Judaism and Christianity.

Other organizational or institutional types. Permanent versus temporary membership correlates with different monastic institutional types, though it does not seem to have any bearing on organizational structure. In the Theravāda Buddhist order (sangha) of Thailand, Myanmar (Burma), and Sri Lanka, men join a monastery for an unspecified period of time, with minimum periods (three months to a year) prescribed in the *Dhammayut,* the smaller and more highly ascetic of the two sections of the Siamese sangha; the Mahāssaṅghikas, who form the monastic majority, do not specify any such duration. Lifelong monastic views are, in those regions, a matter of individual choice. The order itself does not take any official stance on them. This differs radically from all full-fledged Catholic orders, as well as from those Hindu organizations that initiate members by the *virajā-homa* (*i.e.,* the Vedic rite of renunciation); since the initiate is declared dead by this ceremony, he cannot return to the world of the living (*i.e.,* to society). Dispensations, on the other hand, were given, though reluctantly, to Roman and Greek Christian monks and nuns who wanted to leave their orders. In the Hindu monastic code there can be no such dispensation—monks who leave and return to social life are highly stigmatized.

Hierarchical and status types. *Sacerdotal.* In addition to organizational and institutional forms, a typological survey must also include aspects of monastic status and hierarchy. The first and most important division here is between sacerdotal and nonsacerdotal full-time supererogatory specialists. Most of the canon-based (scriptural) religions of the world distinguish between priests and nonpriestly practitioners. In the case of Greek and Roman Catholic Christianity, the distinction is crucial on the sacerdotal but incidental on the monastic end. Monks who have priestly ordination are full priests and full monks; monks and nuns who do not have it are full monastic members nevertheless, sharing the same vows and the same discipline. Islām does not officially recognize monastic status, nor does it have priests—the imam is the leader in prayer, but he has no special vows or ordinations. The dichotomy does not apply in Judaism either.

In the case of the religions originating in India, the situation is markedly different: in Hinduism only a male person born into a Brahman (highest) caste is entitled to perform sacerdotal, Vedic (scriptural) ritual; this requires no further initiation than that given to all high-caste boys. A Hindu priest (*purohita*) must be married. Celibacy is not incumbent on all Hindu monastic orders, though it is in those of high prestige. But the monk cannot perform any sacerdotal service, even if he was born into a Brahman family. The fact of monastic ordination cancels his sacerdotal status. Monastic organizations ordain monks in various ways, and the types of ordination are numerous; but monastic and priestly ordinations are totally different and distinctive in type, scope, and intent. The Brahman priest supports and enhances the mundane well-being of his client and the worldly estate of his society through Vedic and other rituals. The monk, on the other hand, withdraws from the mundane in a radical sense by rejecting sacerdotal commitments, and he recommends such withdrawal to his clients in a long-range perspective.

Separations of sacerdotal and monastic functions in Hinduism

Secondary and tertiary orders. The notion of secondary and tertiary orders was generated in the Roman Catholic world, though by analogy it could be extended to non-Christian cultures. The triple division of the Franciscans and the Dominicans epitomizes this hierarchical and status type: the first order consists of ordained priests and lay brothers, the second of contemplative nuns; the third order incorporates laymen and laywomen, "tertiaries," who live under abridged, or "minor," vows that may include celibacy. In the Theravāda Buddhist world, these tertiaries have parallels in the sangha, which can be viewed as analogous to the first orders of Dominicans and Franciscans. Whereas the full-fledged Buddhist monk has more than 200 vows, part-time monks (*śramaṇas*) have fewer than one-third that number. In Myanmar, quasi-monastic but unordained practitioners (*upāsakas*) may stay at monasteries and participate in the meditative and congregational activities of the monks for a limited period and with payment of a nominal fee to the bursar of the cloister.

Status of nuns

In all monastic traditions of the world, the status of nuns is considerably lower than that of monks. The only conceivable exception is that of certain famous saintly women in Hindu India, today and in the past; some of them, known for their extreme piety or, more importantly, for their physical-mental (yogic) and mesmeric (hypnotic) powers, have gained high charismatic (spiritually influential) status that may place them, as individuals, above male monastics. Yet there is no truly hierarchical superiority wherein an ever so exalted nun could have disciplinary powers over a monk or even over a male novice. Though the Roman Catholic tradition has refused equal status to nuns, because women cannot obtain sacerdotal ordination, the Indian attitude rests on notions of ritualistic impurity—women, being polluted through the menstrual cycle, never have access to the ritual complex owing to their innate defilement, hence their status is much lower—even though some noncanonical texts (*e.g.,* the *Bhagavadgītā*) assert spiritual, though not ritualistic, equality of women and men.

When women postulants approached the Buddha for admission into the order, he was reluctant; finally, when his disciples and sponsors had succeeded in establishing nunneries, the Buddha said that this step augured the decline of the order. This did not discourage women either then or later. Buddhist nunneries are not numerous, and their ratio to male convents does not exceed 1:20 in any of the Buddhist countries. The modal Buddhist monastic attitude toward the nuns is one of embarrassed silence except in Japan, where the general loosening of monastic rules has worked in their favour.

Tertiary orders

Tertiary orders in the Christian world were established by noblewomen who combined piety with pioneering medical knowledge and strong motivation to gear the latter to religious pursuits approaching the monastic in degree of dedication. Though the term tertiary did not originally contain reference to the sex of the membership, its selective denotation was well established by the 13th century, usually referring to women, often of aristocratic background, who led a saintly life in a cenobitic setting but were inspired by humanitarian ideals rather than by that of sheer contemplation. In a very real sense, women belonging to such groups were the first nurses; their tradi-

tion has been continued in all the Christian nursing orders extant today and is being emulated by some new non-Christian orders, such as the Hindu Ramakrishna Mission. Where there are male tertiaries, as in some segments of the Dominican and Carmelite orders, the humanitarian connotation dominates.

Though the hierarchical types in the Christian West must be viewed from an organizational or managerial vantage point, there is much variation in Eastern orders. In the religious world derived from the teachings of India, a true hierarchy comparable to the Christian orders is found only in the Tibetan ecclesiastic setting. Contrary to the popular notion, the lamas are not simply high-ranking monks but are viewed as incarnations of a particular aspect of the Buddha or of a teacher who in turn was such an incarnation. Though Tibetan monasteries prided themselves on the presence of one or more lamas, they really stood above and outside the operational hierarchy, and their function was and is advisory rather than executive.

VARIETIES OF MONASTICISM IN THE RELIGIONS OF THE WORLD

Since monastic systems developed mainly in the Mediterranean monotheistic, and in South Asia's theologically more complex situation, their diffusion into other parts of the Western and Eastern world can generally be viewed as a modification of the two historical core areas that were located in two relatively small regions—*i.e.,* the Judeo-Christian-Islāmic and Hindu-Buddhist-Jaina areas.

Religions of the East. *Hinduism.* With some reservations, the religion of the majority of the population of the Indian subcontinent after the decline and exit of Buddhism can be called "Hinduism," to supplant the erstwhile "Brahmanism" or the Vedic religion of pre-Buddhist days. Buddhism, like Christianity, is an export religion in the sense that it did not survive in an ecclesiastically organized fashion in its homeland. Hinduism, on the other hand, has absorbed so many Buddhist traits that it is virtually impossible to isolate the latter in medieval and later Hinduism. The most important Buddhist-inspired element in Hinduism is no doubt its monastic tradition. Where there were hermitages in ancient, pre-Buddhist India (such as the abodes of the Vedic "seers" [*ṛṣis*] and the *gurukula* [teacher's family]—which was a germinal cenobitic setting comprising the *ṛṣis* and their disciples), typically monastic features such as vows of chastity and an unequivocal rule of monastic comportment were not operative before the time of the Buddhist sangha in the 6th century BC and its little-known early contemporary movements, such as the Ājīvikas, which are viewed as proto-Jaina, and other incipiently monastic institutions.

Śaṅkara and Rāmānuja

The most outstanding Hindu monastic founders and thinkers, comparable in many ways to the Christian St. Benedict of Nursia or the great theologian Thomas Aquinas (1225/26–1274) with regard to their importance in their respective areas, were Śaṅkara (8th century AD) and Rāmānuja (11th century AD). Both of these teachers propounded the Vedānta theology (a religio-philosophical system concerned with the nature of ultimate reality), albeit in very different, mutually incompatible interpretations. Śaṅkara's order of Daśanāmī Sannyasi (Holy Men of Ten Names) was and has remained the monastic order that set the monastic standards for the rest of Hindu India. Based on a scholastic, nondualistic reading of the four "great dicta" (*mahāvākya*) of the canonical Upanishads (one of the Hindu scriptures), the monk, following the example given by the founder, meditates constantly on the numerical identity of his individual soul (atman) with the cosmic soul (*brahman*). All his other observances—such as group incantation of canonical liturgy, participation in the monastic assemblies with other orders (*kumbhamelā*) at various places and at astrologically computed time intervals, alms begging, teaching religious topics to the laity, and conducting scriptural discourse with lay and monastic scholars (*śāstrārtha*)—are ancillary to his main purpose, which is meditation. He does no social work, and there is nothing to parallel the humanitarian deeds of some other orders in the Indian and most orders in the Judeo-Christian world. He cannot conduct ritual, and he has no obligation whatever toward society; but society has its obligation toward him—it feeds and clothes him. For this, he instructs those who wish to be instructed in the methods of meditation leading to emancipation from rebirth, which is believed to be caused by one's good or evil thoughts, words, and deeds in one's previous existence. In a more formal manner, he may or may not initiate lay seekers and monastic postulants (potential monks) into meditation by imparting to them a mantra, a sacred secret formula aiding the emancipatory process. Since the monk's initiation entails the symbolic cremation of his body, he is not cremated at his death as is done in the case of lay Hindus but is interred or immersed in the river.

Most of the prestigious Hindu monastic orders follow this pattern, though their disciplinary codes are often radically different. Thus, the followers of Rāmānuja, referred to as Śrīvaiṣṇavas (worshippers of Vishnu and his spouse), are largely lay, high-caste Hindus. The monastic order relating to this tradition emphasizes ritual and worship of the personally conceived deity; its rules of celibacy, compared with the strict and unexceptional rules in the Daśanāmī Sannyasi order, are somewhat vague and flexible—so that, in theory at least, a person who claims the title of a monk could be a married man.

Celibacy in Hindu monasticism

Of the approximately 90 monastic orders in Hinduism, some 70 impose celibacy and a cenobitic rule on their ordained members. Others—such as the Dādū-panthīs (created by Dādū, an important Indian saint of the 16th century) and a number of other orders whose designation ends in *-panthīs* ("path-goers") of relatively recent origin (14th century and later)—follow specific theistic doctrines of medieval Hinduism. Unlike the Daśanāmī, who accept only Brahmans (highest-caste Hindus) into their order, the *panthīs* do not discriminate on grounds of caste in their recruitment. In fact, most of these orders can be viewed as anti-Brahmanic revival, or even rebellion, movements.

Jainism. Jainism, founded in the 6th century BC by Mahāvīra in reaction to Brahmanic Hinduism, has fewer than 12 million followers today, but, because of its mercantile predilection, it is a wealthy community that has given traditional and substantial support to its monastic organizations. The two main lay divisions of the Jainas derive their primary designation from the monastic setting, which is unique, in India and the West. The Śvetāmbara ("White-Clad") sect is so called because its monks wear a white robe and a white piece of cloth to cover the mouth, thereby preventing the inhaling and annihilation of microbes and insects, and carry a broom with which they sweep the ground in front of them as they walk so as to clear away insects and other living beings that would be hurt or killed by being stepped on. The Digambara ("Sky-Clad"; *i.e.,* nude) sect is so called because its monks used to go naked to signify their complete detachment from worldly things and social trappings. Under Muslim rule this custom was interdicted, and since then (15th century) the Digambara monks have been wearing the white robe, thus becoming sartorially indistinguishable from the Śvetāmbaras. The Jaina monk stresses mendicancy, extreme austerity, and detachment. Both Jainism and Buddhism were purely monastic religions—in fact the only ones in a cross-cultural perspective; the rules for the laity are derived from monastic rules. The founders of Jainism (Mahāvīra, the semilegendary Pārśvanātha), as well as the Buddha, directed their instructions to monks and postulants exclusively—the vows of the monks are more numerous and more intensive, but the way of life enjoined on the laity is simply an abridged monastic rule with more dispensations and compromise.

The sangha of Buddhism

Buddhism. The generic term for the Buddhist monastic order is the sangha; the terms connoting the order in all Buddhist countries are literal translations of the Indian word. Far more than in other monastic traditions of the world—with the possible exception of Jainism—the Buddhist doctrine attaches central importance to the order. The recitation of the "threefold refuge" formula that makes a person a Buddhist, either lay or monastic, imparts a pledge of allegiance to the Buddha, the dharma ("teaching"), and the sangha; most commentaries imply that each of these three constituents is equally important.

In northern Buddhism (*i.e.,* Mahāyāna) the role of the historical Buddha has been and is negligible, and the order has acquired an even more exalted position. Since the Buddha began every single one of his sermons with the address *bhikkhave* ("O ye begging monks"), fundamentalistic Buddhism is tantamount to monastic Buddhism. The monastic discipline of the Buddhist clergy varies widely in the different parts of the Buddhist world. Ideally, the rules are laid down in the vinaya (monastic rules) portion of the Buddha's sermons, but monastic traditions and regulations have been influenced by the ecological and cultural conditions of Buddhist areas. Rules of vicinity or distance from lay settlements had to be interpreted and implemented differently depending on whether tropical, moderate, or (as in the case of Tibet and Mongolia) subarctic climatic situations are involved. Although celibacy is postulated for the Buddhist clergy everywhere, there have been and are liberal exceptions. The married monks of pre-20th-century Ceylon (Sri Lanka) and those of some of the Japanese Buddhist clergies were highly evident exceptions. Since the vows of the Buddhist monk in principle are not permanent, the theoretical emphasis on celibacy became academic in many parts of Asia. In the Theravāda ("Way of the Elders") countries of South and Southeast Asia, the monks were and still are teachers to the people—not only of religious matters but also of basic education—particularly in Myanmar. There appears to be a correlation between the regional stress on educational and other forms of monastic involvement with the lay society and the provision of special amenities for monks who prefer a strictly contemplative life, as in Sri Lanka and Thailand; conversely, where the involvement is not spelled out, as in Japan, no such special division is apparent. The differences in living style between the northern (Mahāyāna, or "Greater Vehicle") and the southern (Theravāda, wrongly called Hīnayāna, or "Lesser Vehicle") monastic institutions are quite radical. The fundamental feature, however, is meditation (Sanskrit *dhyāna,* Pāli *jhāna,* from which is derived the Chinese Ch'an and Japanese Zen). The path of meditation leads positively toward the intuitive understanding of momentariness—the condition of existence—or negatively toward the total rejection of all notions of permanence. The negation of existence is described by the Sanskrit term *nirvāṇa* ("fading away")—Pāli *nibbāna;* Tibetan *thar pa* ("crossing over")—which is identical with the Zen term *satori,* a translation of the Sanskrit *bodhi,* (enlightenment), which was the discovery of the Buddha. All these terms are operational synonyms.

Functions of Buddhist monks

Sikhism. Sikhism, founded by the Punjabi reformer Nānak, was and is a martial version of Hinduism. Of all the indigenous Indian religions, it was the least sympathetic to monastic inspirations. The aforementioned Sikh monastic Nirmal-akhāḍā and the quasi-monastic Niṅaṅg Sāhibs are patently compromises with the overall Indian trend to establish monastic traditions to express full-time involvement with redemptive practice. During the past two centuries only, the monastic Udāsī order (founded by Nānak's elder son Śri Chand) has achieved a most successful rapprochement with Hindu elements. Its disciplinary, sartorial, cenobitic settings are identical with those of the Hindu Sannyasi. They refer to the *Ādi Granth,* the sacred book of the Sikhs, as their basic text, in spite of the fact that their intramonastic and intermonastic discourse proceeds on Sanskritic, Vedāntic (religio-philosophical) lines similar to those of the orthodox Hindu orders. This accounts for the fact that the Udāsī are now respected as equal to the most prestigious and ancient Hindu orders.

Taoism. Taoism, a Chinese religion based on folk religion with Buddhist influences, with some very weak emulation in Japan, holds a middling position with respect to monastic ventures, lying somewhere between the powerfully antimonastic Confucian schools that always represented the official culture and mainstream of sophisticated Chinese opinion and the radically monastic Buddhists. Some scholars believe that since Taoism originated in the southwestern parts of China, Indian influences are conceivable. The chief object of Taoism, however, is not the redemptive, salvational purpose evinced in other scripturally based religions, Eastern and Western. The ultimate aim of the Taoist sage was longevity or ultimate physical immortality rather than salvation from this world or the self. The Taoist quest after the elixir of life, and its expression in cryptic and enigmatic poetry that is well known to, and generally misunderstood by, European and American seekers of the mysterious East in the 20th century, is in no way comparable to the supererogatory search of the monastics thus far discussed. The Taoist settlements of sages, in forests and mountain glades as well as in the cities, are, at best, analogous to the eremitic type of protomonasticism. When Taoist settlements were cenobitic or celibate, these features were indeed incidental to Taoism, which defies and rejects rules of any corporate kind.

Aim of the Taoist sage

Other Eastern varieties. Monastic orders are thus either of Mediterranean or of Indian origin. The connections between the monasticisms of the two regions are immensely complicated, since the Indian subcontinent has absorbed or generated traditions that, though abhorrent to the orthodox Islām of the Arabic *'ulamā',* are nevertheless Muslim rather than Hindu in origin. North Indian Hindus have felt a strong attraction to Ṣūfī (Islāmic mystical) poetry, and dozens of Ṣūfī saints are worshiped by Muslims and Hindus alike. With India as one of the centres of monastic diffusion, new quasi-monastic types can be expected to spring up in the present or near future. In a strictly descriptive account, leaving aside historical development, the Ṣūfī orders must be seen as separate from both the Indian and the Mediterranean backgrounds.

Of the slightly less than 100 monastic and quasi-monastic orders in South Asia, well over half are local and regional growths. They usually lack even a body of rules and conventions that would be recognized or accepted by a wider Hindu-Buddhist-Jaina consensus. About a dozen orders are rejected and feared as heretical, outside the pale of the acceptable, and charged with using religious pretexts to indulge in antisocial behaviour. The Hindu and Buddhist Tantric sects (practicing occult, esoteric, meditative techniques) represent esoteric countermonasticism in India, though they have been accepted fully in certain Tibetan Buddhist hierarchies.

South Asian quasi-monastic groups

Of the not numerous but clearly evident monastic or quasimonastic organizations of recent origin in other parts of Asia, the Vietnamese Cao Dai achieved some recognition. They had a military organization and their own army "regulars" up to the mid-1950s; an eclectic theology with such heterogeneous patron saints as the 19th-century French novelist Victor Hugo, the World War II British prime minister Winston Churchill, and the Buddha; and a monastery-fortress in south-central Vietnam. The members were bound by vows that did not include celibacy or poverty but stressed obedience to the hierarchy.

Religions of the West. *Judaism.* Judaism and Islām, the two Western religions that assume a father-oriented social hierarchy as the supreme secular value, did not generate monastic traditions as part of their official institutions. The Essenes of the Qumrān community were, in their own vision, inimical to the ecclesiastic centre and marginal to the official Judaic complex. The weak eschatology (doctrine of the last times) of Jewish theology might account for the lack of a lasting monastic quest, which is typically inspired by individual salvational expectations.

Islām. In a somewhat milder fashion, this holds for Islāmic orders as well. Whereas some Jewish prophets might have had celibate leanings, the Prophet Muḥammad discouraged celibacy. Non-Arabic Islām, nevertheless, generated monastic orders, "Ṣūfism" being the generic name for a monastic attitude rather than for monastic institutions. The Bektashi (see above) and the Sanūsīyah (conservative order founded in the 19th century) are typical for the marginal status of monastic settings in Islām; vestigial rules and formalized vows are discernible, but the main thrust of these monastics was of an interpersonal type, centring on the relation between the individual teacher of esoteric wisdom (*murshid*) and his disciple (*murīd*) and on the practices of chanting and meditation on the secret or known names of God (*dhikr*) and of other ecstasy-producing methods. The "way" (*ṭarīqah*) meant something that, by implication, was not accessible to the pious, orthodox Muslim alone. The Naqshbandīyah order, which

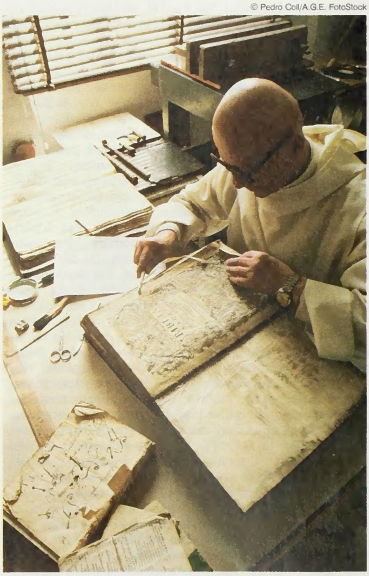
Ṣūfī orders

originated in Turkic-speaking areas of southwestern Central Asia, became widespread in the Islāmic Middle Ages, rediffusing from India into the western reaches of the Ottoman Empire (14th–20th centuries). The actual or alleged ingestion of *Cannabis* drugs (such as hashish) and the nonconformist, antilegalistic doctrines of the order have made it attractive on the popular level. The ritualization of the esoteric, contrasted with that of the social and the civil in the official Muslim prayer, seemed to provide an outlet and an alternative for a large number of devout but nonconformist Muslims, much as the late-20th-century cultic movements (such as spiritualist, hippie, and similar groups) do for the religiously alienated in the West. Nonconformity to the official doctrine was often enhanced by unexpected or deviant behaviour; thus, the Sanūsīyah brethren prepared and used a variety of perfumes for their personal toilets. The element of rebellion, frequently manifested in eccentric behaviour, is typical for a setting where the official religion is antimonastic, as is the case of Islām.

Christianity. Since the terms monk and monastic are historically and etymologically Christian terms, the ideal-type monasticism is the Christian type by a semantic fiat—thus, the denotation of monastic terms has been extended to religions other than Christianity, the lexical donor for monastic terminology.

An overall view of Christian monastic history reveals a shift of emphasis from the contemplative to the socially active; this shift, if observed in a geographic dimension, moves roughly from East to West. The Eastern Orthodox and other Greek-liturgy-based sections of Christianity generated highly meditative orders, the Mount Athos (Greece) complex being the most famous among them. The large variety of Roman Catholic orders displays eclectic emphases: the Benedictines, Franciscans, and all the orders designated as "minor" (in the Latin sense of humble or modest, rather than hierarchical or organizational) emphasize meditation; the Dominicans should be called "major"—though they are not—because the areas of preaching, scholastic continuity, and evangelizing rank far above contemplation in their order; and the Society of Jesus, or Jesuits (founded 1534), stands on the other end of the contemplative–social-centred continuum—teaching, social work, and the active life being regarded as the quintessence of supererogatory piety that is so important to this powerful organization.

© Pedro Coll/A.G.E. FotoStock



A Benedictine monk restoring incunabula at the monastery of Monte Oliveto Maggiore, Tuscany, Italy.

Viewed from the social-centred end of the continuum, certain institutions in the Protestant tradition should be called monastic. The Taizé (France) communities of the Reformed Protestant tradition, founded in the 1940s, initiated an ecumenical monastic movement. Where emphasis is on faith rather than on works, monasticism would seem to be a viable expression of the Protestant tradition; yet, owing to a set of historical accidents whose ideological summation was described in *The Protestant Ethic and the Spirit of Capitalism* by the German sociologist Max Weber (1864–1920), the operational emphasis has been on active engagement in the world rather than on seclusion. This explains why the various kinds of part-time Protestant retreats (linked to places especially designated for recuperation from the Christian involvement in work) are often placed in rural settings, and, where they are urban, there is usually an apology at hand. Professional Western monasticism, by contrast, never made much of the rural-urban division, with important monastic centres randomly built in rural and urban settings; it is only the eremitic and some quasi-monastic types that insisted on rural or sylvan environments.

Ecumenical monastic groups and retreats

MONASTICISM IN THE 20TH CENTURY

A discernible trend can be noticed in modern Western monasticism. It appears to be directed toward an enhanced individual liberalism, covertly or overtly condoned and granted to monks who wish to reinterpret fundamental traditions in the light of individualistic experience. If the work of Teilhard de Chardin, the Jesuit paleontologist, and of the Trappist monk Thomas Merton are in any way typical of this trend, it may be a clue for the future of Christian monasticism. Even though the disciplinary bases of the standard religious orders are likely to remain conservative for a very long time, there is much more ideological permissiveness. Individual monks and their disciples have been interpreting the monastic ideal in psychological, anthropological, phenomenological, and existentialist terms. It is perhaps the Christian monks of the future rather than nonmonastic priests and ministers who will seek a trans-Christian ecumenism through religious experimentation. Of significance is the fact that Thomas Merton was killed in an accident while in Bangkok to visit the Dalai Lama (leader of Tibetan Buddhists). Monks had been the religious entrepreneurs of early Christianity; they may well emerge as innovators in an era of infinitely greater interreligious information and diffusion.

Sponsored largely by college-campus religious organizations (both Protestant and Roman Catholic), a retreat pattern based on ecumenical concerns has been emerging in North America and in Great Britain during the second half of the 20th century, though much less so in other parts of the Western world. There were roughly 1,000 such para-academic, campus-sponsored organizational retreats in North America alone in 1968. Most were Protestant, but with an increasing participation by Catholic and Jewish groups; almost all the participants were of college age. No vows of any sort are enforced or even recommended, but there is emphasis on meditation, discussion, and religious reflection. The projection of ideas about the future of monasticism is easier to discuss than that of the future of religion in general. After the end of World War II (1945) there was a highly visible resurgence of monastic ideals and an overall increase of monastic organizations in all parts of the world outside the communist-dominated countries. The models were admittedly Christian, particularly Jesuit in the case of neo-Hindu orders—such as the Ramakrishna Mission (an eclectic Hindu order founded in the 19th century), which has over a dozen well-established centres in America and Europe. One Hindu monk, a swami—which correctly means an ordained Hindu monk—presides over each of these centres, often assisted by a younger monk; in theory, these centres train monks in the sannyasi tradition, but they actually serve a European and American

laity committed in various degrees to the Vedānta theology. On the model of the Ramakrishna Mission, there are some two dozen other India-originated organizations of this quasi-monastic or semimonastic type that have spread over the Western world, some of them with considerable wealth and popularity. Among such groups are the Self-Realization Fellowship, founded by the late Swami Yogananda Paramahamsa, with its headquarters in Los Angeles; and the Hare Krishna movement with chapters all over North America and western Europe. Founded by A.C. Bhaktivedanta, Swami Prabhupāda, it attracts and assembles young men and women who temporarily dress in the monastic ochre robe of Hindu monks and chant the names of the Lord as incarnated in Rāma and Kṛṣṇa (Krishna), accompanied by simple, arcane dance movements. Zen Buddhism has been established in North America and Europe, with Zen churches and centres in many major cities. Again, it is usually only the leader and the teacher who is a Japanese Buddhist monk, instructing a variously interested laity.

Infusion of Oriental groups in the West

Since the 1960s much attention has been given to the monastic potential of the communes, the hippie and kindred congregations in various parts of the Western world, some of whom have already established settings that have many genuine monastic traits—though with none of the foundational vows of traditional monasticism, Eastern or Western. The growing concern with mystical experience since the mid-20th century, with or without the use of psychedelic ("mind-expanding") drugs, and the general alienation from the official religions parallel many, if not most, of the monastic creations of the past 2,000 years or more.

In the countries of non-Christian monastic traditions (*i.e.,* India, Sri Lanka, Myanmar, Thailand, Laos, Vietnam, and Japan), monasticism is, in varying degrees, on the decline. If the monastic distribution of the future could be assessed, an intelligent appraisal might be that monastic commitments and organizations will increase in the West and decline in Asia, perhaps in direct proportion to the changing religious concerns of the societies involved—the West, often believed to have had a surfeit of material progress, might seek the spiritual; the East, tired of contemplation, might turn to the pursuit of material progress on the former Western model. (A.Bh.)

Sacred kingship

Sacred kingship, a religio-political concept by which a ruler is seen as an incarnation, manifestation, mediator, or agent of the sacred or holy (the transcendent or supernatural realm), traces its origins to prehistoric times and, at the same time, exerts a recognizable influence in the modern world. At one time, when religion was totally connected with the whole existence of the individual as well as that of the community and when kingdoms were in varying degrees connected with religious powers or religious institutions, there could be no kingdom that was not in some sense sacral.

Among the many possible kinds of sacral kingdoms, there was a special type in which the king was regarded and revered as a god—the god-kingdom, a polity of which there were three forms: preliminary, primary, and secondary.

Forms of the god-kingdom

The preliminary form exists in primitive cultures in which the chieftain is regarded as divine. The primary form was the god-kingdom of the large empires of the ancient Middle East and East Asia, of ancient Iran, and of pre-Columbian Meso-America and South America. The secondary form occurred in the Persian, Hellenistic (Greco-Roman cultural), and European empires. Between these three forms there are many transitional types.

PRINCIPAL SCHOOLS OF INTERPRETATION

The phenomenon of sacred kingship was known and described in ancient times by various travelers (*e.g.,* Aristotle, in the 4th century BC, and the 1st-century BC Greek geographers and historians Strabo and Diodorus Siculus). In more recent times Sir James Frazer, a British anthropologist, introduced the study of sacred kingship in *The Golden Bough* (1890–1915). Taking his comprehensive material from ethnological reports and studies, Frazer concentrated on the preliminary stage. With the discovery of texts in cuneiform writing in Mesopotamia and Asia Minor, however, a new stage of research began. Called Pan-Babylonism by some scholars, the theories based on the results of these discoveries placed the god-kingdom of the ancient Middle East in the foreground.

Building on the thesis of Pan-Babylonism that a homogeneous Middle Eastern culture existed and on the theories of cult as a ritual drama, the so-called British and Scandinavian cult-historical schools maintained that the king, as the personified god, played the main role in the overall cultural pattern. The English branch of this school (the "Myth and Ritual School") concentrated on anthropological and folklore studies. The Scandinavian branch (the "Uppsala School") concentrated on Semitic philological, cultural, and history-of-religions studies. It is represented in the latter part of the 20th century by Swedish historians of religion who have theorized that, for the entire ancient Middle East, certain cult patterns existed and that behind those cult patterns lay the sacred-king ideology.

The "Myth and Ritual" and "Uppsala" schools

Rejecting the cult-historical school theory of an unchanging cult pattern and an unchanging sacred-king ideology, many scholars in the latter part of the 20th century have tended to emphasize individual research of case histories. The fundamental differences between the kingdom ideology in Egypt and that in Mesopotamia have been investigated by historians, and questions concerning sacred kingship in the Old Testament by Old Testament scholars. One result of such scholarly research is that the theory and practice of sacred kingship—in a history extending over thousands of years—has undergone immense changes and, thus, because of these widespread and extensive differences, all generalizations and categorizations are difficult to maintain. Though there may be an amazing correspondence among numerous individual phenomena, each individual sacral form of government can be explained only in its own historical, social, and religious context.

STATUS AND FUNCTIONS

The sacred status of kings, leaders, and chieftains. Basic to an understanding of sacred kingship is a recognition that the exercise of power of one person over other persons or over a community (local, regional, or imperial) in early times was general and not divided. Power could be exercised by only one person—one who simultaneously had the necessary physical (individual and corporate) and spiritual (psychic) strength and influence—over both people and objects. Because he was ruler over a community, the king's power extended to everything pertaining to the life of the community. Only gradually did a division of these powers develop.

The sacral status of the ruler differs in form and origins. Three main forms can be distinguished: (1) the possessor of supernatural power, (2) the divine or semidivine king, or (3) the agent of the sacred.

The possessor of supernatural power. The ruler may be viewed as the possessor of supernatural power—both beneficial and malevolent—needed to maintain the welfare and order of the community and to avert danger and damage. In primitive, or preliterate, societies he represents the life force of the tribe, in which worldly and spiritual or political and religious spheres are not distinguished. Concentrated in the chief is the common inheritance of the magical power of the community, and his authority is based solely on the possession and exercise of this supernatural power. The impact and comprehensiveness of such power wielded by a chief, for example, reaches into all areas of life of the tribe: provision of food, fertility, weather, all forms of communal life, and protection against enemies and misfortune. Because the supernatural (magical) power of the chief is identical with his own life force, the chief (or king) of such a society is not allowed to have any physical defects. With the dwindling of his own physical powers (illness, graying of hair, and loss of teeth), his own power to maintain and secure the common welfare and his own ability to rule are believed to be correspondingly diminished.

The basis of authority in tribal communities

This form of sacral status is found mainly among rulers

over one tribe (or several) in primitive cultures—he may be a chief, medicine man, shaman (a religious personage who has healing and psychic transformation powers), or king (as, for example, a rainmaker-king in Africa)—in which a fixed definition or limitation of such functions is not possible.

The divine or semidivine king. In some societies, especially in ancient kingdoms or empires, the king was regarded as a god or identified with some god. In early Egypt he was identified with the sky god (Horus) and with the sun god (Re, Amon, or Aton). Similar identifications were made in early China and early Erech in Mesopotamia. In the Turin Papyrus (a list of kings written *c.* 13th–12th centuries BC), the sun god Ra is viewed as the first king of Egypt and the prototype of the pharaoh (the god-king). The symbol of the sun circle, one of the most prevalent artistic representations of the sacred king, and the practice of addressing the king as "my sun" are well depicted in rock reliefs and inscriptions in areas ruled by the Hittite kings. The Persian king was regarded as the incarnation of the sun god or of the moon god. In addition to sky or sun deities, the sacred king also has been identified with other gods: the town god (Mesopotamia), the gods of the country, the god of the storm, and the weather god. Generally, however, the king was not identified with a specific god but rather was regarded as himself a god. As the incarnation of all that is divine, the Egyptian pharaoh was addressed simultaneously in inscriptions as Aton, Horus, and Re (sun deities). Significant for Egyptian royal theology was the doctrine of the god-kingdom spanning two generations; each king ruled as King Horus and became Osiris (the father of Horus, a fertility god and later god of the dead) after his death.

A broader foundation for the divinity of the king is the view of the king as the son of a god, which can take on different forms. The first king has been regarded as a god and his successors as sons of the god in a number of societies—in Africa, Polynesia, Japan (where the emperor, until the end of World War II, was revered as a descendant of the sun goddess), Peru (where the inca, or ruler, was believed to be a descendant of the sun god), Egypt, Mesopotamia, and Canaan. Because he personifies the divine national hero (as among the Shilluk in Africa), the king can demand divine status, a practice that was taken up in the Greco-Roman world by Alexander the Great and by the Roman emperors. When a king who has been sired by a god or when a god who takes on the external form of the living king approaches his queen, he begets the future king—the queen is thus called the mother of God. An essentially different foundation is the king's divine sonship through adoption, as, for example, in the legend of King Sargon of Akkad in ancient Mesopotamia. The adoption of the crown prince by a god is often part of the coronation ritual, especially in Mesopotamia: the god declares the king as his son when he ascends the throne.

The king as a god's image

An especially frequent expression of the relationship of the king to divinity, in Egypt and Mesopotamia, was that of the king as a god's image. In Egypt the king—addressed by the god as "my living image on earth"—is shown in the likeness of Re, Aton, Amon, or Horus. In Mesopotamia this kind of description was rare.

The myth of divine birth, such as that of Romulus, the legendary founder of Rome, in many places served to legitimatize the claims of the king. Unusual natural phenomena, such as an especially bright star, are sometimes connected with the birth of a divine king.

The conception and practice of making a king divine after his death are very old and widespread. Probably connected with ancestor worship, deification is practiced most often when the living king, although connected with gods, is not regarded as a god in the fullest sense. Only after his death does he become god. Among the Hittites, for example, the expression "the king becomes a god" means that the king has died.

The king as the principal agent of the sacred. In addition to the conception of a king as the incarnation of supernatural power and the possible equality of the king with the divinity, there is also a widespread belief that the king is the executive agent of a god. As the servant of a god, he carries out the work of the god on earth. The divine character of this form of sacred kingship is connected not so much with the individual king as with the institution of kingship. In this emphasis on the institution of kingship lies the difference between kingship in Mesopotamia and Egypt and in India and China. The institution was emphasized in Mesopotamia and China. Sharp distinctions cannot be drawn between the different conceptions of the relationship of a god to kingship. Despite all the different expressions of kingship in the history of Mesopotamia (especially among the empires of Sumer, Babylon, and Assyria), there nevertheless was a continuous theme: the real lord of the city, the country, or the state remains the god, and the king remains in a subservient relationship to him. Even when the king possessed or disposed divine power and had sacral character and sacral duties, he remained subordinate to the god who selected him and put him into his regal position. The king had a mediating position between the gods and man, especially in his significance for the cult (thus, Sargon of Akkad is first described in inscriptions as deputy of Ishtar). The king also had a similar status as agent in Mongolia, where it was believed that the king came from heaven and was enthroned by God to carry out his will.

Functions of the sacred king: the king as the source of cosmic power, order, and control. *The king as bringer of blessed power.* The usual function of a sacred king is to bring blessings to his people and area of control. Because he has a supernatural power over the life and welfare of the tribe, the chief or king is believed to influence the fertility of the soil, cattle, and man but mostly the coming of rain. He has power over the forces of nature. Where rain is vitally necessary for the welfare and continuity of a people, the king can be described primarily in terms of this special function. Protection against evil of all sorts also is important for the welfare of the country. If the tribe or the country is beset by misfortune, epidemic, starvation, bad harvests, or floods, the king can be held responsible. Sometimes the king is believed to have the power to heal sickness by means of touch or contact with his garment.

The king's influence over men and nature

The function of the king as bringer of good fortune is prevalent in primitive cultures. Especially prevalent in Africa, it also has been observed in Polynesia, Scandinavia, and ancient Greece. This power to bring good fortune is also an aspect of sacral kingship in advanced cultures, such as in India, Iran, China, Japan, pre-Columbian Mesoamerica, Egypt, Mesopotamia, and Canaan. The difference between Egypt and Mesopotamia is significant: in Egypt the pharaoh was the direct dispenser of all good fortune in the country, whereas in Mesopotamia the king mediated for good fortune through cultic speeches and actions.

The function of the king as dispenser of good fortune has had an amazingly long influence: the English king was believed to have had healing power over a special disease (the king's evil) until the time of the Stuarts in the 17th century, and until the 20th century a folklore belief persisted in Germany that the ruler has influence over the weather ("emperor weather"). Words sometimes used to symbolize the king as the wielder of beneficial influence are gardener, fisherman, and shepherd.

The king as shepherd. An Egyptian pharaoh once said of himself: "He made me the shepherd of this country." In Mesopotamia the description of the king as a shepherd was quite frequent; in the 3rd millennium BC the term was applied to Sumerian city princes (*e.g.,* Lugalbanda in the 1st dynasty of Uruk [Erech]). The function of the king as shepherd also has been noted in India. The image of the shepherd expresses the most important functions of the king—he provides his people with food; he leads them and protects them from dangers and, at the same time, shows his superiority over them. Christ's description of himself in the New Testament as the "good shepherd" is, in a sense, a description of his official position in the Christian church, which also describes him as king, prince of peace, and Lord.

The king as judge. From earliest times, in addition to other functions, the chief was the judge of his tribe; he personified the protection that the community provided

The king as the personification of protection

for the individual. Providing for a balance of power in the community, mediating quarrels, and protecting individual rights, the chief or king was the lawgiver and the highest administrator for all community affairs. The *ensi,* the lawgiver and the highest judicial authority in the Sumerian city-state, was responsible for order. In Egypt the king was the highest judge, the guarantor of all public order, the lord over life and death. Early Egypt and India developed a high degree of justice that described the activities of the king as *ma'at* in Egypt and dharma in India. Both conceptions may be expressed as "justice" or "order" but actually are more comprehensive. Because the king preserves the god-given world order, the task of being just has been viewed as one of his fundamental functions. The pharaoh of Egypt and the emperor of China were believed to be responsible for the maintenance of cosmic as well as social order.

The king as warlord. Belief in the supernatural power of the ruler caused him to be viewed as the protector of his tribe or his people from enemies. On the one hand, he was the chief warlord and decided on questions of war and peace (as in ancient Sumer). The Egyptian pharaoh was represented, in his divine capacity as warrior, in larger-than-life dimensions. He alone was regarded as the one who triumphed over the enemy. On the other hand, there was the concept that the king, because of his sacral character, should not personally take part in war. These concepts existed, for example, among the Persian kings.

The king as priest and seer. Religious duties quite often are connected with the office of chieftain, who is also priest or seer and rainmaker—all in one. Correspondingly, in nontribal societies, cultic functions belong to the office of the king. In the 3rd dynasty of Uruk, Lugalzaggisi is described as king of the country, priest of the god Anu (the god of the heavens), and prophet of Nisaba (goddess of grasses and writing).

Relationship of kings to ritual and divination

When a division of functions evolved, the intrinsically royal priestly and other cultic functions were transferred to priests, seers, and other servants of the cult; the old concept of the king as priest, however, survived in some fashion for thousands of years. The Egyptian king was the chief priest of the land and the superior of all priests and other cult functionaries. In many images he is portrayed as presiding over the great festivals and bringing offerings to the gods. Later, priests carried out their functions as his representative. In Mesopotamia the king was viewed as the cultic mediator between god and man. As head of all of the priests of the country, he had important cultic functions at the New Year's festival. In critical situations, the king might issue an oracle of blessing; through him the land would be promised salvation, which was often accompanied by the words, "Fear not!" The Persian king performed the sacrifice at the horse offering and was also the "guardian of the fire." In all questions of religion he was the highest authority; he was also the most cultivated of the magicians. The king in Ugarit (in Canaan) also carried out priestly functions and as prophet was the receiver of revelations. Like other ancient Middle Eastern monarchs, the Hittite king was the chief priest.

The relationship between sacred kingship and priestly cultic functions has extended over widespread geographic areas and historical eras: East Asia, China, Japan, India, Europe (among the Germanic and Scandinavian kings), Africa (in the great empires), and Madagascar. Sometimes, the division of functions brought about a transfer of the royal title to those who carried out cultic functions. In Africa, from the earliest times, there was a type of king who was called lord of the earth; he originally combined political and cultic functions but, with changing times, retained only the cultic ones. The strict separation of the priestly office from that of the king, as in India, where king and priest belong to different castes—Kṣatriya and Brahman, respectively—is an unusual exception, however.

The king may be the recipient of a direct revelation of the will of a god. Thus, in Egypt the pharaoh received a divine oracle through dreams in the temple (a practice known as incubation). In Mesopotamia the duty of the king to ascertain the will of the gods was more strongly emphasized; a directive of the gods could result from



A larger-than-life Ramses II towering over his prisoners and clutching them by the hair. Limestone bas-relief from Memphis, Egypt, 1290–24 BC; in the Egyptian Museum, Cairo.

Egyptian Museum, Cairo; photograph, O. Louis Mazzatenta/National Geographic Image Collection

omens, dreams, or reading the entrails of offerings. All major undertakings of the king were dependent on directives of the god, who was to be consulted in advance. A direct divine revelation to a king is related in the Old Testament I Kings, chapter 3, which tells of a dream of the 10th-century-BC Israelite Solomon in which he received the promise of the gift of wisdom. Likewise, in the Old Testament (in Genesis, chapter 41), Yahweh, god of the Hebrews, gives the pharaoh a directive in a dream.

The difference between ruler cults and sacred kingship

The king as the centre of ruler cults. Although a pharaonic cult occasionally existed in Egypt, the ruler cult differs entirely from sacred kingship, because it came into being from political impulses. The ruler cult, generally developed in a country or empire with many peoples and many religions, was one of the ruler's means of power. Syncretism, the fusing of various beliefs and practices, often succeeded in bringing together completely different religious and nonreligious motives. Alexander the Great (who established an empire of many peoples and religions), for example, revealed a conscious effort at continuity with the Egyptian kingdom, inasmuch as the oracle of the Egyptian god Amon at Siwah designated him as the son of Amon and thus the successor of the pharaohs. Among the *diadochoi* (successors to Alexander) of the first generation, the ruler cult remained limited, but, under Ptolemy II Philadelphus of Egypt (reigned 285–246 BC), it became an established institution that was connected with the deified Alexander. When the ruler cult was carried over to Rome, the emperor Augustus (reigned 27 BC–AD 14) allowed it to be practiced only in the east in connection with the worship of the goddess Roma—though he allowed it to be pursued with fewer restrictions in the newly conquered western provinces; the adaptation of honouring the divine Caesar (or emperor of Rome) soon became, however, an important expression of the unity of the empire. Serious resistance to the imperial cult was encountered only among the two radical monotheistic religions: Judaism (*e.g.,* against Antiochus IV Epiphanes of Syria in the mid-2nd century BC) and Christianity. The Hellenistic and Roman ruler cults never generated a strong religious movement. The sacrifices brought to the king (emperor) go back to the custom of bringing tribute

to the king (or chief) of primitive cultures. From this practice, the custom of bringing offerings to the deceased king developed.

Regal ceremonies. *Modes of selection and succession.* In the beginning, succession to rulership was not necessarily connected with the sacral kingship; the sacral king also could be elected or, through a power struggle, also could receive a divine, magical, or supernatural anointment. If the firstborn son of the king was not stipulated to succeed him or if the king left no children, severe struggles for the succession often occurred, generally resulting in a change of dynasty. The death of a king often was kept secret until the succession was assured, because of potential danger to the people and country. To counteract problems of succession, there were rituals to secure the continuity of the sacral power. In primitive cultures the successor of the dead chief was brought into physical connection with his predecessor, his utensils, or his clothing—he had to stay, for example, in the home of the dead chief and use his utensils. The funeral of the dead king took place after the new king was established in his office or just before his coronation (as in Egypt). Efforts to secure the succession show high regulatory standards; in Egypt a complicated succession theology linked the new and old king as Osiris and Horus. During the lifetime of his father, the crown prince could be designated as coregent. The designation of the successor often came through an oracle, a sign, or some other manifestation of the word of the god; Thutmose III of Egypt (reigned 1504–1450 BC), for example, reported how he was designated to the succession through the oracle of the god Amon. In ancient Iran, after an interregnum, the election of the king took place through an omen. The king was chosen by the god; sometimes he was described as divinely predestined in the womb of his mother as the ruler.

Regulatory standards of succession

Forms and types of sacred legitimation of kings. The coronation or ascent to the throne by a king is an official act that most clearly shows the sacral character of the kingdom. Until the 20th century two characteristics in the coronations of kings and emperors remained: through ascent to the throne, the king is placed higher than other men, and the act of accession is connected with supernatural powers. With this action a new era begins. This was expressed in Egypt and Mesopotamia in two acts that marked the beginning of the government of the new ruler. First, upon the old ruler's death, the crown prince took control of the government and soon thereafter established his accession in a festive celebration. The coronation, however, generally had to coincide with a new beginning in nature, such as the New Year's festival. The coronation also was viewed as a cosmic new beginning. The most important initial actions of sacred kingship—the ascent to the throne and coronation with proper insignia and king's robes—have remained the same in primitive cultures in ancient and modern times, as well as in many contemporary Western cultures. The throne, crown, headdress, garment (as sign of dignity), and sceptre (the staff through which the rule is carried out) were originally believed to contain the power through which the king ruled. The star garment of the Persian king symbolized his world rulership, similar to the feather mantle of the kings of Hawaii. In many cultures the throne, the crown, and the sceptre are viewed as divine and identified with gods and goddesses. This view was especially expressed in the Egyptian royal theology: in the hymnal prayer during coronations, the crowns of Upper and Lower Egypt were addressed as goddesses of the red and white crown by the king. In India the throne personified the kingdom. Sometimes, the throne that a new king ascends is viewed as the throne of the god. For example, Hatshepsut, an Egyptian queen (reigned 1503–1482 BC), was announced by Horus: "You have appeared on the throne of Horus." On many Egyptian images the king sits on the throne, and the god is at his side holding a hand over the king. In becoming someone else (a god), the king receives a new name, a throne name. Throne names are known in Africa, Mesopotamia, and Egypt (where the five throne names comprise the whole king theology: birth name, royal name, hawk name, serpent name, and a name that designates the king as heir of the power of the gods of the stars). In Iran, for example, the king is proclaimed by his royal name as world ruler. Immediately upon the proclamation of the new status of the king and his royal name, the subject people generally evoke a jubilant shout, such as "Long live the king." An African variety of a response to the proclamation is "He is our corn and our shield," which shows the importance of the king for his people. Another response to the proclamation is a prayer for the king: African and Polynesian prayers; Egyptian, Mesopotamian, and Israelite psalms; and hymns, such as the British hymn "God Save the King."

Symbols of sacred kingship

The act of adoration of a king is based on the throne rite, which is known only in areas having national kings. Though ascent to the throne and coronation with investiture are worldwide, there are many other rituals connected with sacred kingship. Among these are the anointment of the kings in Israel, India, and Iran, which originally was a ritual that gave strength to the recipient—as noted in primitive cultures (*e.g.,* rubbing with the fat of a lion); pseudo-fights (sham battles), from which the king emerges as victor; ritual cleansing; and ritual meals. The survival of elements of the sacred kingship in the Christian West is especially depicted in coronation rites. In early Christian art Christ is shown as kingly ruler on his throne with a royal court; he is emperor and universal ruler. Sacred kingship also survived in the papacy, as well as in the Holy Roman Empire (until the words Holy Roman were dropped in 1806). In the papacy, for example, the court ceremonies employ forms of address that go back to the imperial language of ancient sacred kingdoms: "Holy Father" or "Holiness."

Ascent, coronation, and other rituals

Ritual roles prescribed for kings in public or state functions. Many and various rituals express the concept that in the chief or king is concentrated the well-being of the country. At the order of a god, a Mesopotamian king might become involved in war; thus, his loot was placed before the god in the temple. If the king made a decision as judge, it was because of his unique wisdom as king. If he mediated a quarrel, the parties recognized his supreme power. The king acted to protect his land against the enemy; after his accession, in some areas, the king shot four arrows to the cardinal directions of the compass. In Africa and Egypt, for example, he then said: "I am shooting down the nations, to overcome them." He acted to ensure fertility and to distribute growth power when he started sowing corn, the seed of the tribe. He also was regarded as the guardian of the hearth fire (*e.g.,* Africa and Rome).

The king's participation in great festivals

The king exercised an important function through his participation in the great festivals, which were of utmost importance to the life of his people. In such festivals, various functions were differentiated: (1) priestly, as when the king presided over the sacrifices, said prayers, and gave the benediction, and (2) cultic participation, as when the king took part in the cult drama. The origin of the cult drama as a spontaneous event is known in primitive cultures (*e.g.,* the Ashanti in Africa), but it had its fullest expression in Mesopotamia and Egypt. In the Sed festival (in Egypt), the king, as ruler, renewed his rulership over the whole world. In the New Year's festival his ascent to the throne was renewed; and at the festival of Min, the god of life force and reproduction, the king played a significant role.

In Mesopotamia, festivals originating in cultic drama had great importance, especially the Babylonian New Year's festival. The events of the epic *Enuma elish,* which describes the sun god Marduk's victory over the powers of chaos and the resulting creation of the universe, were recreated in the cultic drama of the New Year's festival, in which the king represented Marduk, the victor and creator. Another cult drama represented the death and resurrection of the god of vegetation, in which the participants in mourning processions searched for the vanished god (represented by the king) and rejoiced at this triumphant return. Another Mesopotamian cult drama was the sacred marriage that the god Dumuzi celebrated with the goddess Innana. In the "holy wedding" the king and a priestess represented the god and the goddess and through this sexual union the forces of growth and fertility in nature were renewed. These cult dramas originated in early

prehistory, when gods were identified with the forces of nature and the cultic actions were understood as exerting direct influences on nature. At the Persian New Year's festival the king appeared as a killer of the dragon, whose rule was identified with the dry season.

The theory of the kingship in which the king occupied the position of mediator between people and gods also implies that the king may have to atone and suffer for the people of the cult. Under such a theory the absolution and reinstatement of the king meant the renewal of land and people. Perhaps behind this theory was a ritual similar to that found among African coronations of prehistory (as, for example, among the Ashanti) in which the king was beaten by priests before his installation.

Private ritual forms peculiar to kings and their families. The special status of the sacral king necessarily also influences his private life. In order to keep the supernatural force dwelling within him, the king had to observe a number of regulations and taboos in the details of his daily life. To this belongs temporary separation—in some cases, the king lived completely separated (*e.g.,* in Africa). The king often appeared only for audiences, on great festivals, or special occasions—sometimes veiled (as in Iran) or with a mask. There also have been special food taboos: he was not allowed to eat certain foods or may have had to drink only from a certain well. The custom of the king taking his meals alone is widespread. The isolation-separation theme in sacred kingship also appears in court ceremonials: the king must be addressed only from a certain measured distance; a person approaching the king must kneel; if the king is encountered, the head of the subject must be covered with the hands (as in Iran); the king must not be touched; and he must not touch the ground. Inasmuch as the king is filled with supernatural power, everything he touches can take on some of that power (as in Tahiti). Such proscriptions and taboos for the private life of the king are especially evident in Africa but also occur in Polynesia and Micronesia, East Asia, and the ancient Middle East. The divine or superhuman character pertained not only to the king but also, in lesser measure, to his family. The king's consecration could involve ritual incest. The participation of the family in the sacral status of the king was evidenced in several places (*e.g.,* Egypt), where, upon the death of the king or queen, members of the royal family and the court were killed or buried with them. Brother-sister marriages in some areas give evidence to this kind of royal ideology.

Personal regulations for a king's daily life

When a king began to grow old, it was said in Africa: "The grass is fading." To preserve the growth and well-being of the land, it was necessary to kill the aging king so that his power could be transferred to a successor. The compulsory killing of the king was widespread among many of the non-Semitic peoples in northern Africa; and among some peoples the killing of the king occurred after a specified period of time and was integrated into the cosmic ritualistic rhythm. The real meaning of the killing of the king showed itself in rituals in which the blood of the murdered king is mixed with the seed corn, which then became especially fertile.

CONCLUSION

With the increasing secularization of all areas of life, sacred kingship will of necessity disappear everywhere. Where monarchies are still retained, the sacred features will most likely diminish or vanish altogether. Only in the great regal ceremonies—especially coronations—are traces of sacred kingship still retained in the 20th century.

Certain elements of sacred kingship, however, are still encountered in modern life: in state ceremonies (*e.g.,* the red carpet at formal receptions), in governmental and ecclesiastical titles, and in the honour and respect given to particular personalities in different walks of life, such as in politics, art, and sports, in which the title king is often used to designate the secular veneration of an individual in the modern cult of personalities. (C.W.)

BIBLIOGRAPHY

Priesthood. The nature, characteristics, and significance of priesthood in primitive cultures ancient and modern are discussed in GRAHAME CLARK, *Archaeology and Society,* 3rd ed., rev. (1957); and ÉMILE DURKHEIM, *The Elementary Forms of the Religious Life,* 2nd ed. (1976; originally published in French, 1912). Cultural and scientific developments are examined in JACQUETTA HOPKINS HAWKES and LEONARD WOOLLEY, *Prehistory and the Beginnings of Civilization* (1963). The prominent features of priests and kings in the rise of civilization are considered by HAROLD PEAKE and HERBERT JOHN FLEURE, *The Corridors of Time,* vol. 4, *Priests & Kings* (1927). GUNNAR LANDTMAN, *The Origin of Priesthood* (1905); and E.O. JAMES, *The Nature and Function of Priesthood* (1955, reissued 1961), are comparative and anthropological studies of the subject in its wider aspects, with full bibliographies.

The organized priesthoods in ancient Egypt and the Middle East are treated in JAMES HENRY BREASTED (ed.), *Development of Religion and Thought in Ancient Egypt* (1912, reissued 1986); and HENRI FRANKFORT, *Ancient Egyptian Religion* (1948, reissued 1961). The available Sumerian evidence is produced by S.N. KRAMER, *Sumerian Mythology* (1944, reprinted 1988). S.H. HOOKE, *The Origins of Early Semitic Ritual* (1938, reissued 1980), surveys the priestly aspects of the autumnal festival. The Canaanite counterparts are recorded in CYRUS HERZL GORDON, *Ugaritic Literature* (1949); and GODFREY ROLLES DRIVER, *Canaanite Myths and Legends,* 2nd ed. (1978). A full bibliography on the Hellenic conceptions of priesthood is appended to the article by W.K.C. GUTHRIE, "The Religion and Mythology of the Greeks," in *The Cambridge Ancient History,* 3rd ed., vol. 2, part 2 (1961), pp. 851–905. The origin and status of the Levites and the Levitical code in the Hebrew priesthood are investigated by THEOPHILE JAMES MEEK, *Hebrew Origins,* rev. ed. (1950, reissued 1973). JULIA M. O'BRIEN, *Priest and Levite in Malachi* (1990), contains an up-to-date survey of research on the Israelite priesthood. The origin of the Christian priesthood is addressed in KENNETH E. KIRK (ed.), *The Apostolic Ministry* (1946, reprinted 1962).

The Vedic, Brahmanic, and Upanishadic conceptions of priesthood and the predominance of the Brahman caste in Hinduism are discussed in ARTHUR BERRIEDALE KEITH, *The Religion and Philosophy of the Veda and Upanishads,* 2 vol. (1925, reprinted 1989); S. RADHAKRISHNAN, *The Hindu View of Life* (1927, reissued 1980), and *Eastern Religions and Western Thought,* 2nd ed. (1940, reissued 1991); J.H. HUTTON, *Caste in India,* 2nd ed. (1951); and R.C. ZAEHNER, *Hinduism* (1962, reissued 1977), with a full bibliography. V. BOUILLIER and G. TOFFIN (eds.), *Priesthood, Power, and Authority in the Himalayas* (1989), in English and French, treats the role of the priesthood in several religions. The sublimation of priesthood in Buddhism in India, China, and Japan is treated in PAUL DAHLKE, *Buddhism and Its Place in the Mental Life of Mankind* (1927); and EDWARD CONZE, *Buddhism* (1951, reissued 1975). D. HOWARD SMITH, *Chinese Religions* (1968), introduces religious thought and sacerdotal practice in China. Texts concerning the Zen sect include DAISETZ TEITARO SUZUKI, *Manual of Zen Buddhism,* 2nd ed. (1950, reissued 1983); and ALAN WATTS, *The Way of Zen* (1957, reprinted 1989); while reference is made to it in R.C. ZAEHNER, *Mysticism, Sacred and Profane* (1957, reissued 1980). The priesthood in Shintō is discussed in D.C. HOLTOM, *The National Faith of Japan* (1938, reissued 1965).

Studies of the modern situation of the priesthood in Christianity include BERNARD COOKE, *Ministry to Word and Sacraments: History and Theology* (1976); MAX THURIAN, *Priesthood and Ministry: Ecumenical Research* (1983; originally published in French, 1970); and KENAN B. OSBORNE, *Priesthood: A History of the Ordained Ministry in the Roman Catholic Church* (1988). An overview of research on the Roman Catholic priesthood is DEAN R. HOGE, RAYMOND H. POTVIN, and KATHLEEN M. FERRY, *Research on Men's Vocations to the Priesthood and the Religious Life* (1984); a more recent essay is ANDREW M. GREELEY, "A Sea of Paradoxes: Two Surveys of Priests," *America,* 171(2):6–10 (July 16, 1994). Much contemporary Christian, and especially Roman Catholic, discussion regarding the priesthood focuses on the question of women's ordination. JACQUELINE FIELD-BIBB, *Women Towards Priesthood* (1991), argues in favour; while MANFRED HAUKE, *Women in the Priesthood?* (1988; originally published in German, 1982), argues against. (E.O.J./Ed.)

Shamanism. I.M. LEWIS, *Ecstatic Religion: A Study of Shamanism and Spirit Possession,* 2nd ed. (1989), provides an excellent introduction. A thorough description of the shamanism of the peoples of Siberia is given in M.A. CZAPLICKA, *Aboriginal Siberia* (1914, reissued 1969); and MIRCEA ELIADE, *Shamanism: Archaic Techniques of Ecstacy,* rev. and enlarged ed. (1964, reissued 1989; originally published in French, 1951), with an extensive bibliography. UNO HOLMBERG, *Finno-Ugric, Siberian,* vol. 4 in LOUIS HERBERT GRAY and GEORGE FOOT MOORE, *The Mythology of All Races* (1927, reissued 1964), describes shamanism among these peoples. A very thorough summary of the worldview and specific traits of shamanism in North Asia may be found in GEORG NIORADZE, *Der*

Schamanismus bei den sibirischen Völkern (1925); and the traits considered most significant are discussed by ÅKE OHLMARKS, *Studien zum Problem des Schamanismus* (1939). V. DIÓSZEGI (ed.), *Popular Beliefs and Folklore Tradition in Siberia* (1968; originally published in German, 1963), contains studies on the shamanistic conceptions of the Sami, Hungarian, and Siberian peoples. MARJORIE MANDELSTAM BALZER, *Shamanism: Soviet Studies of Traditional Religion in Siberia and Central Asia* (1990), summarizes contemporary Soviet research. Several studies explore Latin American shamanism, including JOHANNES WILBERT, *Tobacco and Shamanism in South America* (1987); and LUIS EDUARDO LUNA and PABLO AMARINGO, *Ayahuasca Visions: The Religious Iconography of a Peruvian Shaman* (1991). Women's roles in Korean shamanism are explored in LAUREL KENDALL, *Shamans, Housewives, and Other Restless Spirits* (1985), while *The Life and Hard Times of a Korean Shaman* (1988) chronicles the shamanic career of one South Korean woman through extensive use of her own words and stories. DAVID LAN, *Guns & Rain: Guerrillas & Spirit Mediums in Zimbabwe* (1985), examines the complex interaction between the Zimbabwe African National Liberation Army and the traditional shamanic religious leaders of the Dande and Korekore subgroups of Shona during the 1970s. (V.D./M.Ee./Ed.)

Monasticism. WALTER NIGG, *Warriors of God: The Great Religious Orders and Their Founders* (1959; originally published in German, 1953), provides an excellent account of the inceptors of monastic traditions, with special reference to the paramilitary trends in early monastic attitudes. Religions of the East are treated in a number of works. Hindu monasticism is studied in J.N. FARQUHAR, *The Fighting Ascetics of India* (1925); and AGEHANANDA BHARATI, *The Tantric Tradition* (1965, reprinted 1977), and *The Ochre Robe,* 2nd ed. (1980), partly autobiographical, which analyze the official and the esoteric monastic traditions in Hinduism. P.S. JAINI, "Śramanas: Their Conflict with Brahmanical Society," in JOSEPH ELDER (ed.), *Chapters in Indian Civilization,* rev. ed., 2 vol. (1970), is an excellent short survey of the Jaina monastic tradition juxtaposed with the Hindu and Buddhist orders. Surveys of Buddhist monasticism are L. AUSTINE WADDELL, *The Buddhism of Tibet, or, Lamaism* (1895, reprinted as *Buddhism & Lamaism of Tibet,* 1985), a classic, by way of an overall view of lamaism; DAISETZ TEITARO SUZUKI, *The Training of the Zen Buddhist Monk* (1934, reissued 1991), another classic, by a famous Zen scholar; ROBERT JAMES MILLER, *Monasteries and Culture Change in Inner Mongolia* (1959), an anthropological account of Buddhist monasticism in Mongolia (one of the only studies in English); E. ZÜRCHER, *The Buddhist Conquest of China,* 2 vol. (1959, reprinted 1972), an excellent, learned study of Chinese monastic and lay Buddhism and its conflicts with the official Chinese culture; SUKUMAR DUTT, *Buddhist Monks and Monasteries of India* (1962, reissued 1988), a classic work relating the monastic tradition to the other cultural traditions of India; J.A. NIELS MULDER, *Monks, Merit, and Motivation,* 2nd rev. ed. (1973), a report on monastic behaviour in contemporary South and Southeast Asian Buddhism; R.A.L.H. GUNAWARDANA, *Robe and Plough: Monasticism and Economic Interest in Early Medieval Sri Lanka* (1979), an analysis of the role of monastic institutions in economic development in agrarian societies; J. PRIPMØLLER, *Chinese Buddhist Monasteries,* 3rd ed. (1982); HEINZ BECHERT and RICHARD GOMBRICH (eds.), *The World of Buddhism: Buddhist Monks and Nuns in Society and Culture* (1984, reissued 1991); DAVID N. GELLNER, *Monk, Householder, and Tantric Priest: Newar Buddhism and Its Hierarchy of Ritual* (1992); and ROBERT E. BUSWELL, JR., *The Zen Monastic Experience: Buddhist Practice in Contemporary Korea* (1992), which gives readers an inside look at a Korean Zen monastery and challenges Western stereotypes regarding Zen Buddhism. Other works on Eastern monasticism include JOHN CAMPBELL OMAN, *The Mystics, Ascetics, and Saints of India* (1903, reprinted 1984), a classic work that is a fair account of the monastic situation in both ancient and contemporary India; and G.S. GHURYE, *Indian Sadhus,* 2nd ed. (1964), an English-language survey of the monastic situation in India.

Religions of the West are also treated in a number of texts. E.E. EVANS-PRITCHARD, *The Sanusi of Cyrenaica* (1949, reissued 1968), is an important anthropological study of a regional Ṣūfī tradition. Christian monastic history is examined in J. BRUCE WILLIAMSON (compiler), *The History of the Temple, London,* 2nd ed. (1925), an erudite historical account of the Templars; NORBERT MCMAHON, *The Story of the Hospitallers of St. John of God* (1958); OWEN CHADWICK (ed.), *Western Asceticism* (1958, reprinted 1979), providing a good survey, especially of the Roman Catholic orders; LOUIS BOUYER *et al., A History of Christian Spirituality,* 3 vol. (1963–69, reprinted 1982; originally published in French, 1960), a comprehensive survey of monastic developments; DAVID KNOWLES, *Christian Monasticism* (1969, reissued 1977); KARL SUSO FRANK, *With Greater Liberty: A Short History of Christian Monasticism and Religious Orders,* trans. by JOSEPH T. LIENHARD (1993; originally published in German, 1975), with discussions of both Orthodox and Reformed forms of monasticism; JOHN A. NICHOLS and LILLIAN THOMAS SHANK (eds.), *Medieval Religious Women,* 3 vol. (1984–92); BERNARD MCGINN, JOHN MEYENDORFF, and JEAN LECLERCQ (eds.), *Christian Spirituality: Origins to the Twelfth Century* (1985); JILL RAITT, BERNARD MCGINN, and JOHN MEYENDORFF (eds.), *Christian Spirituality: High Middle Ages and Reformation* (1987); LOUIS DUPRÉ, DON E. SALIERS, and JOHN MEYENDORFF (eds.), *Christian Spirituality: Post-Reformation and Modern* (1989); and MARIE AUGUSTA NEAL, *From Nuns to Sisters: An Expanding Vocation* (1990). (A.Bh./Ed.)

Sacred kingship. JAMES GEORGE FRAZER, *The Golden Bough,* 3rd ed., 12 vol. (1911–15, reprinted 1990), also available in an abridged ed., *The New Golden Bough* (1959, reissued 1972), is the fundamental and classical work. Also significant is A.M. HOCART, *Kingship* (1927, reprinted 1969), which considers the apotheosis of the king for the oldest religions of humankind. General works are G. VAN DER LEEUW, *Religion in Essence and Manifestation* (1938, reissued 1986; originally published in German, 1933); and GEO WIDENGREN, "Das sakrales Königtum," in his *Religionsphänomenologie* (1969), pp. 360–393. Also of interest is LILY ROSS TAYLOR, *The Divinity of the Roman Emperor* (1931, reprinted 1981). DAVID CANNADINE and SIMON PRICE (eds.), *Rituals of Royalty: Power and Ceremonial in Traditional Societies* (1987), is a collection of historical and anthropological essays that examines ritual as a form of power and includes studies of several different civilizations.

S.H. HOOKE (ed.), *Myth, Ritual, and Kingship: Essays on the Theory and Practice of Kingship in the Ancient Near East and in Israel* (1958), is a study from the English school. Studies from the Scandinavian school are found in GEO WIDENGREN, *Sakrales Königtum im Alten Testament und im Judentum* (1955); S. MOWINKEL, *He That Cometh* (1956), pp. 21–95; IVAN ENGNELL, *Studies in Divine Kingship in the Ancient Near East,* 2nd ed. (1967); and AAGE BENTZEN, *King and Messiah,* 2nd ed. (1970; originally published in German, 1948).

Sacred kingship in Egypt is explored in ALEXANDRE MORET, *Du Caractère religieux de la royauté pharaonique* (1902). Mesopotamia is the focus of RENÉ LABAT, *Le Caractère religieux de la royauté assyro-babylonienne* (1939); and THORKILD JACOBSEN, *Toward the Image of Tammuz and Other Essays on Mesopotamian History and Culture* (1970). HENRI FRANKFORT, *Kingship and the Gods* (1948, reissued 1978), an important exposition on Egypt and Mesopotamia, ushers in the new epoch of research. Discussion of sacred kingship in Canaan can be found in JOHN GRAY, "Canaanite Kingship in Theory and Practice," *Vetus Testamentum,* 2:193–220 (1952). Works on Africa include LEO FROBENIUS, *Erythräa: Länder und Zeiten des heiligen Königsmordes* (1931); and DIEDRICH WESTERMANN, *Geschichte Afrikas* (1952, reissued 1968), pp. 20–46. Texts examining sacred kingship in Israel are JEAN DE FRAINE, *L'Aspect religieux de la royauté israélite* (1954); *The Sacral Kingship* (1959), papers from the eighth International Congress for the History of Religions; KARL-HEINZ BERNHARDT, *Das Problem der altorientalischen Königsideologie im Alten Testament* (1961); MARTIN NOTH, "God, King, and Nation in the Old Testament," in his *The Laws in the Pentateuch and Other Studies* (1966, reissued 1984; originally published in German, 2nd enlarged ed., 1960), pp. 145–178; AUBREY R. JOHNSON, *Sacral Kingship in Ancient Israel,* 2nd ed. (1967); H.H. ROWLEY, *Worship in Ancient Israel: Its Forms and Meaning* (1967, reissued 1976), pp. 186–202; J. ALBERTO SOGGIN, *Das Königtum in Israel* (1967); BARUCH HALPERN, *The Constitution of the Monarchy in Israel* (1981); and RICHARD HARLIN LOWERY, *The Reforming Kings: Cults and Society in First Temple Judah* (1991).

The Roman Catholic papacy represents the single most important attempt at the construction of a sacral monarchy in European civilization. COLIN MORRIS, *The Papal Monarchy: The Western Church from 1050 to 1250* (1989), includes a good general introduction to the subject. PAOLO PRODI, *The Papal Prince: One Body and Two Souls: The Papal Monarchy in Early Modern Europe* (1987), examines the interrelation of spiritual and temporal power in the ideology of papal absolutism. Papal claims to sacral kingship faced competition from secular rulers who saw themselves as the heirs of Christ's royal office. J.H. BURNS, *Lordship, Kingship, and Empire: The Idea of Monarchy, 1400–1525* (1992), examines the role of Christian theology in the development of monarchical ideology in Europe.

(C.W./Ed.)

Saint Petersburg

The second largest city of Russia and one of the world's major cities, St. Petersburg has played a vital role in Russian history—for two centuries as capital of the Russian Empire—and is important today as an industrial and cultural centre and as a seaport. Founded by Peter I the Great in 1703 as St. Petersburg (Russian: Sankt Peterburg), it was renamed Petrograd in 1914 and Leningrad in 1924. The name St. Petersburg was restored in 1991.

The city is particularly renowned as the scene of the February and October revolutions in 1917 and as the besieged and fiercely defended city of World War II. Architecturally, it ranks as one of the most splendid and harmonious cities of Europe. St. Petersburg lies in the far northwestern corner of Russia, about 398 miles (640 kilometres) northwest of Moscow and only about 7° south of the Arctic Circle.

This article is divided into the following sections:

Physical and human geography 1037
- The landscape 1037
  - The city site
  - Climate
  - The city layout
  - The outer region
- The people 1041
- The economy 1041
  - Industry
  - Transportation
- Administration and social conditions 1042
  - Government
  - Services and health
  - Education
- Cultural life 1042

History 1042
- The early period 1042
  - Foundation and early growth
  - The rise to splendour
- Evolution of the modern city 1043
  - The road to revolution
  - The Soviet period

Bibliography 1044

Physical and human geography

THE LANDSCAPE

The city site. St. Petersburg is located on the delta of the Neva River, at the head of the Gulf of Finland. The city spreads across nearly 42 islands of the delta and across adjacent parts of the mainland floodplain. The very low and originally marshy site has made the city subject to recurrent flooding, especially in the fall, when strong cyclonic winds drive gulf waters upstream, and also at the time of the spring thaw. Exceptionally severe inundations occurred in 1777, 1824, and 1924, the last two being the highest on record and flooding most of the city. To control the destructive floodwaters, the city began construction in the 1980s of an 18-mile-long dike across the Gulf of Finland. A number of canals also have been cut to assist drainage. These, together with the many natural channels, make St. Petersburg a city of waterways and bridges and have earned it the nickname "Venice of the North."

Floods

Greater St. Petersburg, the city itself with its satellites, forms a horseshoe shape around the head of the Gulf of Finland and includes the island of Kotlin in the gulf. On the north it stretches westward along the shores for nearly 50 miles to include Zelenogorsk. This northern extension is an area of dormitory towns, resorts, sanatoriums, and children's camps set among extensive coniferous forests and fringed by fine beaches and sand dunes. Some influential citizens also have summer cottages, or dachas, in this area. On the southern side of the gulf, the metropolitan limits extend westward to include Petrodvorets and Lomonosov. Eastward, Greater St. Petersburg stretches up the Neva River to Ivanovskoye.

Climate. The mitigating effect of the Atlantic Ocean provides St. Petersburg with a milder climate than might be expected for its far northern site. Nevertheless, winters are rather cold, with a mean January temperature of about 18° F (−8° C), a few degrees warmer than Moscow. Winter temperatures can drop below −40° F

© Guy Bouchet, Paris



Bridge at the confluence of the Moyka and Fontanka rivers.

(−40° C), however. Snow cover lasts on the average about 132 days. The Neva begins to freeze normally about mid-November, and the ice is solid by the start of December; breakup begins in mid-April and usually is completed by the end of the month. Icebreakers prolong the navigation season. Summers are moderately warm, with an average temperature of 64° F (18° C) in July. The mean annual precipitation amounts to about 23 inches (584 millimetres), the maximum coming in summer. Because of St. Petersburg's northerly location, winter nights are long; in early summer, on the other hand, the city enjoys the half-light "white nights," one of its most renowned features.

The city layout. Central St. Petersburg is divided by distributaries of the Neva River into four sections. The Admiralty Side lies along the left (south) bank of the Neva itself. Between the two major arms of the Neva, the Bolshaya (Great) Neva and the Malaya (Little) Neva, is Vasilyevsky Island. The Malaya Neva and the distributary known as the Bolshaya Nevka enclose a group of islands known as the Petrograd Side, while east of the Bolshaya Nevka and north of the Neva lies the Vyborg Side.

The Admiralty Side. Much of St. Petersburg's historical and cultural heritage is concentrated on the Admiralty Side. The district centres on the Admiralty. This, the nucleus of Peter's original city, was reconstructed in 1806–23 by Andreyan D. Zakharov as a development of the earlier building of Ivan K. Korobov, remodeled in 1727–38, but retaining the layout of the original. Its elegant spire, topped by a weather vane in the form of a ship, is one of the principal landmarks of the city. The building today houses a naval college.

Admiralty

Just to the east lies the great Palace Square, the city's oldest. The 600-ton granite monolith of the Alexander Column (1830–34), the tallest of its kind in the world and so finely set that its base is not fastened, thrusts up for 165 feet (50 metres) near the centre of the square.

Between the square and the river rises the huge and massive rectangle of the Winter Palace, the former principal residence of the tsars. The present structure, the fifth to be built, was the Baroque masterpiece of Bartolomeo F. Rastrelli. Construction of it began in 1754 and was completed in 1762. Both the exterior and the interior of the palace were designed in dazzlingly luxurious style. In 1837 the building was destroyed by fire, and only the adjoining Hermitage survived; the Winter Palace was recreated in 1839 almost exactly according to Rastrelli's plans. The striking appearance of the palace is highlighted by white columns against a green background, with golden stucco moldings; 176 sculptured figures line the roof. The whole complex, now called the Hermitage, or State Hermitage Museum, is a treasure-house of fine art of worldwide significance that originated in 1764 as the private collection of the tsarina Catherine II.

Opposite the Winter Palace, the great crescent of Carlo Rossi's General Staff building (1811–29) dominates the square. The two wings of the building are joined by a huge triumphal arch, topped by heroic figures and crowned by a chariot carrying a figure representing Glory, expressing the Russian victory in the campaign of 1812.

On the western (downstream) side of the Admiralty stretches the expanse that was called Senate Square when the Senate moved there in 1763; it is now called Decembrists' (or Dekabrists') Square in commemoration of the revolt in 1825. The buildings of the former Senate and Synod (now housing archives) dominate the western side of the square, their decorated facades dating from the 1830s and representing the last great work of Rossi. They are separated by an arch looking across to the centre of the square where stands the equestrian statue of Peter the Great, known as the "Bronze Horseman" and created in 1782 by Étienne Falconet. Near the Senate and Synod buildings to the south rises the classical front of the Horse Guards Riding School, or Manezh (1804–07); beyond, dominating the south side of St. Isaac's Square, is the cathedral of the same name. An outstanding monument of late classical Russian architecture built by Auguste Montferrand (1818–58), St. Isaac's is one of the largest domed buildings in the world; its golden cupola, gilded with about 220 pounds (100 kilograms) of pure gold, soars to 331 feet in height and is visible all over St. Petersburg. It is now a museum.

Decembrists' Square

From the Admiralty and its surrounding squares radiate three great avenues, of which the most important and best known is the Nevsky Prospekt. One of the world's great thoroughfares, the Nevsky Prospekt cuts southeastward across the peninsula formed by the northward loop of the Neva to the vicinity of the Alexander Nevsky Abbey, crossing the smaller Moyka and Fontanka rivers. The Anichkov Bridge across the latter is graced by four sculptured horses. The street has a special beauty: the architecture is majestic, the buildings are graceful and finely proportioned, the construction is complex. On the Nevsky Prospekt stand the Stroganov, Shuvalov, and Anichkov palaces (former private residences of the nobility) and several churches, of which the most prominent are St. Peter's Lutheran Church (1833–38), St. Catherine's Roman Catholic Church (1763–83), and the Kazan Cathedral (1801–11). The latter edifice, undoubtedly the street's finest feature, was designed by Andrey Voronikhin in Russian classical style and has an interior rich in sculptures and paintings behind a magnificent semicircular frontal colonnade. Another interesting building is the department store Gostiny Dvor (1761–85), originally designed by Jean-Baptiste M. Vallin de la Mothe. This building forms an irregular square and opens onto four streets; formerly it was a mercantile centre. Other department stores line the Nevsky Prospekt, as

Nevsky Prospekt

© Wolfgang Kaehler



Nevsky Prospekt looking northwest toward the spire atop the Admiralty.



St. Petersburg and its metropolitan area and (inset) the central city.

do many theatres—most notably the Pushkin Academic Drama Theatre—restaurants, and cafés.

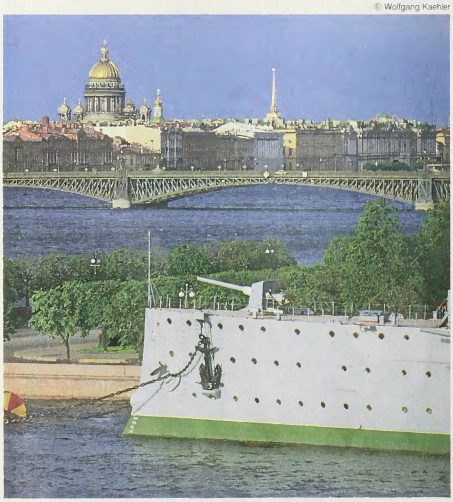
At the eastern end of the Nevsky Prospekt, Alexander Nevsky Square fronts the main entrance to the abbey of the same name and its surrounding gardens. Beyond the square's main entrance lie, on the left and right, respectively, monuments and sculptures of the 18th-century Lazarus Cemetery (where Mikhail V. Lomonosov and many of the city's architects are buried) and the 19th-century Tikhvin Cemetery (containing the graves of such writers and composers as Dostoyevsky, Mussorgsky, and Tchaikovsky). Behind rise the spires and cupolas of the Church of the Annunciation (1720, designed by Domenico Trezzini), which is now a museum, and Holy Trinity Cathedral (1778–90, designed by Ivan Starov).

Through the Admiralty Side and intersecting the radial avenues cut the natural channels and canals that so characterize the city. The most important, in outward order from the Admiralty, are the Moyka and Fontanka rivers and the Griboyedov and Obvodny (Bypass) canals. Downstream from the northern entrance of the Fontanka into the Neva lies the Field of Mars, one of the city's beautiful open spaces. Begun under Peter the Great (when it was known as the Field of Amusement), it was intended for popular festivities and fireworks. It was a favourite haunt of the 18th-century nobility, but its present name derives from a monument erected in 1801 and portraying the great Russian military leader Aleksandr V. Suvorov (buried in the Church of the Annunciation) as the god of war. In the 19th century the space was used for military parades and exercises. The fallen of the February Revolution of 1917 and the defenders of the city during the civil war and foreign military intervention (1918–20) were buried there. They are commemorated by an eternal flame.

Just to the east lies the Summer Garden. Founded on an island in 1704, it has parks and gardens that, by the end of the 18th century, contained more than 250 statues and busts, mostly the work of Venetian masters. The Summer Palace, Peter's first in the city, erected 1710–14 in early Russian Baroque style and designed by Trezzini, stands in the northeastern portion. The Neva embankment is fronted by a fence (1784), the iron grille of which is reputed to be among the world's finest examples of wrought iron work. So light and delicate is its design that the grillwork almost seems to be suspended in air.

Summer Palace

© Wolfgang Kaehler



The bow of the cruiser *Aurora*, anchored in the Bolshaya Nevka River, and (centre) the Troitsky (Trinity) Bridge crossing the Neva River. Beyond (left background) is the dome of St. Isaac's cathedral.

At the extreme eastern side of the central city, within the sharp bend of the Neva itself, lies the Smolny complex of buildings; these include the former convent, with the five-domed cathedral, designed by Rastrelli and begun in 1748, and the classical building of the Smolny Institute, constructed by Giacomo Quarenghi in 1806–08. The institute was used as Lenin's headquarters in 1917.

Vasilyevsky Island. One of the first areas of St. Petersburg to be developed because of its defendable position, Vasilyevsky Island forms the northwestern corner of the central city. Opposite the Admiralty and Winter Palace, at the island's eastern tip, is the remarkable architectural complex known as the Strelka (literally, "Pointer"), facing the bifurcation of the Neva. Behind the two great Rostral Columns, decorated by carved ships' prows, and across Pushkin Square, the point rises majestically to the former Exchange building (Thomas de Thomon, 1805–10), the city's finest example of early 19th-century style and reminiscent of a classical Greek temple in appearance; it now houses the Central Naval Museum.

Farther back, the Twelve Colleges building (Trezzini, 1722–42), originally intended to house the supreme governmental bodies of Peter the Great, is now the home of the city's state university. The building is divided into 12 identical but independent sections and runs at right angles to the Neva embankment, which is fronted at that point by the facades of the main building of the Academy of Sciences, the Menshikov Palace, and the Academy of Arts. On the far, or northern, side of the Exchange is the Customs House (now the literary museum and the Institute of Russian Literature known as Pushkin House), designed by Giovanni Luchini (1829–32).

The Petrograd Side. Upstream of the bifurcation of the Neva is the Petrograd Side, where the great Peter-Paul Fortress faces the Strelka across the Malaya Neva. Founded in 1703, this fortification, the city's first structure, initially had earthen walls, but these were soon replaced by stone walls 40 feet high and 12 feet thick, with 300 cannons mounted on the bastions. Above the squat horizontal lines of the fortress's massive walls soars the slender, arrow-like spire of the Cathedral of St. Peter and St. Paul, a golden landmark for the city. The cathedral was built in 1712–33 by Trezzini, and the tsars and tsarinas of Russia from the time of Peter the Great (except for Peter II and Nicholas II) are buried in it. Trezzini also designed St. Peter's (Petrovsky) Gate (1718) as the eastern entrance to the fortress. The Neva Gate, designed by Nikolay A. Lvov, dates from 1787. From the early 19th century the fortress was used as a prison, chiefly for political prisoners. Today it is a museum. At noon each day a cannon is fired from its battlements.

Cathedral of St. Peter and St. Paul

Just to the east of the Peter-Paul Fortress, where the Bolshaya Nevka River begins, the cruiser *Aurora* is permanently moored as a museum and training vessel for the Naval College. It was the *Aurora* that in 1917 fired the blank shot that served as the signal to storm the Winter Palace.

The Vyborg Side. The northeastern part of the central city had by the late 19th century developed into an industrial suburb. One of its most famous features is the Finland Railway Station, which faces the Admiralty Side across the Neva. Lenin returned to Russia in April 1917 via this station, and there he made his initial pronouncement of a new course that would bring the Bolsheviks to power. A major street of the Vyborg Side is the Prospekt Karla Marksa, along which stand such buildings as the Cathedral of St. Sampson.

The outer region. St. Petersburg now extends well to the north and south of the original delta site, with arms of growth extending westward along the banks of the Gulf of Finland. The newer outer suburbs include extensive open areas, which help to reduce the overall population density, and parts of the periphery are designated as greenbelt. However, the multiplicity of large housing blocks containing numerous two- or three-room apartments means that population densities in the built-up areas remain extremely high. As in virtually all modern cities, commuting

over long distances is the price paid for more living space and the cleaner air of the suburbs. Among the suburbs noteworthy for their historic and cultural value are Petrodvorets, Pushkin, Pavlovsk, and Gatchina.

Petrodvorets. The most famous of the communities around St. Petersburg is Petrodvorets (Peterhof before 1944), whose unique garden-park setting, stretching in terraces rising above the Gulf of Finland, contains representative works from two centuries of Russian architectural and park styles. The Great Palace, the former residence of Peter the Great, stands at the edge of the second terrace, its bright yellow walls contrasting with white stucco decorations and the gilt domes of its lateral wings. Built in the Baroque style (1714–28), it was reconstructed and expanded by Rastrelli from the mid-1740s to the mid-1750s. On the north the building commands a view of the Grand Cascade, a grandiose structure including a grotto, 64 fountains, and two cascading staircases, which lead to an enormous semicircular basin that contains a giant statue of Samson wrestling with a lion. This statue, symbolizing the military glory of Russia, is a copy of the original statue by Mikhail I. Kozlovsky, which was carried off by the Nazis during World War II. In fact, much of the town's treasure was plundered, and this magnificent vista becomes all the more remarkable when it is remembered that much of it is a post-World War II restoration.

Great Palace

Pushkin. The town of Pushkin (called Tsarskoye Selo before 1917, Detskoye Selo in 1918–37) arose in the early 18th century as one of the tsarist residences. The Catherine Palace (1717–23; enlarged by Aleksey V. Kvasov and Savva I. Chevakinsky, 1743–48; rebuilt by Rastrelli, 1752–57) is notable for its dimensions, the beauty and majesty of its form, and the wealth of its sculptural decoration. The golden suite of splendid halls (including the Amber Room) exemplifies Russian Baroque at its peak. The community also is the site of the Chinese Village (1782–96) in Alexander Park and the gallery (1780–90) named after its architect, Charles Cameron, the terraces of which contain more than 50 busts of figures from ancient Greek and Roman history. The Lycée, a school for the offspring of the nobility, had the great Aleksandr Pushkin as a student, and a famous statue of the poet stands near the town's Egyptian Gates. The town suffered severe damage during the German onslaught but has been restored.

Pavlovsk. Pavlovsk, a southern suburb, is the site of a late 18th- and early 19th-century palace and park in the classical style that was created as a country residence for Tsar Paul I. The central Great Palace (1782–86; Cameron) is crowned by a dome supported by 64 columns. It was severely damaged by the Nazis and has been restored.

Gatchina. Another southern suburb, Gatchina, is noted for the palace built in 1766–81 by Antonio Rinaldi for Count Grigory Orlov, a favourite of Catherine II. Gatchina Park was created at the same time. Its monuments, sculptures, and gardens, like those of all St. Petersburg, are preserved by the state.

THE PEOPLE

The population of St. Petersburg is overwhelmingly Russian. Before the Revolution the city had sizable Polish, Baltic, and German and smaller Tatar, Jewish, and Chinese communities. In the interwar period it continued to act as a magnet for Russian peasant labour, and, even in the more homogeneous postwar city, newcomers have tended to outnumber those native to St. Petersburg. Nevertheless, observers have commonly ascribed certain traits to the people—politeness, a sophistication, a slight reserve—that have seemingly passed from generation to generation. The old intelligentsia is no more, but many people in St. Petersburg, living in a city designed as a cultural centre, consider themselves to be the most cultivated of Russians.

THE ECONOMY

Industry. St. Petersburg is one of the major industrial centres in Russia, with more than half of its working population employed in factories and the building trades; its products are distributed throughout Russia and beyond. The city also makes important contributions in scientific and technical research.



The Grand Cascade at Petrodvorets with (centre) the statue of Samson (a copy of the original by Mikhail I. Kozlovsky).
David B.A. Jones/Robert Harding Picture Library

Engineering accounts for more than half of the city's industrial output. Special emphasis is placed on those branches of engineering that require a skilled labour force and relatively small quantities of metal. St. Petersburg's original industry, shipbuilding, is still important and remains one of the largest of its kind in Russia; it produces icebreakers (some atomic-powered), tankers, timber carriers, and fishing vessels. Other sectors of heavy engineering make armaments and rolling stock. Of national importance are a plant producing nuclear reactors and others that manufacture electrical and power machinery, such as steam, hydraulic, and gas turbines, armatures, and generators.

Precision engineering includes tools and instruments, refrigerators, radios, televisions, and other electrical and electronic goods. Machinery is produced for automated factories and for the knitwear and footwear industries.

Chemical-based industries rank second to engineering in importance. These make a wide range of items, such as superphosphate fertilizers (using apatite from the Kola Peninsula), tires and other synthetic rubber goods, plastics and plastic goods, artificial and synthetic fibres, paints, and pharmaceutical preparations. There are industries producing a wide range of consumer goods, with the city itself as their principal market. These include the production of cotton and woolen textiles, clothing, footwear, tobacco products, beer, and foodstuffs. St. Petersburg has a long-established and sizable printing industry.

Electrical power for these industries comes from hydroelectric plants on the Volkhov, Svir, and Vuoksa rivers and, more recently, from nuclear power stations. Natural gas is piped to St. Petersburg from the southern regions of European Russia and Central Asia.

Power sources

Transportation. St. Petersburg is one of Russia's most important hubs of transportation. Its port, the nation's largest, is of international significance. The main harbour is protected by breakwaters and is reached by a dredged channel from Kronshtadt. Imports include metal pipes, factory equipment, chemicals, sugar, cotton, and fruit, while machinery, timber, coal, potassium salts, and pyrites form the bulk of exports. Passenger liners maintain reg-

ular summer services to Stockholm and to Tilbury, Eng. Smaller seagoing ships have access by way of the Neva to Lake Ladoga and thence throughout the inland waterway system of European Russia. From Lake Ladoga the Svir River, Lake Onega, and the White Sea Canal (Belomor Kanal) lead to the White Sea and the Russian Arctic coast. From Lake Onega the Volga-Baltic Waterway system leads to the Volga basin and Caspian Sea and, via the Volga-Don Canal, to the Black and Azov seas. The main airfield, Pulkovo International Airport, is located 11 miles south of the city.

The city is a focus of rail routes, with trunk lines radiating to Helsinki, Fin., and Warsaw as well as to Moscow and other major Russian cities. There are five principal passenger rail terminals—the Moscow, Vitebsk, Warsaw, Baltic, and Finland stations. A network of suburban electric services connects the outer parts of St. Petersburg and its satellite towns. Internal city traffic is carried by a subway system (opened in 1955) and a well-developed network of bus, streetcar, and trolleybus lines.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. Since 1931 St. Petersburg has been designated a "city of republican subordination"—that is, its city council is directly subordinate to the government of the Russian republic. The council members are elected for two-year terms. The city itself is divided into 16 administrative wards. In addition to these, there are five towns in the outer part of Greater St. Petersburg administered by the local government, which have status equivalent to that of the administrative wards; these include Kolpino, Petrodvorets, Pushkin, Sestroretsk, and Kronshtadt, on the island of Kotlin. Two other towns, Zelenogorsk and Pavlovsk, and 17 urban districts are subordinate to the administrative ward councils. St. Petersburg is also the administrative centre of St. Petersburg *oblast* (province). As a residue of the city's former status as capital, certain organizations still maintain their national headquarters in St. Petersburg, among them the all-union geographic, chemical, and medical societies.

Services and health. Much of the city's housing was destroyed during World War II and was replaced by massive postwar building programs. As a result, a large proportion of the people live in relatively new and modern apartments, some in high-rise buildings. Virtually all housing has central heat and is tied to the city's sanitation and power services.

St. Petersburg is fully equipped with the health services of a modern city. There are several hundred clinics providing medical and dental care and maternity and nursing services. Medical care also is provided by more than 150 general and specialized hospitals. As in the rest of Russia, health care in St. Petersburg is free.

Education. St. Petersburg is one of Russia's most important centres for education and scientific research; a sizable proportion of the employed population is engaged in education, the arts, and the sciences. Heading the educational establishments is the city's state university, founded in 1819. No less renowned and even older are the Academy of Arts, founded in 1757, the Institute of Mines (1773), and the Military Medical Academy (1798). St. Petersburg also has a polytechnic institute (1899), as well as numerous other higher educational establishments, general schools, and specialist and technical secondary schools.

Library of the Academy of Sciences

A focus for research is the library of the Russian Academy of Sciences (from 1925 to 1991, Academy of Sciences of the U.S.S.R.), which remained in the city when the academy's headquarters moved back to Moscow after the Revolution. The research establishments of the Academy of Sciences in the city include the Botanical, Geological, Forestry, and Zoological institutes as well as the Pulkovo Observatory. The city is the principal centre in Russia for Arctic research, notably at the Arctic and Antarctic Research Institute and the Institute for the Study of Permafrost.

CULTURAL LIFE

St. Petersburg evolved as a city of culture, and the number and quality of its cultural institutions remain one of its enduring attractions. It has many large and grand theatres and auditoriums. The Kirov State Academic Theatre of Opera and Ballet, formerly the Mariinsky Theatre, has long enjoyed an international reputation, and its resident company is frequently on tour abroad. Other important theatres are the Maly (Little), Gorky, Pushkin, and Musical Comedy theatres. The largest of several concert halls is the October Great Concert Hall, which seats 4,000 people. The city's musical tradition has been enhanced by the Rimsky-Korsakov Conservatory. Notable museums include the Hermitage and the State Russian Museum, which specializes in Russian painting. St. Petersburg is a significant centre of the country's motion-picture industry.

Museums

There are a large number of libraries in the city, headed by the Saltykov-Shchedrin Public Library on the Nevsky Prospekt, established in 1795; among libraries in Russia, it is second only to Moscow's Lenin Library. Another important specialized collection is in the Pushkin House literary museum on Vasilyevsky Island.

St. Petersburg has abundant recreational facilities. Among the notable stadiums in the area is Kirov Stadium, which seats approximately 100,000 spectators. Other opportunities for outdoor recreation are provided by the Kirov Park of Culture and Rest, the zoo, the botanical gardens, and numerous other smaller parks and gardens.

History

THE EARLY PERIOD

Foundation and early growth. Settlement of the region around the head of the Gulf of Finland by Russians began in the 8th or 9th century AD. Known then as Izhorskaya Zemlya, or more commonly as Ingermanland, or Ingria, the region came under the control of Novgorod, but it long remained thinly populated. In the 15th century the area passed, with Novgorod, into the possession of the grand princes of Moscow. Sweden annexed Ingria in 1617 and established fortresses along the Neva. During the Second Northern War (1700–21) Peter the Great, seeking a sea outlet to the west, constructed a fleet on the Svir River (which connects Lakes Onega and Ladoga) and, sailing across Lake Ladoga, launched an attack on the fortress of Noteburg (now Petrokrepost), where the Neva flows out of Ladoga. In 1703 Noteburg fell to Peter; afterward he captured the Swedish fortress of Nienshants on the lower Neva, thus gaining control of the delta.

On May 16 (May 27, New Style), 1703, shortly after the fall of Nienshants, Peter himself laid the foundation stones for the Peter-Paul Fortress on Zayachy Island. This date is taken as the founding date of St. Petersburg. In the spring of the following year, Peter established the fortress of Kronshlot, later Kronshtadt, on Kotlin Island in the Gulf of Finland, to protect the approaches to the delta. At the same time, he founded the Admiralty shipyard on the riverbank opposite the Peter-Paul Fortress; in 1706 its first warship was launched. Around the fortress and shipyard Peter began the building of a new city to serve as his "window on Europe." Just upstream of the Peter-Paul Fortress, the first small house, built for Peter himself in the early days of the city's construction, is preserved as a museum.

Peter-Paul Fortress

Although the first dwellings were single-storied and made of wood, it was not long before stone buildings were erected. The first stone palace, still preserved, was completed in 1714 for Prince Aleksandr Danilovich Menshikov, first governor of the city. From the start the city was planned as an imposing capital, on a regular street pattern, with spacious squares and broad avenues radiating out from the Admiralty. Architects, craftsmen, and artisans were brought from all over Russia and from many foreign countries to construct and embellish the new town. In 1712 the capital of Russia was transferred there from Moscow, although it was not until 1721 that Sweden, in the Peace of Nystad, formally ceded sovereignty of the area to Russia. Members of the nobility and merchant class were compelled by Peter to move to the new capital and to build houses for themselves. Government buildings and private palaces and houses arose swiftly; among the earliest buildings were the Merchants' Exchange (now the

Naval Museum), Customs House (now the Museum of Literature), and marine hospital, together with the Summer Palace. Canals for drainage were cut through the marshy left bank of the Admiralty Side. The first floating bridge over the Neva was constructed in 1727, and soon more than 370 bridges had been built across the many canals and river channels. Marshy, flood-prone land and an inhospitable climate made construction expensive in terms of human life; St. Petersburg, it was later suggested, rested on a swamp of human bones.

A harbour was constructed, and Peter took measures to curtail traffic through Arkhangelsk on the White Sea, previously Russia's major port. In consequence, as early as 1726 St. Petersburg was handling 90 percent of Russia's foreign trade. In 1703 work began on the Vyshnevolotsky Canal in the Valdai Hills, the first link in a chain that by 1709 gave the capital a direct water route to central Russia and all of the Volga basin. Industry soon began to develop. The original and flourishing Admiralty shipyard was joined by enterprises to supply its needs and those of the growing fleet—a foundry to produce cannons, a gunpowder factory, and a tar works. Merchantmen as well as warships were built, and before the end of the 18th century papermaking, printing, and food, clothing, and footwear industries had been established; as early as the 1740s a factory was set up to make china. By 1765 the population numbered 150,300, and by the end of the century it had reached 220,200, of whom more than a third were in the armed forces or the administration.

The rise to splendour. The growing city displayed a remarkable richness of architecture and harmony of style. Initially the style was one of simple but elegant restraint, represented in the cathedral of the Peter-Paul Fortress and in the Summer Palace. In the mid-18th century an indelible stamp was put on the city's appearance by the architects Bartolomeo F. Rastrelli, Savva I. Chevakinsky, and Vasily P. Stasov, working in the Russian Baroque style, which combined clear-cut, even austere lines with richness of decoration and use of colour. To this period belong the Winter Palace, the Smolny Convent, and the Vorontsov and Stroganov palaces, among others; outside the city were built the summer palaces of Peterhof and of Tsarskoye Selo. After a transitional period, dominated by the architecture of Jean-Baptiste M. Vallin de la Mothe and Aleksandr Kokorinov, toward the end of the 18th century a pure classical style emerged under the architects Giacomo Quarenghi, Carlo Rossi, Andrey Voronikhin, and others. The Kazan and St. Isaac's cathedrals, the Smolny Institute, the new Admiralty, the Senate, and the Mikhaylovsky Palace (now the State Russian Museum) are representative of the splendid buildings of this period.

Russian Baroque architecture

Within this grand architectural setting, cultural life developed and flourished. In 1773 the Institute of Mines was established. The University of St. Petersburg was founded in 1819. Many of the most celebrated names in Russia in the spheres of learning, science, and the arts are associated with the city: Mikhail V. Lomonosov, Dmitry I. Mendeleyev, Ivan Pavlov, Aleksandr Pushkin, Leo Tolstoy, and Fyodor Dostoyevsky, among others. Dostoyevsky's *Crime and Punishment* was set in the city, and the buildings described in the novel are a focus of tourism. As early as 1738 the first ballet school in Russia was opened in St. Petersburg; in the 19th century, under Marius Petipa, the Russian ballet rose to worldwide renown and produced such dancers as Vaslav Nijinsky, Tamara Karsavina, and Anna Pavlova. In 1862 the first conservatory of music in Russia opened its doors, and there the premieres of works by Tchaikovsky, Rimsky-Korsakov, Rachmaninoff, and other composers were performed. Over all, as focus and patron of the city's cultural life, stood the imperial court; its ostentatious splendour and wealth were legendary throughout Europe.



The Winter Palace, with the Alexander Column in the foreground, Palace Square.
Philip Craven/Robert Harding Picture Library

EVOLUTION OF THE MODERN CITY

The road to revolution. The imperial magnificence, centred on the tsarist autocracy, was in sharp contrast to the other side of St. Petersburg's development, the growth of its industrial proletariat. During the 19th century there was much industrial growth in the city, accelerated by improvements in communications and extension of trade. Navigation was opened on the Mariinsky (1810) and Tikhvin (1811) canal systems, which replaced an old and inadequate system. In 1813 the first Russian steamship was built in St. Petersburg, and in 1837 the first Russian railway was constructed from the city to the Summer Palace in Tsarskoye Selo (now Pushkin). Five years later work started on the railway to Moscow, opened to traffic in 1851. A line to Warsaw was built in 1861–62, followed by still others. In particular, the cotton textile and metal-working industries flourished, the former using imported raw materials. By the 1840s more than 60 percent of Russian imports were entering by way of St. Petersburg. In 1885 a channel was dredged to give larger ships access to the port. City growth and industrialization were stimulated by the emancipation of the Russian serfs in 1861, which allowed far greater mobility of labour. From 539,400 inhabitants in 1864, the population rose to about 1,500,000 in 1900, largely by migration from rural areas (as late as 1910 only one-third of the population had been born in the city). By 1917 the total had risen to about 2,500,000.

The factory environment in St. Petersburg became a breeding ground for revolution. With the development of metalworking and engineering as the primary industries, there arose a skilled labour force, increasingly alert politically. Moreover, the factory workers, who numbered nearly 250,000 in 1914, tended to be concentrated in plants of far larger size than was usual in Russia; the Putilov (later renamed Kirov) armaments works alone employed about 13,000. It was thus easier for revolutionaries to spread their ideas and for workers' groups to organize than it was elsewhere in Russia. At the same time, the growth of the city was characterized by a belated and slow development of public transport, making it necessary for workers to live close to their place of work, in conditions of appalling overcrowding (more than 180,000 per square mile in the centre), squalor, and lack of sanitation. Throughout the period before 1917 the city administration was lacking in efficiency and often in funds, and the provision of all public services—even a water supply—was inadequate. Outbreaks of serious epidemics were commonplace.

The first serious revolutionary outbreak in St. Petersburg came on Dec. 14 (Dec. 26), 1825: the Decembrist insur-

Decembrist insurrection

rection, organized largely by liberal aristocrats and army officers seeking a liberal constitution and an end to serfdom. It was ruthlessly suppressed. During the rest of the 19th century, workers' revolutionary activity and unrest steadily increased, with ever more frequent strikes and outbreaks of violence. These culminated in the general strike of January 1905, when some 150,000 workers took part. On what became known as Bloody Sunday, January 9 (January 22), a mass march to the Winter Palace, bearing a petition to the tsar, was met by troops, who opened fire; more than 100 people were killed and hundreds more wounded. The situation developed into revolution, spreading throughout Russia. Although it was again crushed, underground revolutionary activity continued.

The outbreak of World War I in 1914 brought an upsurge of patriotic fervour centred on the tsar. The Germanic form of the city's name was changed to its Russian version, Petrograd. The military disasters of the war and the worsening economic situation, however, revived and intensified discontent. Transport inefficiencies led to severe shortages of food and other supplies. On Feb. 26 (March 11), 1917, with a general strike in effect, disorder broke out. The authorities were slow to act and lost all control. The Petrograd Soviet of Workers' and Soldiers' Deputies was formed on February 27 (March 12). On March 2 (March 15) the tsar abdicated. A provisional government was set up, eventually under the premiership of Aleksandr Kerensky. On April 3 (April 16) Lenin returned to Petrograd from Switzerland and set about organizing the overthrow of the provisional government. Demonstrations in July were suppressed, but on October 25 (November 7) Bolshevik-led workers and sailors stormed the Winter Palace, deposing the provisional government and establishing the Bolshevik Party in power.

Return of Lenin

The Russian Revolution of 1917, which changed the course of history, was spearheaded by the Petrograd proletariat and the sailors from Kronshtadt. In January 1918 a Constituent Assembly met in Petrograd, but the Bolsheviks, who had won only a minority of seats, dispersed it and consolidated their authority.

The Soviet period. Civil war reigned in Russia from 1918 to 1920, during which the Bolsheviks successfully defended their government against various Russian and foreign elements. In March 1918 the capital of the young Soviet state had been moved back to Moscow. The years of the civil war after the Revolution had a disastrous effect on the city's economy. Industry came very nearly to a standstill. The population fell sharply to 722,000 in 1920, a mere third of the pre-Revolutionary size. Many died of starvation. Recovery began when the war ended. In 1924, following Lenin's death, the city was renamed Leningrad. When in 1928 the era of five-year plans began, much of the initial burden of developing the national economy fell on the city and its established industrial plant and work force, especially in the provision of power equipment and machinery. This stimulated further growth; by 1939 the city was responsible for 11 percent of all Soviet industrial output, and its population had passed three million.

Then once again the city was struck by a period of loss and destruction. It was one of the initial targets of the German invasion in 1941; by September of that year, German troops were on the outskirts of the city and had cut off communication with the rest of the U.S.S.R., while Finnish troops advanced from the north. Many of the inhabitants and nearly three-quarters of the industrial plant were evacuated eastward ahead of the German advance. The remainder of the population and the garrison then began to endure what has become known as the 900-day siege; the German blockade in fact lasted 872 days, from Sept. 8, 1941, to Jan. 27, 1944. Leningrad put up a desperate and courageous resistance in the face of many assaults, constant artillery and air bombardment, and appalling suffering from shortages of supplies. An estimated 660,000 people died, a very high proportion from scurvy and starvation. In particular the exceptionally bitter winter of 1941–42, when temperatures fell to −40° F (−40° C), was one of extreme hardship and loss of life. The blockade proper was broken in January 1943, but it was another year before the Germans had been driven back from the outskirts. Enormous damage had been caused by the bombardment, and, before retreating, the Germans destroyed the palaces at Petrodvorets and Pushkin. For its role in the war, Leningrad was honoured with the title of "Hero City," a special defense medal, and the Order of Lenin. Not until the 1960s did the city regain its prewar size of three million; by the 1980s the population had passed four million.

900-day siege

The first postwar Soviet five-year plan was devoted in part to reconstruction of the city's industry and restoration of its architectural heritage. In the late 1950s a program of housing construction in the new suburbs got under way; renovation of highly sought-after inner-city apartments began in the 1970s. In the face of continuing construction and expansion, maintenance of the old city and modernization of the infrastructure became major problems. In response the city planners pioneered new forms of industrial administration, drawing on the city's strength as a scientific and technical centre and emphasizing the need to preserve its unique cultural heritage.

Major changes in the city's political life began to occur in the late 1980s, when the central government of the U.S.S.R. introduced reforms that encouraged greater democratization and openness. The nationwide legalization in March 1990 of political parties other than the Communist Party had an especially sharp impact on Leningrad, as two months later elections for the city council gave a group of reform-minded Communists and reformers outside the Communist Party a substantial majority of the council seats. The council quickly pushed for a variety of free-market measures and initiated a process of stripping the Communist Party of assets and privileges in the city. Reflecting this movement away from the city's Communist past, voters in a June 1991 citywide referendum chose to restore the city's name of St. Petersburg.

After the fall of the Soviet Union later that year, the 1990s brought the city an increase in crime and governmental corruption. Assassinations and kidnappings became common. Also during the decade the financial and economic focus of the country turned more toward Moscow and away from St. Petersburg. In 1994 election turnout was so poor that the city spent most of the year without a council; meanwhile, mayor Anatoly Sobchak (elected in 1990) hosted the Goodwill Games and lavish state visits. Sobchak was defeated in 1996. The city was hard hit by the 1998 Russian financial crisis.

BIBLIOGRAPHY. General descriptive works in English are NIGEL GOSLING, *Leningrad: History, Art, Architecture* (1965); and K. NEUBERT and J. NEUBERT, *Portrait of Leningrad* (1966). Two helpful books, both translated from Russian, are PAVEL KANN, *The Environs of Leningrad* (1981); and A. BEREZINA, *Leningrad: A Short Guide* (1980). Practical information can be found in EVAN MAWDSLEY and MARGARET MAWDSLEY (eds.), *Moscow and Leningrad* (1980). For historical background, see JAMES H. BATER, *St. Petersburg: Industrialization and Change* (1967); and E.M. ALMEDINGEN, *Tomorrow Will Come* (1941, reprinted 1983), a memoir spanning the revolutionary period. On the Siege of Leningrad, see HARRISON E. SALISBURY, *The 900 Days* (1969, reprinted 1985). An account of the city government is provided by DAVID T. CATTELL, *Leningrad: A Case Study of Soviet Urban Government* (1968). For politics and city planning in the 1970s, see DENIS J.B. SHAW, "Planning Leningrad," *Geographical Review*, 68(2):183–200 (April 1978); and BLAIR A. RUBLE, "Romanov's Leningrad," *Problems of Communism*, 32(6):36–48 (Nov.–Dec. 1983). The cultural history is discussed in JOHN GREGORY and ALEXANDER UKLADNIKOV, *Leningrad's Ballet: Maryinsky to Kirov* (1980). LOGAN ROBINSON, *An American in Leningrad* (1982), is an account by a Harvard law student.

Works in Russian include L.S. SHAUMIAN (ed.), *Leningrad: entsiklopedicheskiĭ spravochnik* (1957), a survey of the city's history, economics, public education, cultural and educational institutions, architecture and construction, public health service, science, literature, and art; S.M. SERPOKRYL (compiler), *Leningrad: Dostoprimechatel'nosti*, 3rd enlarged ed. (1974), on the basic architectural aggregations, squares, embankments, avenues, and suburban parks; and M.P. VIATKIN (ed.), *Ocherki istorii Leningrada*, 6 vol. (1955–70), an analysis of the development of the city and its construction, with an account of the stages of its revolutionary development. See also IURIĬ P. SELIVERSTOV (ed.), *Leningrad: istoriko-geograficheskiĭ atlas* (1981), a historical-geographic atlas.

(Y.M.D./R.A.F./M.McA./Ed.)

BOSTON PUBLIC LIBRARY
3 9999 04738 030 6
WITHDRAWN
No longer the property of the
Boston Public Library.
Sale of this material benefits the Library.